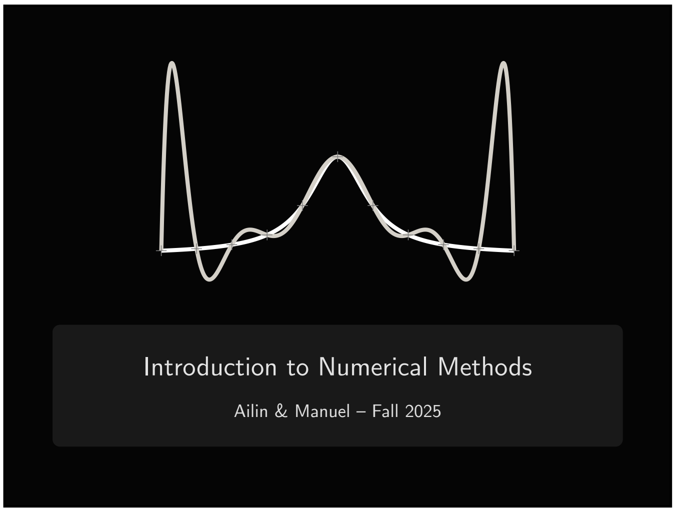


Introduction to Numerical Methods

Ailin & Manuel – Fall 2025

Notes

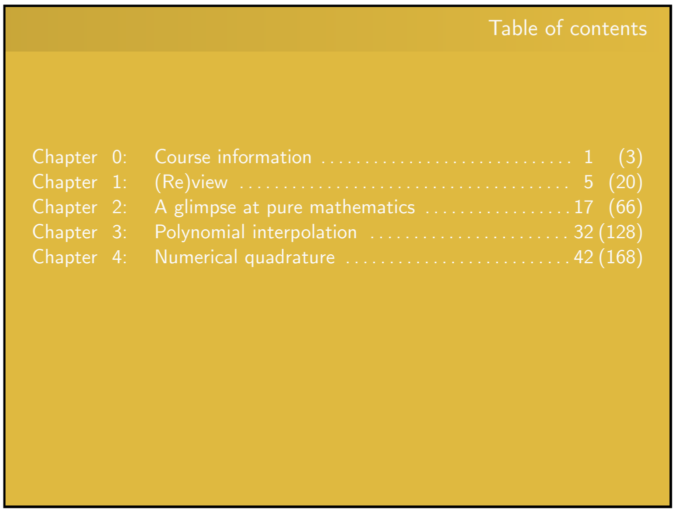
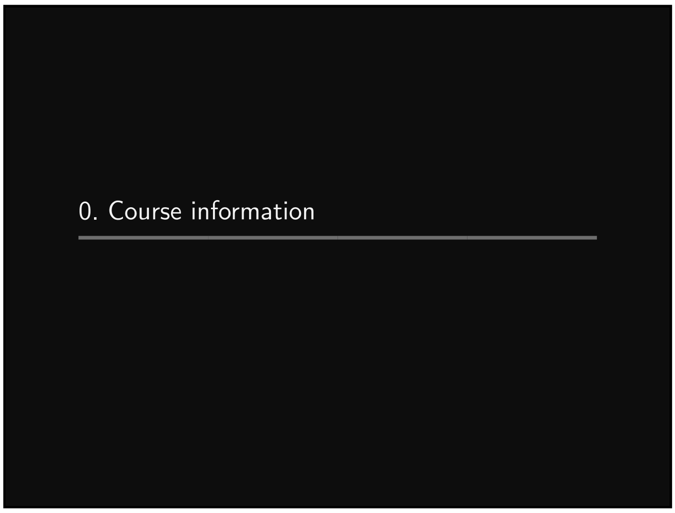


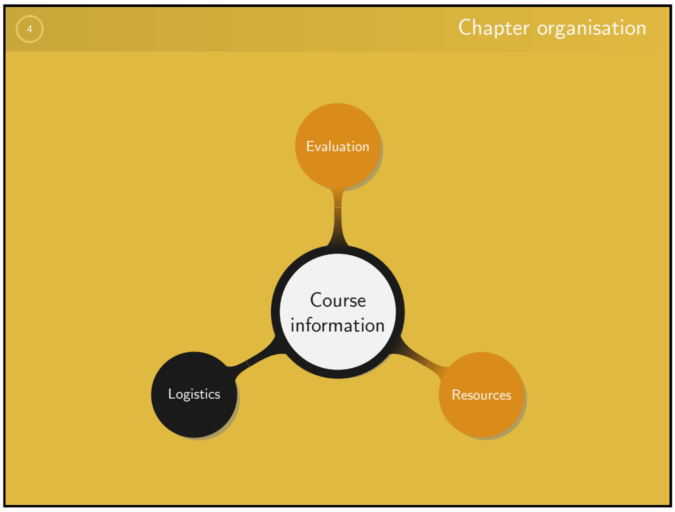
Table of contents

Notes



0. Course information

Notes



4

Chapter organisation

Notes

Basic information

Teaching team:

- Instructors:
  - Ailin (ailin.zhang@sjtu.edu.cn)
  - Manuel (charlem@sjtu.edu.cn)
- Teaching assistant: Zhiyuan (zhiyuan_wang@sjtu.edu.cn)

Course arrangements:

- Lectures:
  - Monday 14:00 – 15:40
  - Wednesday 14:00 – 15:40
- Refer to Canvas homepage for office hours times

Notes

Communication rules

To ensure smooth communication:

- Carefully comply with the course communication policies
- Prepend [MATH471] to the email subject, e.g. `[MATH471] h2 grade issue`
- When contacting a TA for an important matter, CC the instructor
- Use SJTU NetDisk service to share large files ($>$ 2 MB)
- Ensure Canvas email address is set to your SJTU address
- Check Canvas and SJTU emails at least once a day

- Never change your Canvas email address
- Never send large files by email
- Never post screenshots

Notes

Course objectives

This course splits into two parts:

- Numerical Analysis
  - Become familiar with the rigorous analysis of most common methods
  - Get solid mathematical foundations to tackle harder problems
- Numerical Methods
  - Understand how to apply numerical methods to solve problems
  - Be able to connect numerical analysis to numerical methods

Construct and assess the quality of methods to solve a given problem

Notes

Course workflow

Learning strategy:

- Course side:
  1. Understand the basic concept of numerical analysis
  2. Know the most common problems and their solutions
  3. Get an overview of the wide applications of numerical methods
- Personal side:
  1. Prove mathematical results
  2. Derive clean and clear algorithms from mathematical results
  3. Relate known strategies to new problems
  4. Perform extra research

Notes

9

Course outcomes

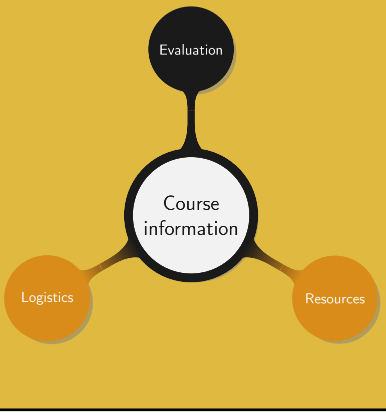
Detailed goals:

- Understand the mathematics behind numerical analysis
- Be proficient at using all the basic numerical methods
- Be able to assess the quality of a method
- Know how to perform function interpolation
- Know the common methods for numerical integration
- Solve various nonlinear equations numerically
- Solve various nonlinear optimisation problems numerically
- Solve systems of linear equations numerically
- Solve eigenvalue problems numerically
- Find various matrix decomposition numerically
- Solve various ordinary differential equations numerically

Notes

10

Chapter organisation



Evaluation

Course information

Logistics

Resources

Notes

11

Assignments

Homework:

- Total: $6 + 4$
- Content: basic concepts, prove results, derive algorithms

Projects:

- Total: 1
- Content: neural networks, theory and applications

Notes

12

Grading policy

Grade weighting:

- Projects: 25%
- Homework: 20%
- Midterm exam: 27.5%
- Final exam: 27.5%

Assignment submissions:

- Bonus: +10% for a work fully written in LaTeX, limited to 100%
- Penalty: −10% for a work not written in a neat and legible fashion
- Late policy: −10% per day, not accepted after three days

Grades will be curved with the median in the range $[\![B, B+]\!]$

Notes

13

Honor Code

General rules:

- Not allowed:
  - Reuse the work from other students
  - Reuse the work from the internet
  - Give too many details on how to solve an exercise
- Allowed:
  - Share ideas and understandings on the course
  - Provide general directions on where or how to find information

Shared details should never solve a question

Notes

14

Honor Code

Documents allowed during the exams: an A4 paper sheet with original handwritten notes

Default Honor Code policy for group works:

- Every student in a group is responsible for his group submission
- In case of violation, the whole group is sent to Honor Council

Default Honor Code policy for generative AI: not allowed

Notes

15

Special circumstances

Contact us as early as possible when:

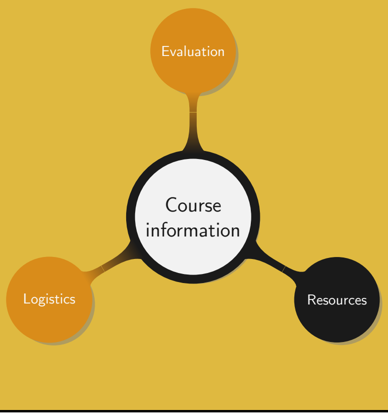
- Facing special circumstances, e.g. full time work, illness
- Feeling late in the course
- Feeling to work hard without any result

Any late request will be rejected

Notes

16

Chapter organisation



Notes

17

Course resources

Canvas:

- Syllabus
- Lecture slides
- Homework
- Projects
- Announcements
- Grades
- Surveys

Gitea:

- Extra documents
- Course support

Mattermost:

- Announcements
- Quick questions

Notes

18

References

The course is self-contained, do not abuse external resources

Useful places where to find information:

- *Fundamentals of Engineering Numerical Analysis*, Moin
- *Introduction to Algorithms*, Cormen, Leiserson, Rivest, and Stein
- *Numerical Methods using MATLAB*, Mathews, and Fink
- *Numerical Recipes*, Press, Teukolsky, Vetterling, and Flannery
- Search information online, i.e. $\{\textit{websites} \setminus \{\textit{non-English websites}\}\}$

Notes

19

Key points

- Work regularly, do not wait the last minute
- Respect the Honor Code
- Go beyond what is taught
- Do not learn, understand
- Keep in touch with us
- Feedback and suggestions are always much appreciated

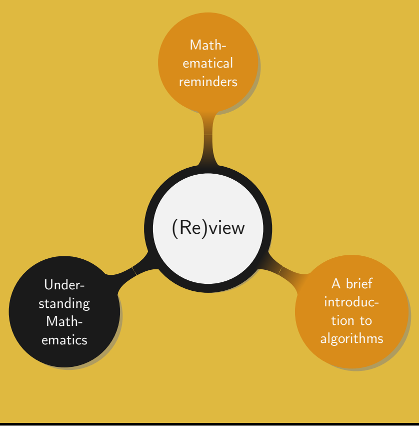
Notes

1. (Re)view

Notes

21

Chapter organisation



Notes

Mathematics is easy

Notes

23

What is Mathematics?

> *The universe cannot be read until we have learnt its language and become familiar with the characters in which it is written. It is written in mathematical language, the letters being the triangles, circles, and other geometrical figures without which it would be humanly impossible to comprehend a single word.*
>
> Galileo Galilei

Notes

24

Applied and Pure Mathematics

Mathematics splits into two main types:

- Applied: develop mathematical methods to answer the needs from other fields of science such as physics, engineering, or finance.
- Pure: study of abstract concepts that are totally disconnected from the real world.

Relation between pure and applied Mathematics:

- Applied Mathematics highly relies on pure Mathematics to model the real world
- Pure Mathematics is often the result of real world abstractions

Notes

25

Mathematical abstraction

Abstraction process:

1. Take a real world problem
2. Phrase it into mathematical terms
3. Eliminate any "physical dependence"
4. Widen the application by generalizing the concept

Notes

26

Numerical Analysis

Goal: use numerical approximations to solve mathematical analysis problems

Applications: physics, engineering, biology, weather prediction, finance

Remark. Numerical analysis is not to be confused with numerical methods: the latter focuses on the study of methods to solve a given problem, while the goal of the former is to develop them. As a result numerical analysis requires a better understanding of the underlying mathematics.

Notes

27

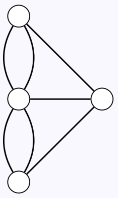
Topology

Goal: classify objects with respect to their behavior under smooth deformation

Applications: study the inherent structure of an object with respect to properties such as convergence, connectedness, and continuity

Example.

- Euler's seven bridges of Königsberg
- Is it possible to walk across each bridge exactly once?
- Non-affecting factors: distance between bridges, shape of the "islands"
- Affecting factor: number of "islands" and bridges, as well as their arrangement



Notes

28

Algebra

Goal: classify sets with respect to how their elements behave under some given operations

Applications: number theory, analysis, geometry, theoretical physics

Remark. Although from a historical point of view algebra started with the study of objects with the goal of discovering useful properties, this approach changed during the 20-th century. Nowadays sets are studied from an abstract and formal perspective without taking into account any application.

Notes

29

Measure theory

Goal: find a systematic way to measure the size of any set

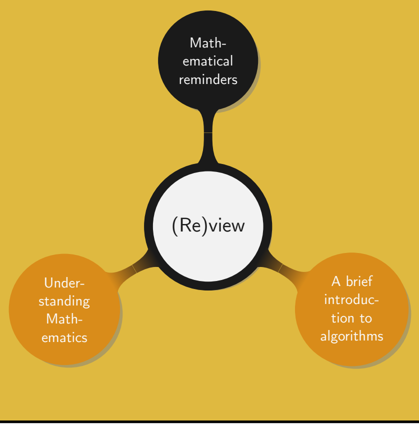
Applications: probability theory, analysis

Remark. The idea consists in defining a function called *measure* which maps subsets of a set to positive real values. The main issue is how to keep the size consistent, i.e. given a countable number of disjoint subsets the measure of their union must be equal to the sum of their measures.

Notes

30

Chapter organisation



Notes

31

Algebraic structures

Definition (Group)

A *group* is a pair $(G, +)$ consisting of a set G and a *group operation* $+: G \times G \to G$ that verifies the following properties:

- i *Associativity:* $a + (b + c) = (a + b) + c$ for all $a, b, c \in G$;
- ii *Existence of a unit element:* there exists an element $e_+ \in G$ such that for all $a \in G$, $a + e_+ = e_+ + a = a$;
- iii *Existence of an inverse:* for any $a \in G$ there exists an element $a^{-1} \in G$ such that $a + a^{-1} = a^{-1} + a = e_+$;

A group is said to be *abelian* if it also verifies

- iv *Commutativity:* $a + b = b + a$ for all $a, b \in G$;

Examples.

- $(\mathbb{Z}, +)$ is an abelian group but $(\mathbb{Z}, \times)$ is not.
- Let $\mathcal{F}(X, \mathbb{R})$ be the set of the functions from a set X into $\mathbb{R}$. Then $(\mathcal{F}(X, \mathbb{R}), +)$ is a group but $(\mathcal{F}(X, \mathbb{R}), \circ)$ is not.

Notes

32

Algebraic structures

Definition (Ring)

A *ring* is a triple $(R, +, \cdot)$ consisting of a set R and two *binary operations* $(+)$ and $(\times)$ defined from $R \times R$ to R , and such that

- i $(R, +)$ is an abelian group;
- ii *Multiplicative unit:* there exists an element $e_\times \in R$ such that for all $a \in R$
$$a \times e_\times = e_\times \times a = a;$$
- iii *Associativity:* for any $a, b, c \in R$, $a \times (b \times c) = (a \times b) \times c$;
- iv *Distributivity:* for any $a, b, c \in R$,
$$a \times (b + c) = (a \times b) + (a \times c), (b + c) \times a = (b \times a) + (c \times a);$$

A ring is said to *commutative* if it also verifies

- v *Commutativity:* $a \times b = b \times a$ for all $a, b \in R$;

Notes

Algebraic structures

Definition (Field)

Let $(\mathbb{K}, +, \times)$ be a commutative ring with unit element of addition 0 and unit element of multiplication 1. Then $\mathbb{K}$ is a *field* if

- i $0 \neq 1$
- ii For every $a \in \mathbb{K}^*$ there exists an element a^{-1} such that
$$a \times a^{-1} = 1.$$

Remark. Another way of writing this definition is to say that $(\mathbb{K}, +, \times)$ is a field if $(\mathbb{K}, +)$ and $(\mathbb{K} \setminus \{0\}, \times)$ are abelian groups, $0 \neq 1$, and $(\times)$ distributes over $(+)$.

Notes

Algebraic structures

Definition (Vector space)

A *vector space* $(V, +, \cdot)$ over a field $(\mathbb{K}, +, \times)$ is a set on which the $(+)$ and $(\cdot)$ operations are defined and closed. Moreover, for any $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $\alpha, \beta \in \mathbb{K}$ the following conditions hold.

- i $(V, +)$ is an abelian group;
- ii *Compatibility of scalar and field multiplications:*
$$\alpha \cdot (\beta \cdot \mathbf{v}) = (\alpha \times \beta) \cdot \mathbf{v};$$
- iii *Distributivity of scalar multiplication for vector addition:*
$$\alpha \cdot (\mathbf{u} + \mathbf{v}) = \alpha \cdot \mathbf{u} + \alpha \cdot \mathbf{v};$$
- iv *Distributivity of scalar multiplication for field addition:*
$$(\alpha + \beta) \cdot \mathbf{v} = \alpha \cdot \mathbf{v} + \beta \cdot \mathbf{v};$$
- v *Scalar multiplication identity:* for all $\mathbf{v} \in V, 1_F \cdot \mathbf{v} = \mathbf{v}$;

Elements of V are called *vectors* and the ones of F *scalar*.

Notes

Vector spaces

For the sake of simplicity we do not denote vectors using bold fonts, for instance we now write u instead of $\mathbf{u}$.

Definitions

1. Let V be a vector space. A subset of V whose elements are linearly independent and span V is called a *basis* of V.
2. If V is spanned by a finite number of vectors then V is said to have a *finite dimension*. The *dimension* of V is defined as the cardinal of the basis.
3. Let V_1 and V_2 be two vector spaces over a field $\mathbb{K}$. A function f defined from V_1 into V_2 is a *linear map* or *linear map* if:
   - For any $x, y \in V_1$, $f(x + y) = f(x) + f(y)$;
   - For any $x \in V_1$ and $a \in \mathbb{K}$, $f(ax) = af(x)$;
4. A bijective linear map is called an *isomorphism*.

Notes

Kernel of a linear map

Definition

Let E and F be two vector spaces, and f be a linear map from E to F. The *kernel* of f, noted $\ker f$, is the set of all the elements x from E such that $f(x) = 0$.

Lemma

For two vector spaces E and F of similar dimension, a linear map f from E into F is bijective if and only if $\ker f = \{0\}$.

Proof. If f is bijective then clearly $\ker f = \{0\}$.

Since $\ker f = \{0\}$, f is at least injective. Moreover as E and F have similar dimension, a basis from E is mapped by f into a basis of F. Hence f is bijective. □

Notes

37

Algebraic structures

Let $\mathbb{K}$ be $\mathbb{R}$ the field of the reals, or $\mathbb{C}$ the field of complexes.

Definitions

Let V be a vector space over a field $\mathbb{K}$.

1. The map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{K}$ is an *inner product*, if for $u, v, w \in V$ and $a \in \mathbb{K}$:
   - i. *Conjugate symmetry:* $\langle u, v \rangle = \overline{\langle v, u \rangle}$;
   - ii. *Linearity:* $\langle au, v \rangle = a\langle u, v \rangle$ and $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$;
   - iii. *Positivity:* $\langle u, u \rangle \geq 0$ and $\langle u, u \rangle = 0 \Leftrightarrow u = 0$;

   A vector space with an inner product is an *inner product space*.
2. The map $\| \cdot \| : V \to \mathbb{R}_+$ is a *norm*, if for any $u, v \in V$ and $a \in \mathbb{K}$:
   - i. $\|u\| = 0 \Leftrightarrow u = 0$;
   - ii. $\|au\| = |a|\|u\|$;
   - iii. *Triangle inequality:* $\|u + v\| \leq \|u\| + \|v\|$;

   A vector space with an norm is an *normed vector space*.

Notes

38

Normed vector space

We now show that inner product spaces have a naturally defined norm, inducing a normed vector space. In this course we restrict our attention to the simpler case $\mathbb{K} = \mathbb{R}$.

Proposition

Let E be an inner-product space over $\mathbb{R}$. Then E is a normed vector space and for $u \in E$ the norm is defined by $|u| = \sqrt{\langle u, u \rangle}$.

Proof. We simply verify definition 1.37.

(i) Clearly $|u| = 0$ if and only if $u = 0$.

(ii) For $\lambda \in \mathbb{R}$, $|\lambda u| = \sqrt{\langle \lambda u, \lambda u \rangle} = |\lambda| \sqrt{\langle u, u \rangle} = |\lambda||u|$.

(iii) We have

$$\begin{aligned} |u + v|^2 &= \langle u + v, u + v \rangle = |u|^2 + |v|^2 + 2\langle u, v \rangle \\ &\leq |u|^2 + |v|^2 + 2|u||v| = (|u| + |v|)^2 \end{aligned} \qquad (1.1)$$

Hence from equation (**??**) we obtain $|u + v| \leq |u| + |v|$. □

Notes

39

Cauchy-Schwarz inequality

Since $| \cdot |$ is a norm we denote it $\| \cdot \|$.

Theorem (Cauchy-Schwarz inequality)

Let E be an inner product space over $\mathbb{R}$ and $u, v \in E$. Then

$$|\langle u, v \rangle| \leq \|u\| \cdot \|v\|.$$

Proof. Let $\lambda \in \mathbb{R}$. Using the notation from the theorem we have

$$0 \leq |u - \lambda v|^2 = \langle u, u \rangle - 2\lambda \langle u, v \rangle + \lambda^2 \langle v, v \rangle.$$

Viewing the equation as a quadratic polynomial, its discriminant $\langle u, v \rangle^2 - \langle u, u \rangle \langle v, v \rangle$ cannot be positive since it is over $\mathbb{R}$ and we do not want $|u - \lambda v|^2$ to be negative.

Hence we obtain

$$|\langle u, v \rangle| \leq \sqrt{\langle u, u \rangle} \cdot \sqrt{\langle v, v \rangle} = \|u\| \cdot \|v\|.$$

□

Notes

40

Cauchy-Schwarz inequality

Remarks.

- Although the previous proof of the Cauchy-Schwarz inequality is only valid when $\mathbb{K} = \mathbb{R}$, the result can be proven when $\mathbb{K} = \mathbb{C}$ at the cost of a slightly more complex strategy.
- The equality only happens when the vectors u and v are linearly dependent.

Proposition

Let E be an inner-product space over $\mathbb{R}$ and $u, v \in E$.

1. Parallelogram law: $\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2)$;
2. $\|u + v\|^2 - \|u - v\|^2 = 4\langle u, v \rangle$;

Notes

41

Normed vector spaces

Proof. The proofs is trivial as we have

$$\begin{aligned}\|u+v\|^2+\|u-v\|^2 &= \|u\|^2+\|v\|^2+2\langle u,v\rangle+\|u\|^2+\|v\|^2-2\langle u,v\rangle\\ &= 2(\|u\|^2+\|v\|^2),\end{aligned}$$

and

$$\begin{aligned}\|u+v\|^2-\|u-v\|^2 &= \|u\|^2+\|v\|^2+2\langle u,v\rangle-\|u\|^2-\|v\|^2+2\langle u,v\rangle\\ &= 4\langle u,v\rangle.\end{aligned}$$

□

Remark. Most normed vector spaces over $\mathbb{R}$ and $\mathbb{C}$ do not have an inner-product. In other words, the converse of proposition 1.38 is false. However if for a norm the parallelogram law is verified then this norm arises from an inner-product.

Notes

42

Normed vector spaces

Definition

For two normed vector spaces E and F, we denote by $\mathrm{L}(E,F)$ the $\mathbb{K}$-vector space of the linear maps from E to F. We define the supremum norm of a linear map u as

$$\|u\| = \sup_{x\in E\setminus\{0\}} \frac{\|u(x)\|}{\|x\|}.$$

An element of $\mathrm{L}(E,F)$ is said to be *bounded* if it has finite norm. All bounded linear maps form a subspace $\mathcal{L}(E,F)$ of $\mathrm{L}(E,F)$.

Remark. Note that E are F are two vectors spaces which can be functions spaces themselves, i.e. their vectors are functions. This means that u is a linear map which is applied to a function. Differentiation and integration are common examples of such u.

Notes

43

Linear maps

Proposition

Let E and F be two vector spaces over $\mathbb{K}$ and $u \in \mathcal{L}(E,F)$. If $(x_n)_{n\in\mathbb{N}}$ is a sequence converging to l, then $\big(u(x_n)\big)_n$ converges to $u(l)$.

Proof. Let n be a positive integer. Then

$$\|u(x_n)-u(l)\| = \|u(x_n-l)\| \leq \|u\|\|x_n-l\|,$$

i.e. if $(x_n)_n$ converges, then so does $\big(u(x_n)\big)_n$. □

Remark. This result is only valid for $u \in \mathcal{L}(E,F)$, i.e. when u has finite norm. If u had infinite norm then we would not be able to bound $\|u(x_n)-u(l)\|$. In particular to find a counter example it is necessary to think of an infinite dimensional space, for instance by looking at a sequence of converging functions.[1]

[1] Refer to h1.6 for an example of discontinuous liner map.

Notes

44

Metric spaces and continuity

Definitions

1. Let X be a set. A *distance* or *metric* on X is a function $d : X\times X \to \mathbb{R}_+$ such that for all $x,y,x \in X$:
   - i. $d(x,y)=0$ if and only if $x=y$;
   - ii. $d(x,y)=d(y,x)$;
   - iii. $d(x,z)\leq d(x,y)+d(y,z)$ (triangular inequality);

   A set endowed with a distance is called a *metric space*.
2. Let X and Y be two metric spaces endowed with distances d and d', respectively. A function $f : X \to Y$ is *continuous* at a, if for all $\varepsilon \in \mathbb{R}_+^*$, there exists $\eta \in \mathbb{R}_+^*$, such that $d(a,x) < \eta$ implies $d'(f(a),f(x)) < \varepsilon$.
3. Let X and Y be two metric spaces. A function $f : X \to Y$ is *continuous* if it is continuous at any point of X.

Notes

45

More on continuity

We now introduce more refined notions of continuity.

Definitions

Let X and Y be two metric spaces endowed with metrics d and d', respectively, and $f : X \to Y$.

1. The function f is *uniformly continuous* if for all $\varepsilon \in \mathbb{R}_+^*$, there exists $\eta \in \mathbb{R}_+^*$, such that $d(x, y) < \eta$ implies $d'(f(x), f(y)) < \varepsilon$, for $x, y \in X$.
2. The function f is *Lipschitz continuous* if there exists k such that for any $x, y \in X$, $d'(f(x), f(y)) \leq kd(x, y)$.
3. Let X be an interval of $\mathbb{R}$. The function f is *absolutely continuous* if for all $\varepsilon \in \mathbb{R}_+^*$, there exists $\eta \in \mathbb{R}_+^*$, such that whenever $\{[x_i, y_i] : i = 1, \cdots, n\}$ is a finite collection of mutually disjoint subintervals of X, $\sum_{i=0}^n |y_i - x_i| < \eta$ implies $\sum_{i=0}^n d'(f(x_i), f(y_i)) < \varepsilon$.

Notes

46

More on continuity

Remark.

- i. A Lipschitz continuous function is uniformly continuous.
- ii. A uniformly continuous function is continuous.
- iii. A continuous function is rarely uniformly continuous.
- iv. An absolutely continuous function on a finite union of closed real intervals is uniformly continuous.
- v. A Lipschitz continuous function on a closed real interval is absolutely continuous.

To derive a useful result on the continuity of linear maps over a normed vector space we introduce the notion of ball.

Definition

Let (X, d) be a metric space, $a \in X$ and $r \in \mathbb{R}_+^*$. The set of all the points at a distance strictly less than r from a is called an *open ball*, or simply a *ball*, of radius r and is denoted $B_d(a, r)$.

Notes

47

Normed vector spaces and continuity

Theorem

Let E and F be two normed vector spaces over $\mathbb{K}$ and u be a linear map. The following conditions are equivalent.

- i. $u \in \mathcal{L}(E, F)$;
- ii. u is uniformly continuous;
- iii. u is continuous;
- iv. u is continuous at any $x \in E$;
- v. u is bounded over any bounded subset of E;
- vi. There exists $a \in E$ and $r \in \mathbb{R}_+^*$ such that u is bounded over $B(a, r)$;

Proof. The implications (ii) ⇒ (iii) ⇒ (iv) and (v) ⇒ (vi) are trivial.

Notes

48

Normed vector spaces and continuity

(i) ⇒ (ii). If $u \in \mathcal{L}(E, F)$, then by linearity, for $x, y \in E$ we have

$$\|u(x) - u(y)\| = \|u(x - y)\| \leq \|u\|\|x - y\|.$$

So u is $\|u\|$-Lipschitz continuous, and in particular uniformly continuous.

(iv) ⇒ (v). If u is continuous at a then there exists $r > 0$ such that $\|u(x) - u(a)\| < 1$ when $\|x - a\| < r$. Thus for $x \in B(a, r)$ we have $\|u(x)\| \leq 1 + \|u(a)\|$, meaning that u is bounded over $B(a, r)$. But as any bounded subset of E is contained in a ball, this shows that u is bounded over any bounded subset of E.

(vi) ⇒ (i). Without loss of generality we prove the result for $a = 0$. Let $x \in E$ and $y \in B(0, r)$ such that $\|y\|x = \|x\|y$. Then we apply u and by linearity we get $\|y\| \cdot \|u(x)\| = \|x\| \cdot \|u(y)\|$.

Since y is in $B(0, r)$ we know the existence of m such that $\|u(y)\| \leq m$ and $\|y\| < r$. Setting $M = \max(m, r)$, we get $\|u(x)\| \leq \frac{M}{r}\|x\|$, meaning that $\|u(x)\|$ is bounded. As expected, this shows that $\|u\| = \sup \frac{\|u(x)\|}{\|x\|}$ is bounded. □

Notes

We now turn our attention to the notion of differentiability. Let E and F be two vector spaces of finite dimension over $\mathbb{R}$ and U be a non-empty open from E. We then define f as a function from U into F.

Lemma

If for $u \in \mathrm{L}(E, F)$, $\lim_{x \to 0} \frac{u(x)}{\|x\|} = 0$, then u is the null function.

Proof. By definition of the limit when x tends to 0, for any $\varepsilon > 0$, there exists $\eta > 0$ such that $\|u(x)\| \leq \varepsilon \|x\|$, if $\|x\| \leq \eta$. To prove that u is null, we need to show that this remains true for any $y \in E \setminus \{0\}$. We define $z = \frac{\eta}{\|y\|} y$. Then by the linearity of u

$$\|u(z)\| = \eta \frac{\|u(y)\|}{\|y\|}.$$

But as $\|z\| = \eta$, we can write $\|u(z)\| \leq \varepsilon \|z\| = \varepsilon \eta$. Hence, for any $y \neq 0$ from E, $\|u\| = \frac{\|u(y)\|}{\|y\|} \leq \varepsilon$. □

Notes

Proposition

There exists at most one linear map $\ell \in \mathrm{L}(E, F)$ such that, for $a \in U \subset E$,

$$\lim_{x \to 0} \frac{1}{\|x\|} \left[f(a + x) - f(a) - \ell(x) \right] = 0.$$

Proof. Let $\ell_1, \ell_2 \in \mathrm{L}(E, F)$ matching the condition from the proposition. Then

$$\lim_{x \to 0} \frac{(\ell_1 - \ell_2)(x)}{\|x\|} = 0.$$

From lemma 1.49 it means ℓ_1 and ℓ_2 are equal. □

Remark. From this result we see that ℓ is tangent to f at a.

Notes

Definitions

1. When it exists, the linear map ℓ is called the *differential of f at a*. Then the function f is said to be *differentiable* at a. The differential of f at a is denoted $\mathrm{d}f(a)$.
2. If for all $a \in U \subset E$, $\mathrm{d}f(a)$ exists then we define the application
$$\begin{aligned} \mathrm{d}f : U &\longrightarrow \mathrm{L}(E, F) \\ x &\longmapsto \mathrm{d}f(x). \end{aligned}$$
This application is called the *differential of f*.
3. When f is *n times differentiable* on $U \subset E$, $n \in \mathbb{N}^*$, the differential of order n of f is n-linear and given by
$$\begin{aligned} \mathrm{d}^n f : U &\longrightarrow \mathrm{L}_n(E, F) = \mathrm{L}(E, \mathrm{L}_{n-1}(E, F)) \\ x &\longmapsto \mathrm{d}^n f(x). \end{aligned}$$

Notes

Definitions

Let f be a function n times differentiable over $U \subset E$.

1. The function f is said to be in the differentiability class C^n if $\mathrm{d}^n f$ is continuous over U.
2. The function f is C^∞ if it is C^n for all $n \in \mathbb{N}$.

Remark.

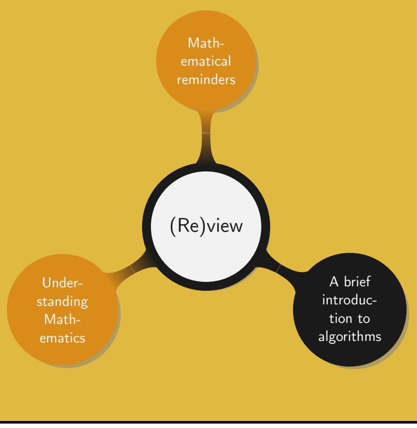
- i. We will often write $f \in C^n[a, b]$ to say that f is defined over $[a, b] \subset \mathbb{R}$, and is in the differentiability class C^n.
- ii. The set of all the $C^n[a, b]$ functions is a vector space.
- iii. A continuous function is sometimes said to be C^0.

In the next section we focus on the less abstract notion of algorithm. In particular we will see how they should be designed and presented in the assignments and exams.

Notes

53

Chapter organisation


Math-ematical reminders
(Re)view
Under-standing Math-ematics
A brief introduc-tion to algorithms

Notes

54

Algorithm

Algorithm: Recipe telling the computer how to solve a problem.

Example. Detail an algorithm to prepare a jam sandwich.

Actions: `cut, listen, spread, sleep, read, take, eat, dip`
Objects: `knife, guitar, bread, honey, jam jar, sword`

Algorithm. (*Sandwich making*)

```
Input  : 1 bread, 1 jamjar, 1 knife
Output: 1 jam sandwich
take the knife and cut 2 slices of bread;
dip the knife into the jamjar;
spread the jam on the bread, using the knife;
assemble the 2 slices together, jam on the inside;
```

Notes

55

Basic algorithms

An algorithm systematically solves a *well-defined* problem:

- The *Input* is clearly expressed
- The *Output* solves the problem
- The *Algorithm* provides a precise step-by-step procedure starting from the Input and leading to the Output

Algorithms can be described using one of the three following ways:

- English
- Pseudocode
- Programming language

Notes

56

A first example

Algorithm. (*Insertion Sort*)

```
Input  : $a_1, \dots, a_n$, $n$ unsorted elements
Output: the $a_i, 1 \leq i \leq n$, in increasing order
for $j \leftarrow 2$ to $n$ do
    $i \leftarrow 1$;
    while $a_j > a_i$ do $i \leftarrow i + 1$;
    $m \leftarrow a_j$;
    for $k \leftarrow 0$ to $j - i - 1$ do $a_{j-k} \leftarrow a_{j-k-1}$;
    $a_i \leftarrow m$
end for
return $(a_1, \dots, a_n)$
```

Notes

57

A first problem

Setup: a robot arm solders chips on a board in n contact points

Goal: minimize the time to attach a chip to the board

Assumptions:

- The arm moves at constant speed
- Once a chip has been attached another one is soldered

Defining the Input and Output:

- Input: a set S of n points in the plane
- Output: the shortest path visiting all the points in S

Notes

58

A first solution

Algorithm. (*Nearest neighbor*)

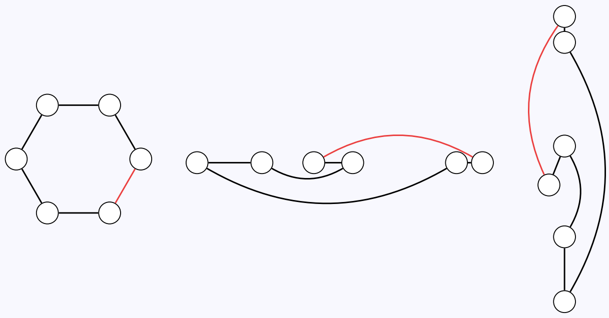
```
Input  : a set $S = \{s_0, \cdots, s_{n-1}\}$ of $n$ points in the plane
Output: the shortest cycle visiting all the points in $S$
$p_0 \leftarrow s_0$;
for $i \leftarrow 1$ to $n-1$ do
    $p_i \leftarrow$ closest unvisited neighbor to $p_{i-1}$;
    Visit $p_i$;
end for
return $\langle p_0, \dots, p_{n-1} \rangle$
```

Notes

59

A first failure

How does the nearest neighbor algorithm (1.58) perform in the following three cases?



Notes

60

A second solution

Algorithm. (*Closest pair*)

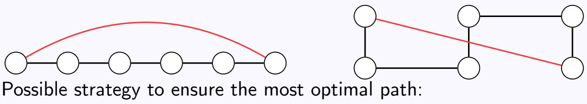
```
Input  : a set $S$ of $n$ points in the plane
Output: the shortest cycle visiting all the points in $S$
for $i \leftarrow 1$ to $n-1$ do
    $d \leftarrow \infty$;
    foreach pair of end points $\langle s, t \rangle$ from distinct vertex chains do
        if $dist(s,t) \leq d$ then
            $s_m \leftarrow s$; $t_m \leftarrow t$; $d \leftarrow dist(s,t)$;
        end if
    end foreach
    Connect $s_m$ and $t_m$ by an edge;
end for
return all the points starting from one of the two end points
```

Notes

61

A second failure

Applying the closest pair algorithm (1.60) on the following vertices arrangement yields the two graphs:



Possible strategy to ensure the most optimal path:

- Enumerate all the possible paths
- Select the one that minimizes the total length

Drawback: for only 20 vertices 20! = 2432902008176640000 paths have to be explored...

Notes

62

The first few lessons

A difference:

- Algorithm: always output a correct result
- Heuristic: idea serving as a guide to solve a problem with no guarantee of always providing a good solution

Correctness and efficiency:

- An algorithm working on a set of input does not imply it will work on all instances
- Efficient algorithm totally solving a problem might not exist

Notes

63

Solving problems using algorithms

Common traps when defining the Input and Output:

- Are they precise enough?
- Can all the Input be easily and efficiently generated?
- Could there be any confusion on the expected Output?

Example. For an Output, what does it mean to “find the best route”?

The shortest in distance, the fastest in time, or the one minimizing the number of turns?

Conclusion: where to start (Input) and where to go (Ouput) must be expressed in simple, precise, and clear terms.

Notes

64

Incorrectness

Finding good counter-examples:

- Seek simplicity: make it clear why the algorithm fails
- Think small: algorithms failing for large Input often fail for smaller one
- Test the extremes: study special cases, e.g. inputs equal, tiny, huge...
- Think exhaustively: test whether all the possible cases are covered by the algorithm
- Track weaknesses: check if the underlying idea behind the algorithm has any “unexpected” impact on the output

Notes

Key points

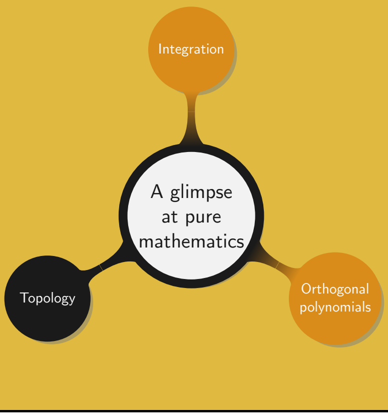
- How hard mathematics is?
- What are topology, algebra, and measure theory?
- For E and F two normed vector spaces, what is $\mathcal{L}(E, F)$?
- How to write pseudocode in the assignments and exams
- How to ensure the correctness of an algorithm

Notes

2. A glimpse at pure mathematics

Notes

Chapter organisation



Notes

Motivation

In this course the two main goals are to:

- Clearly understand why numerical analysis "works"
- Precisely know how to apply numerical methods

To properly complete those two goals we need some results from pure mathematics. The aim of this chapter is to provide solid mathematical bases for the rest of the course.

In particular while some results will be proven in the lectures others will only be presented in the slides and their proof addressed in the assignments.

Notes

69

Topological spaces

Since topology focuses on how the elements of a set relate to each others without really considering these elements themselves very little can be said about general topological spaces.

Definitions

1. A topological space is a pair $(X, \mathcal{T})$ where X is a set and $\mathcal{T}$ is a family of subsets of X called the *topology* of X. Its elements are called *open sets* and satisfy the following three properties:
   - i $\emptyset, X \in \mathcal{T}$;
   - ii Any union, finite or not, of open subsets of X is open;
   - iii The intersection of finitely many open subsets of X is open;
2. Let x be in X. Any open set containing x is called a *neighborhood* of x.

Notes

70

Topological spaces

Example.

- If all the subsets of $\mathbb{Z}$ are viewed as open then $\mathbb{Z}$ becomes a topological space.
- If the open sets of $\mathbb{R}$ are defined in the usual way then $\mathbb{R}$ is a topological space.
- If we decide that the open sets of $\mathbb{R}$ are $\emptyset$, $\mathbb{R}$, and all the half lines $\{x \in \mathbb{R} : x \geq a \text{ and } a \in \mathbb{R}\}$, then $\mathbb{R}$ is not a topological space.

In this course we restrict our consideration of topological spaces to metric spaces. Therefore now need to prove that any metric space is a topological space. But first we introduce a few new definitions.

Notes

71

Metric spaces as topological spaces

Definitions

1. Let X be a metric space. A subset of X is said to be *open* if it can be written as a union of open balls.
2. Let X be a metric space. A subset U of X is said to be *closed* if $X \setminus U$ is an open subset of X.

Theorem

A metric space is a topological space.

Proof. In order to prove the result we only need to verify that the three properties from definition 2.69 apply to metric spaces.

Notes

72

Metric spaces as topological spaces

(i) Since the empty set can be written as the union of the empty set of open balls, $\emptyset$ is open. Similarly X is open for it can be written as $X = \bigcup_{a \in X} B(a, 1)$.

(ii) This is exactly the definition of an open (definition 2.71).

(iii) We reason by induction. When there is a single open U_1, the result is straightforward.

Assume the result is true for $n \in \mathbb{N}^*$, i.e. the intersection of n open subsets of X is an open of X, and prove it is also true for $n + 1$ open subsets.

By applying the induction hypothesis the problem boils down to proving that the intersection of two opens is an open subset of X.

Notes

73

Metric spaces as topological spaces

If $U_1 \cap U_2 = \emptyset$, then the result is given by (i). Therefore lets assume $U_1 \cap U_2 \neq \emptyset$ and a be in this intersection. Then there exist $r_1, r_2 \in \mathbb{R}_+^*$ such that $B(a, r_1) \subset U_1$ and $B(a, r_2) \subset U_2$. Moreover for $r = \min(r_1, r_2)$ we have $B(a, r) \subset U_1 \cap U_2$.

Hence for any $a \in U_1 \cap U_2$, there exists $r_a \in \mathbb{R}_+^*$ such that $B(a, r_a) \subset U_1 \cap U_2$. Finally writing the intersection

$$U_1 \cap U_2 = \bigcup_{a \in U_1 \cap U_2} B(a, r_a),$$

shows that $U_1 \cap U_2$ is an open subset of X.

By the induction principle the intersection of finitely many open subsets of X is open. □

Notes

74

Continuous maps

As topology focuses on the classification of objects with respect to their behaviour under the action of a continuous map we now introduce and prove a fundamental theorem relating the continuity of a function to the closed and open sets.

Theorem

Let X and Y be two metric spaces and $f : X \to Y$. Then the following conditions are equivalent

- i f is continuous;
- ii For any open $V \subset Y$, $f^{-1}(V)$ is an open of X;
- iii For any closed $V \subset Y$, $f^{-1}(V)$ is closed in X;

Proof. (ii) ⇔ (iii) This is trivial: if $V \subset Y$, then $f^{-1}({}^cV) =^c (f^{-1}(V))$.

Notes

75

Continuous maps

(i) ⇒ (ii) Let V be an open from Y and $U = f^{-1}(V)$. If we take $a \in U$, then $b = f(a) \in V$. And since V is open, it is a neighborhood of b. Using the continuity of f we conclude that U is a neighborhood of a^2. Hence U is a neighborhood for all its points, i.e. U is open.

(ii) ⇒ (i) Let $a \in X$, $b = f(a)$. Then $V = B(b, \varepsilon)$, for some $\varepsilon > 0$, is a neighborhood of b. There exists an open subset W of Y containing b and such that $W \subset V$. Then $U = f^{-1}(W)$ is an open from X which contains a. And since $U \subset f^{-1}(V)$, $f^{-1}(V)$ is a neighborhood of a. Therefore we can find $\eta > 0$ such that $B(a, \eta) \subset f^{-1}(V)$. Hence $f(B(a, \eta)) \subset B(b, \varepsilon)$. This proves the continuity of f.[2] □

[2] These results will be proven in homework 2.

Notes

76

Connected spaces

Definition

Let X be a metric space. If the only subsets of X to be both open and closed are X and $\emptyset$, then X is *connected*.

Theorem (Intermediate value theorem)

Let X be a connected metric space, and $a, b \in X$. If $f : X \to \mathbb{R}$ is continuous then it takes all the values between $f(a)$ and $f(b)$.

Sketch of proof (i) Prove that for any continuous function, the image of a connected set is connected;
(ii) Show that the connected subsets of $\mathbb{R}$ are all the intervals;
Then $f(X)$ is an interval of $\mathbb{R}$ which contains both $f(a)$ and $f(b)$, so also any value between them. □

Notes

77

Compact spaces

Definitions

1. A collection U is called an *open cover* of a topological space X if the elements of U are open subsets of X and the union of all elements in U is X.
2. A space X is said to be *compact* if every open cover of X contains a finite subcover.

Example.

- i. Endowed with the standard topology $\mathbb{R}$ is not compact since the family $U = \left\{(-n, n), n \geq 1\right\}$ is an open cover that does not have a finite subcover.
- ii. The empty set, as well as any subset composed of a single element of a metric space, are compact.

Notes

78

Compactness and continuity

We now prove that compactness is preserved through a continuous application.

Theorem

Let X and Y be two metric spaces and $f : X \to Y$ be a continuous function. Then for any compact subset $A \subset X$, $f(A) \subset Y$ is compact.

Proof. Let $(V_i)_{i\in I}$ be a collection of opens from Y covering $f(A)$, and $U_i = f^{-1}(V_i)$. Since f is continuous we know that U_i is an open from X, implying that the collection of $(U_i)_{i\in I}$ covers A. Hence there exists a finite subset J of I such that $(U_i)_{i\in J}$ covers A. Finally note that $(V_i)_{i\in J}$ is finite and covers $f(A)$. □

Notes

79

Extreme value theorem

The extreme value theorem is an important result from analysis, which in fact relies on the topological structure of the set over which the function is defined.

Theorem (Extreme value)

Let X be a metric space, $f : X \to \mathbb{R}$ a continuous map, and A be a non-empty compact of X. Then f is bounded and reaches both its upper and lower bounds.

Sketch of proof (i) Prove that a subset of $\mathbb{R}$ is compact if and only if it is closed and bounded
(ii) Apply theorem 2.78 to conclude that $f(A)$ is closed and bounded in $\mathbb{R}$
(iii) $f(A)$ being closed and non-empty its infinimum and supremum belong to $f(A)$ □

Notes

80

Extreme value theorem

Corollary

If A is a compact subset of a metric space X, then A is bounded.

Proof. Let $a \in X$. For all $x \in X$ we define $f_a(x) = d(a, x)$. The function f is continuous since

$$\begin{aligned} d(f_a(x), f_a(y)) &= d(d(a, x), d(a, y)) \\ &= |d(a, x) - d(a, y)| \leq d(x, y). \end{aligned}$$

Then we can apply the extreme value theorem (2.79) and A is bounded. □

We now present a characterisation of the compacts of $\mathbb{R}^n$, $n \in \mathbb{N}$.

Notes

Bolzano-Weierstrass theorem

Theorem (Bolzano-Weierstrass)

A subset A of $\mathbb{R}^n$, $n \in \mathbb{N}$, is compact if and only if any sequence in A has a sub-sequence converging to an element in A.

Sketch of proof. (i) Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X. Suppose $(x_n)_{n\in\mathbb{N}}$ converges to a limit point l. Then for $a \in X$, there exists $N \in \mathbb{N}^*$ such that $d(l, x_n) < 1$ if $n \geq N$. Thus for $n \in \mathbb{N}$

$$d(a, x_n) \leq \max\{d(a, x_0), \cdots, d(a, x_{N-1}), 1 + d(a, l)\}.$$

This shows that any convergent sequence $(x_n)_{n\in\mathbb{N}}$ is bounded.

(ii) It then suffices to prove that A is compact if and only if any sequence of A has a limit point in A. □

Notes

Heine-Borel theorem

Another important characterisation of the compact of $\mathbb{R}^n$, $n \in \mathbb{N}$, is given by the following theorem.

Theorem (Heine-Borel)

A subset A of $\mathbb{R}^n$, $n \in \mathbb{N}$, is compact if and only if A is closed and bounded.

Sketch of proof
(i) Prove that if A is compact in a metric space X then A is closed in X.
(ii) Apply corollary 2.80 to get the boundedness.

(iii) For the converse we consider A a closed and bounded subset of $\mathbb{R}$. There exist $a_1, b_1, \cdots, a_n, b_n \in \mathbb{R}$ with $a_i \leq b_i$ for all i, and such that $A \subset B = [a_1, b_1] \times \cdots \times [a_n, b_n]$.
(iv) Prove that the compacts of $\mathbb{R}$ are the intervals $[a, b]$, $a, b \in \mathbb{R}$.

Notes

Rolle's theorem

(v) Show that the Cartesian product of two compact is a compact.
(vi) Prove that a closed subset of a compact is compact.

Conclude that A is compact since it is closed and B is compact. □

We now introduce Rolle's theorem that will be needed at a later stage in the course.

Theorem (Rolle's theorem)

Let f be a $C^n[a, b]$ function, for $a, b \in \mathbb{R}$, which has $n + 1$ distinct roots in $[a, b]$. Then there exists $c \in (a, b)$ such that $f^{(n)}(c) = 0$.

Notes

Rolle's theorem

Sketch of proof. The result is proven by induction on n.

(i) Case $n = 1$: apply the extreme value theorem (2.79) to see that f reaches its maximum and minimum in $[a, b]$. Study its left and right limits in those points and observe that they agree.

(ii) Use induction to extend the result to any n; □

Definitions

Let X be a metric space.

1. A sequence $(x_n)_n$ is a *Cauchy sequence* if for any $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_m, x_k) \leq \varepsilon$ when $m, k \geq N$.
2. If any Cauchy sequence on X converges, then X is *complete*.
3. A *Banach space* is a complete normed vector space.

Notes

85

Banach-Steinhaus theorem

Theorem (Banach-Steinhaus)

Let E be a Banach space, F be a normed vector space over $\mathbb{K}$, and H be a non-empty subset of $\mathcal{L}(E, F)$. Then $\sup\{\|u\| : u \in H\}$ is finite if and only if for all $x \in E$, $\sup\{\|u(x)\| : u \in H\}$ is also finite.

Sketch of proof. The first implication is trivial since for $x \in E$, $\|u(x)\| \leq \|u\|\|x\|$. Showing the converse is a bit more complex.

(i) For $x \in E$ and $n \in \mathbb{N}^*$, define $\theta(x) = \sup\{\|u(x)\| : u \in H\}$, which is finite, and $V_n = \{x \in E : \theta(x) > n\}$. Prove that V_n is open in E.

(ii) Show the existence of $p \in \mathbb{N}^*$, $a \in E$, and $r \in \mathbb{R}_+^*$ such that the intersection of V_p with the closed ball $B'(a, r)$ is empty.

(iii) Take $v \in H$ and use θ to upper bound it on $B'(a, r)$. It works as for $x \in E$ with norm 1, $\theta(a)$ and $\theta(a + rx)$ are less than p. □

Notes

86

Stone-Weierstrass theorem

Before introducing our final result for this section we need some more definitions.

Definitions

1. Let $\mathcal{E}$ be a set and $\mathcal{F}(\mathcal{E}, \mathbb{K})$ be the set of the functions defined from $\mathcal{E}$ into $\mathbb{K}$.
   The set $\mathcal{H}$ is a *subalgebra* of $\mathcal{F}(\mathcal{E}, \mathbb{K})$ if
   1. $\mathcal{H}$ is a vector-subspace of $\mathcal{F}(\mathcal{E}, \mathbb{K})$;
   2. The constant applications are in $\mathcal{H}$;
   3. If $f, g \in \mathcal{H}$, then for all x, $f(x)g(x)$ is also in $\mathcal{H}$;
2. The subalgebra $\mathcal{H}$ *separates* $\mathcal{F}(\mathcal{E}, \mathbb{K})$, if for any two distinct elements $x, y \in \mathcal{E}$, there exists $u \in \mathcal{H}$ such that $u(x) \neq u(y)$.
3. Let X be a metric space. A sequence of function $(f_n)_n$, *uniformly converges*, if it converges in $\mathcal{F}(\mathcal{E}, X)$.

Notes

87

Stone-Weierstrass theorem

Remark. Uniform convergence is not to be confused with simple convergence, which only expects $(f_n)_n$ to converge pointwise. Uniform convergence somehow means that "the shape of f_n gets closer and closer from the shape of its limit as n increases".

Theorem (Stone-Weierstrass)

Let X be a compact metric space and $\mathcal{H}$ be a subalgebra that separates $C(X, \mathbb{R})$. Then any element from $C(X, \mathbb{R})$ is the uniform limit of a sequence of elements from $\mathcal{H}$.

Sketch of proof. The proof uses Dini's theorem which provides conditions for a sequence which converges simply to also converge uniformly.

Notes

88

Stone-Weierstrass theorem

Another important lemma gives some conditions on a subset $\mathcal{H}$ of $C(X, \mathbb{R})$ for any element of $C(X, \mathbb{R})$ to be the uniform limit of a sequence of elements of $\mathcal{H}$.

Finally note that the two most important conditions for the theorem to be correct are the compactness of X and $C(X, \mathbb{R})$ being separable. □

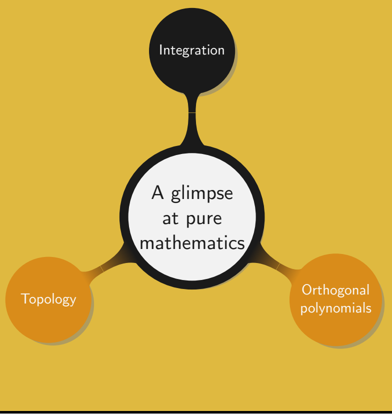
Corollary

Any continuous function defined over $[a, b] \subset \mathbb{R}$ is the uniform limit of a sequence of polynomials.

Proof. This is trivial since closed and bounded real intervals are compact, and the set of polynomials clearly is a subalgebra that separates $C([a, b])$. □

Notes

Chapter organisation



Notes

Riemann sums

Definition

Let $I = [a, b]$ be an interval in $\mathbb{R}$ and $f : I \to \mathbb{R}$, be continuous. Partitioning I using $n + 1$ nodes $a = x_0 < x_1 < \cdots < x_n = b$ the *Riemann sum* of f is defined as

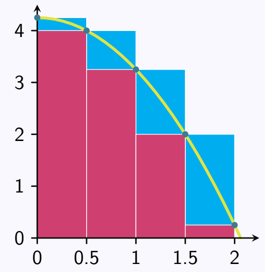
$$\sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}),$$

for $\xi_i \in [x_{i-1}, x_i]$. Each choice of ξ_i defines a different Riemann sum.

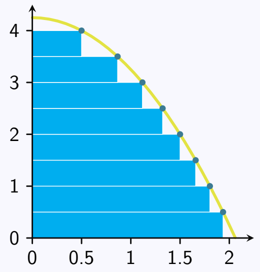
Example. If for all i, $\xi_i = x_{i-1}$, the sum is called *left Riemann sum*. When $\xi_i = x_i$, for any i, this defines the *right Riemann sum*. The *middle Riemann* sum corresponds to taking ξ_i at the middle of each subinterval for all i.

Notes

Riemann vs. Lebesgue integral



Riemann integral



Lebesgue integral

Notes

Riemann vs. Lebesgue integral

Lebesgue explained his integral in simple terms:

> *I have to pay a certain sum, which I have collected in my pocket. I take the bills and coins out of my pocket and give them to the creditor in the order I find them until I have reached the total sum. This is the Riemann integral. But I can proceed differently. After I have taken all the money out of my pocket I order the bills and coins according to identical values and then I pay the several heaps one after the other to the creditor. This is my integral.*
>
> R. Siegmund-Schultze, *Henri Lebesgue*

Notes

93

σ-algebra

Initially Riemann's idea was to split the area under the curve into small rectangles. In other words he explicitly used the notion of "length". As a result some more general or complex sets could not be "measured". Therefore more general tools were needed.

Definition (σ-algebra)

Let Ω be a non empty set and $\mathcal{A} \subset 2^{\Omega}$ be a class of subsets of Ω. Then $\mathcal{A}$ is a σ-algebra if

i. $\Omega \in \mathcal{A}$;
ii. $\mathcal{A}$ is *closed under complements:* if $A \in \mathcal{A}$, then so is $\Omega \setminus A$;
iii. $\mathcal{A}$ is *closed under countable unions:* if $A_0, A_1, \cdots$ are in $\mathcal{A}$, then so is their union;

Notes

94

Measure

Remark. At first glance a topology looks similar to a σ-algebra, but they are two totally different concepts. A major difference is related to the finiteness of the intersection for the topology, while it is countable for a σ-algebra. For instance openness is not preserved in the case

$$\bigcap_{n\in\mathbb{N}} \left(a - \frac{1}{n}, b + \frac{1}{n}\right) \quad = \quad [a, b].$$

Definition

Let Ω be a set and $\mathcal{A}$ be a σ-algebra over Ω. A *measure* is a function $\mu : \mathcal{A} \to \mathbb{R}_+ \cup \{\infty\}$ which is

i. *Non-negative:* for any $A \in \mathcal{A}$, $\mu(A) \geq 0$;
ii. *Null for the empty set:* $\mu(\emptyset) = 0$;
iii. *σ-additive:* for any countable pairwise disjoint set $A_i \in \mathcal{A}$, $\mu(\bigcup_i A_i) = \sum_{i=0}^{\infty} \mu(A_i)$;

Notes

95

Measurable spaces and functions

Definitions

1. Let Ω be a set and $\mathcal{A}$ be a σ-algebra over Ω. Then the pair $(\Omega, \mathcal{A})$ is called *measurable space*.
2. Let $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ be two measurable spaces. A function $f : \Omega_1 \to \Omega_2$ is measurable if for any $A \in \mathcal{A}_2$, the pre-image of A under f is in $\mathcal{A}_1$.
3. A set in a measure space is said to have *σ-finite measure* if it is a countable union of measurable sets with finite measure.
4. In a measurable space, a property is said to *hold almost everywhere* if it holds everywhere but on a subset of measure zero.

We now briefly explain how to construct the Lebesgue integral.

Notes

96

Lebesgue integral

Let $(\Omega, \mathcal{A}, \mu)$ be a measurable space. The integral of a function f, not necessarily continuous, is constructed as follows.

1. For a measurable set $A \in \mathcal{A}$, we have $\int \mathbb{1}_A \, d\mu = \mu(A)$, where $\mathbb{1}_A$ is the *indicator function* of A, i.e. $\mathbb{1}_A(x)$ is 0 if $x \notin A$ and 1 otherwise.
2. Extend indicator functions to *simple functions*, a finite linear combinations of indicator functions. Then by linearity of the integral we have for $a_i \in \mathbb{R}$
$$\int \left(\sum_i a_i \mathbb{1}_{A_i} \, d\mu \right) = \sum_i a_i \mu(A_i).$$
3. A monotonic increasing sequence of positive simple functions converges pointwise, or uniformly if f is bounded, to f. Thus
$$\int_\Omega f \, d\mu = \sup \left\{ \int_\Omega f_n \, d\mu : 0 \leq f_n \leq f \text{ and } f_n \text{ simple} \right\}.$$

Notes

Lebesgue integral

4. Finally for general functions we consider $f = f^+ - f^-$, where $f^+(x) = f(x)$, if $f(x) > 0$ and 0 otherwise, and $f^-(x) = f(x)$ if $f(x) < 0$ and 0 otherwise. Then the Lebesgue integral of f is defined if at least one of $\int_\Omega f^+ \,\mathrm{d}\mu$ and $\int_\Omega f^- \,\mathrm{d}\mu$ is finite. In that case
$$\int_\Omega f \,\mathrm{d}\mu = \int_\Omega f^+ \,\mathrm{d}\mu - \int_\Omega f^- \,\mathrm{d}\mu.$$
5. The function f is said to be *Lebesgue integrable* if $\int_\Omega |f| \,\mathrm{d}\mu$ is finite.

Remark. On figures 2.91 the Riemann integral partitions the domain into subintervals and approximates f using step functions. In Lebesgue integral, the range is partitioned into subintervals I_n, and the measurable sets $A_n = f^{-1}(I_n)$ defined. Thus point 3 states that as n increases, f is more precisely approximated by a simple function, i.e. a linear combination of the characteristic functions of the A_n.

Notes

Fubini's theorem

Theorem (Fubini)

Let (E, X, μ) and (F, Y, ν) be two σ-finite measure spaces, and $(E \times F, X \times Y, \mu\nu)$ be their product measure. If $f(x, y)$ is $X \times Y$ integrable, i.e. it is measurable and $\int_{X\times Y} |f(x, y)| \,\mathrm{d}\mu \,\mathrm{d}\nu$ is finite, then
$$\int_X \left(\int_Y f(x, y) \,\mathrm{d}\nu \right) \mathrm{d}\mu = \int_Y \left(\int_X f(x, y) \,\mathrm{d}\mu \right) \mathrm{d}\nu = \int_{X\times Y} f(x, y) \,\mathrm{d}\mu \,\mathrm{d}\nu.$$

Sketch of proof. (i) Prove that the measure of $X \times Y$ is the product of the measures of X and Y. Use this to prove the theorem in the case of indicator functions.
(ii) Since the spaces have σ-finite measure, prove the theorem for simple functions by using the characteristic function of measurable sets.

Notes

Fubini's theorem

(iii) As f is measurable, consider the case f positive and approximate it by simple measurable functions. Finally prove the result in that case.
(iv) Since the function has a finite integral, write it as the difference of two integrals. Conclude from step (iii). □

Theorem (First mean value theorem)

Let f be a continuous function over $[a, b] \subset \mathbb{R}$ and g is an integrable function which keeps a constant sign on $[a, b]$. Then there exists c in (a, b) such that
$$\int_a^b f(x)g(x) \,\mathrm{d}x = f(c) \int_a^b g(x) \,\mathrm{d}x.$$

Notes

First mean value theorem

Proof. By the extreme value theorem (2.79) we have the existence of m and M, such that $f([a, b]) = [m, M]$.
For g positive we have
$$m \int_a^b g(x) \,\mathrm{d}x \leq \int_a^b f(x)g(x) \,\mathrm{d}x \leq M \int_a^b g(x) \,\mathrm{d}x. \tag{2.1}$$
If $g = 0$, then the results holds. Otherwise we can divide (**??**) by $\int_a^b g(x) \,\mathrm{d}x$, to obtain
$$m \leq \underbrace{\frac{1}{\int_a^b g(x) \,\mathrm{d}x} \int_a^b f(x)g(x) \,\mathrm{d}x}_{A} \leq M.$$

Notes

101

First mean value theorem

Then by the intermediate value theorem (2.76), there exists $c \in (m, M)$ such that $f(c) = A$, i.e.

$$f(c) \int_a^b g(x)\,\mathrm{d}x = \int_a^b f(x)g(x)\,\mathrm{d}x.$$

Finally observe that if g is negative, then the role of m and M in (**??**) is reversed but the rest of the proof remains valid. □

Theorem (Taylor's theorem)

Given a function f such that $f^{(n)}$ is absolutely continuous on the compact $[a, x] \subset \mathbb{R}$,

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n\,\mathrm{d}t.$$

Notes

102

Taylor's theorem

Sketch of proof. As a general remark notice that as $f^{(n)}$ is absolutely continuous on the compact $[a, x] \subset \mathbb{R}$, both its differential and integral exist.

(i) First apply the fundamental theorem of calculus to get

$$f(x) = f(a) + \int_a^x f'(t)\,\mathrm{d}t.$$

(ii) Prove the result by induction, recursively integrating by parts.

□

Definition (Weight function)

A weight function is a continuous function over (a, b) such that

- i For all $t \in (a, b)$, $w(t)$ is strictly positive;
- ii The integral $\int_a^b w(t)\,\mathrm{d}t$ is finite;

Notes

103

Weight function

Example. Let $w(t) = \frac{1}{\sqrt{1-t^2}}$ for $t \in (-1, 1)$. While it is clear that w is strictly positive on $(-1, 1)$, the second condition is also quickly checked by applying the change of variable $x = -\arccos t$.

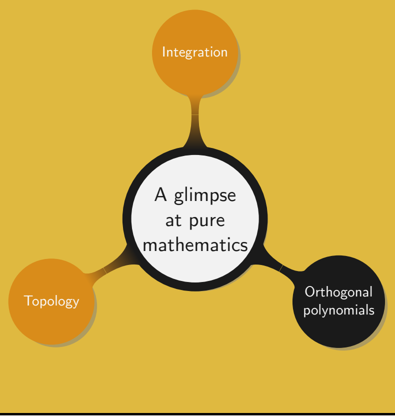
$$\int_{-1}^{1} \frac{1}{\sqrt{1-t^2}}\,\mathrm{d}t = \int_{-\pi}^{0} \mathrm{d}x = \pi < \infty.$$

Remark. A weight functions is in fact a strictly positive measurable function. They are especially useful on intervals where the product of two polynomials is not integrable. In such a case the product can be weighted such as to get convergence in the integral. In particular the weight is expected to decrease faster than the product of polynomials grows.

Notes

104

Chapter organisation



Notes

105

Projection on an inner product space

Theorem

Let F be a finite dimensional vector subspace of an inner product space E. Then for all $e \in E$, there exists a unique element $u \in F$, called the *projection of* e *on* F and denoted $P_F(e)$, such that

$$\|e - u\| = \min_{f \in F} \|e - f\|.$$

Moreover for any $f \in F$, $e - P_F(e)$ and f are orthogonal.

Proof. From lemma 1.36 it is easy to derive that F is isomorphic to $\mathbb{R}^d$, for some $d \in \mathbb{N}$. Let v be the isomorphism from $\mathbb{R}^d$ into F, and note that v is continuous. Indeed, since $\mathbb{R}^d$ is of finite dimension it has a finite basis $(e_1, \cdots, e_d)$ and by linearity for $v \in (\mathbb{R}^d, \|\cdot\|_1)$

$$\|v(x)\| = \|\sum_{i=1}^{d} x_i v(e_i)\| \leq \sum_{i=1}^{d} |x_i| \|v(e_i)\| \leq \max_{1 \leq i \leq d} \|v(e_i)\| \|x\|_1.$$

Notes

106

Projection on an inner product space

As this shows that v is continuous, we can apply theorem 1.47 to obtain that $\|v\|$ is finite. Using a similar reasoning v^{-1} is also continuous and has finite norm.

Let $(x_n)_{n\in\mathbb{N}}$ be a bounded sequence in F, i.e. there exists a closed ball $B'(a, r)$, $a \in F, r \in \mathbb{R}_+^*$, that contains all the elements of the sequence. The image of this ball by v^{-1} is a closed ball from $\mathbb{R}^d$, so it is compact by Heine-Borel (2.82). As v is continuous, $B'(a, r)$ is compact (theorem 2.78) and by Bolzano-Weierstrass (2.81) we can extract a convergent subsequence of $(x_n)_{n\in\mathbb{N}}$.

Our goal is now to use this strategy in order to prove the existence of the unique element u.

Let $e \in E$. We are interested in finding an $f \in F$ that minimizes $\|e - f\|$. We call this infimum m and consider a sequence $(f_n)_{n\in\mathbb{N}}$ from F such that $\|e - f_n\|$ tends to m.

Notes

107

Projection on an inner product space

First we have to prove that our sequence is bounded, then we will be able to apply Bolzano-Weierstrass (2.81).
By the minimality of m, $m < m + 1/n$, for any $n \in \mathbb{N}^*$, and there exists at least one f_n such that $m \leq \|e - f_n\| < m + 1/n$. Finally as

$$\big|\|f_n\| - \|e\|\big| \leq \|f_n - e\|$$

we see that $\|f_n\| \leq \|f_n - e\| + \|e\|$, which basically means that $(f_n)_n$ is bounded.

Thus we are allowed to apply Bolzano-Weierstrass (2.81) to extract a subsequence $(f_{n_k})_{k\geq 0}$ that converges to $\widetilde{f}$.
Then we have

$$\left|\|f_{n_k} - e\| - \|e - \widetilde{f}\|\right| \leq \left\|f_{n_k} - e + e - \widetilde{f}\right\|.$$

Notes

108

Projection on an inner product space

This means that when k tends to infinity, $\|f_{n_k} - e\|$ tends to $\|\widetilde{f} - e\|$. But as by construction $\|f_{n_k} - e\|$ tends to m, we have by uniqueness of the limit, $m = \|\widetilde{f} - e\|$, and $\widetilde{f} = f$.

We now prove the uniqueness. Let f_1, f_2 be two elements from F such that $m = \|e - f_1\| = \|e - f_2\|$. By the parallelogram law (prop. 1.41)

$$\left\|\frac{e - f_1}{2} + \frac{e - f_2}{2}\right\|^2 + \left\|\frac{e - f_1}{2} - \frac{e - f_2}{2}\right\|^2 = 2\left(\left\|\frac{e - f_1}{2}\right\|^2 + \left\|\frac{e - f_2}{2}\right\|^2\right).$$

Rearranging the terms and using m yields

$$\left\|\frac{f_1 - f_2}{2}\right\|^2 = m^2 - \left\|e - \frac{f_1 + f_2}{2}\right\|^2. \qquad (2.2)$$

Notes

Projection on an inner product space

Since $(f_1 + f_2)/2$ is in F, we apply the minimality of m to (**??**) and obtain the unicity as

$$\left\| \frac{f_1 - f_2}{2} \right\|^2 \le 0.$$

As we have proven the existence and uniqueness of f, it makes sense to define the application

$$\begin{aligned} P_F : E &\longrightarrow F \\ e &\longmapsto P_F(e), \end{aligned}$$

such that $\|e - P_F(e)\| = \min_{f \in F} \|e - f\|$.
The only thing left to prove is that for any $f \in F$, $e - P_F(e)$ and f are orthogonal, i.e. for any $f \in F$, $\langle e - P_F(e), f\rangle = 0$.

Notes

Projection on an inner product space

First we assume that $\|e - P_F(e)\|$ is minimum and show the equality. Then for all $f \in F$ and $\lambda \in \mathbb{R}$,

$$\begin{aligned} \|e - P_F(e)\|^2 &\le \|e - P_F(e) - \lambda f\|^2 \\ &\le \|e - P_F(e)\|^2 + \lambda^2\|f\|^2 - 2\lambda\langle e - P_F(e), f\rangle. \end{aligned}$$

And $\lambda^2\|f\|^2 - 2\lambda\langle e - P_F(e), f\rangle$ is positive for all $\lambda \in \mathbb{R}$. Thus

- for $\lambda > 0$, $\lambda\|f\|^2 \ge 2\langle e - P_F(e), f\rangle$, and
- for $\lambda < 0$, $\lambda\|f\|^2 \le 2\langle e - P_F(e), f\rangle$.

Since this is also true for any λ tending to 0 we conclude that $\langle e - P_F(e), f\rangle$ is 0.

Notes

Projection on an inner product space

As a final step we need to check the converse, i.e. making sure $\|e - P_F(e)\|$ is indeed minimal.

Denoting $P_F(e)$ by u we want to show that $\|e - u\|$ is minimal. Note that for all $f \in F$

$$\|e - u - f\|^2 = -2\langle e - u, f\rangle + \|e - u\|^2 + \|f\|^2.$$

So it is clear that for any $f \in F$, $\|e - u - f\| \ge \|e - u\|$. Observing that $u + f \in F$ since u and f are both in the vector space F it means that for any $f \in F$

$$\|e - f\| \ge \|e - u\|.$$

This proves the minimality of $\|e - u\|$, and concludes the proof of the theorem.

□

Notes

Projection on an inner product space

Example. Let E be $C[a, b]$ and w be a weight function over E. For two functions f and g in E, we define their inner-product in E by

$$\langle f, g\rangle = \int_a^b f(x)g(x)w(x)\,\mathrm{d}x.$$

This integral is meaningful since

$$\int_a^b |f(x)g(x)|w(x)\,\mathrm{d}x \le \sup_{x \in [a,b]} |fg|(x) \int_a^b w(x)\,\mathrm{d}x < \infty.$$

Moreover this is an inner-product as over $[a, b] \subset \mathbb{R}$ we have $\langle f, g\rangle = \langle g, f\rangle$. Then as the integral is a linear operation the only thing left to check is the last point from definition 1.37. Again this is clear as by definition of a weight function, $w(x)$ is strictly positive over (a, b).

Notes

113

Projection on an inner product space

Let F be the set of all the polynomials defined over $[a, b]$, with degree less than $n \in \mathbb{N}$. Both E and F are vector spaces, F has finite dimension $n+1$, and $F \subset E$. If given a function $e \in E$, then by theorem 2.105 we know the existence of a unique polynomial P_n such that

$$\int_a^b (f(x) - P_n(x))^2 w(x)\,\mathrm{d}x = \min_{Q \in F} \int_a^b (f - Q)^2(x) w(x)\,\mathrm{d}x.$$

Definition

An *orthogonal basis* of an inner product space F of dimension $d \in \mathbb{N}$ is a basis of vectors $\varphi_1, \cdots, \varphi_d$ such that

$$\langle \varphi_i, \varphi_j \rangle = \begin{cases} 0 & \text{if } i \neq j \\ a & \text{if } i = j, \end{cases}$$

for some a. If a is always 1 then it is called an *orthonormal basis*.

Notes

114

Projection in an orthonormal basis

Theorem

Let $(\varphi_1, \cdots, \varphi_d)$ be an orthonormal basis for a vector subspace F of an inner product space E. Then for all $e \in E$,

$$P_F(e) = \sum_{j=1}^{d} \langle e, \varphi_j \rangle \varphi_j.$$

The $\langle e, \varphi_j \rangle$ are called the *Fourier coefficients* of $P_F(e)$.

Proof. As by definition $P_F(e)$ is in F, there exist $a_1, \cdots, a_d$ such that $P_F(e) = \sum_{j=1}^{d} a_j \varphi_j$. So we only need to prove that $a_j = \langle e, \varphi_j \rangle$ for all $1 \leq j \leq d$.
For $i \in \{1, \cdots, d\}$ we have

$$\langle P_F(e), \varphi_i \rangle = \sum_{j=1}^{d} a_j \langle \varphi_j, \varphi_i \rangle = a_i.$$

Notes

115

Projection in a orthonormal basis

By the characterisation of the projection (theorem 2.105), we know that $\langle e - P_F(e), \varphi_i \rangle = 0$. Hence we obtain

$$\langle e, \varphi_i \rangle = \langle P_F(e), \varphi_i \rangle = a_i.$$

□

Theorem (Gram-Schmidt)

Let F be a finite dimension inner-product space, with a known basis $(e_1, \cdots, e_d)$, where d is the dimension of F. Then one can construct an orthogonal basis over F.

Notes

116

Gram-Schmidt process

Proof. Let F_k be $\operatorname{span}(e_1, \cdots, e_k)$. For $k = 1$ we simply take $\varphi_1 = e_1$.
Assume $(\varphi_1, \cdots, \varphi_k)$ has been constructed and $F_k = \operatorname{span}(\varphi_1, \cdots, \varphi_k)$, with $\langle \varphi_i, \varphi_j \rangle = 0$, for $i \neq j$, where i and j are less than k. Then it suffices to take $\varphi_{k+1} = e_{k+1} - P_F(e_{k+1})$.
Clearly φ_{k+1} is in $F_{k+1} = \operatorname{span}(\varphi_1, \cdots, \varphi_k, e_{k+1})$, and by construction is orthogonal to all the elements of F_k. □

Remark. An orthonormal basis is trivially obtained by dividing each vector by its norm.

Assembling all the previous results together we are now armed to construct orthogonal polynomials.

Notes

117

Construction of orthogonal polynomials

Let $E = C[a,b]$. From example 2.112, we know this is an inner product space. Moreover F, the set of monic polynomials defined over $[a,b]$, with degree less than n, is of finite dimension $n+1$ and included in E. The canonical base for F is $(1, x, \cdots, x^n)$. If we apply the Gram-Schmidt process we can construct a family of orthogonal polynomials $P_i \in F$, $0 \le i \le n$, such that

$$\begin{cases} \mathrm{span}(P_0, \cdots, P_n) = \mathrm{span}(e_0, \cdots, e_n) \\ \langle P_i, P_j \rangle = \displaystyle\int_a^b P_i(x) P_j(x) w(x)\, \mathrm{d}x = 0, \quad \text{if } i \neq j \\ P_k(x) = x^k + Q_{k-1}(x), \quad \text{for } Q_{k-1} \in \mathbb{R}_{k-1}[x] \end{cases} \tag{2.3}$$

Definition

The polynomials P_i, $0 \le i \le n$, such that (**??**) is verified are called *orthogonal polynomials*.

Notes

118

Orthogonal polynomials

Remark. In the previous definition the polynomials are expected to be monic, in fact defining a sequence of orthonormal polynomials.

Proposition

Let P_n be the $(n+1)$st orthogonal polynomial associated to a weight function w.

1. Orthogonal polynomials associated to a weight w are unique;
2. P_n has degree n;
3. P_n is orthogonal to any polynomial $Q \in \mathbb{R}_{n-1}[x]$, for $n \ge 1$;
4. For any $n \ge 1$ there exist $\lambda_n \in \mathbb{R}$ and $\mu_{n-1} > 0$ such that
$$P_{n+1}(x) = (x - \lambda_n) P_n(x) - \mu_{n-1} P_{n-1}(x);$$
5. P_n has n distinct real roots in (a,b);

Notes

119

Orthogonal polynomials

Proof. 1. Let P_n and $\widetilde{P}_n$ be two orthogonal polynomials of degree n. Since both are monic their difference $P = P_n - \widetilde{P}_n$ is of degree at most $n-1$. Therefore by definition we have $\langle P_n, P \rangle = \langle \widetilde{P}_n, P \rangle = 0$. This implies

$$0 = \langle P_n, P \rangle - \langle \widetilde{P}_n, P \rangle = \langle P_n - \widetilde{P}_n, P \rangle = \langle P, P \rangle.$$

Since $\langle \cdot \rangle$ is a scalar product, it means that $P = 0$ and $P_n = \widetilde{P}_n$.

2. This is clear since $P_n(x) = x^n + Q(x)$ with $\deg Q < n$.

3. Since Q is of degree at most $n-1$, it can be expressed in the basis of the P_i, $0 \le i \le n-1$. This yields the result.

Notes

120

Orthogonal polynomials

4. We consider $P(x) = P_{n+1}(x) - xP_n(x)$. It is a polynomial of degree at most n, so there are some c_k such that P can be written

$$P(x) = \sum_{k=0}^{n} c_k P_k(x).$$

We now prove that for any $k \le n-2$, $c_k = 0$. For this we calculate

$$\langle P, P_k \rangle = \sum_{j=0}^{n} c_j \langle P_j, P_k \rangle = c_k \| P_k \|^2. \tag{2.4}$$

But we also have $\langle P, P_k \rangle = \langle P_{n+1}, P_k \rangle - \langle xP_n, P_k \rangle$. As $k < n+1$, $\langle P_{n+1}, P_k \rangle = 0$. All the polynomial being defined over the reals $\langle xP_n, P_k \rangle = \langle P_n, xP_k \rangle$, which is 0 since xP_k has degree $k+1$, smaller than $n+1$.

Notes

Orthogonal polynomials

Hence from eq. (**??**) we have $c_k\|P_k\|^2 = 0$, implying $c_k = 0$. In the end we have obtained

$$P_{n+1}(x) - xP_n(x) = c_nP_n(x) + c_{n-1}P_{n-1}(x).$$

The only thing left to prove is that $c_{n-1} = -\mu_{n-1}$ is strictly positive.
On the one hand looking at $P_{n+1} - xP_n$ as $\lambda_n P_n + \mu_{n-1}P_{n-1}$ yields $\langle P_{n+1} - xP_n, P_{n-1}\rangle = -\mu_{n-1}\|P_{n-1}\|^2$. On the other hand we have

$$\begin{aligned}\langle P_{n+1} - xP_n, P_{n-1}\rangle &= \langle P_{n+1}, P_{n-1}\rangle - \langle xP_n, P_{n-1}\rangle\\ &= -\langle xP_n, P_{n-1}\rangle\\ &= -\langle P_n, xP_{n-1} - P_n + P_n\rangle\\ &= -\langle P_n, xP_{n-1} - P_n\rangle - \|P_n\|^2 = -\|P_n\|^2.\end{aligned}$$

Hence $-\mu_{n-1}\|P_{n-1}\|^2 = -\|P_n\|^2$ and $\mu_{n-1} > 0$.

Notes

Orthogonal polynomials

5. Assume that P_n has no root in (a, b). Then P_n has a constant sign in (a, b). But as $\int_a^b 1 \cdot P_n(x)w(x)\,\mathrm{d}x = 0$ it means $P_n = 0$.↯

As P_n has between one and n roots, let $x_0 < x_1 < \cdots < x_r$ be all its roots in (a, b), each one having multiplicity α_i, $0 \le i \le r$. Then

$$P_n(x) = \underbrace{\prod_{j=0}^{r}(x - x_j)^{\alpha_j}}_{P(x)} Q(x),$$

where $Q(x)$ keeps a constant sign on (a, b).

Thus $P(x)P_n(x) = P^2(x)Q(x)$ also keeps a constant sign on (a, b). But as $\deg P < n$, $\langle P, P_n\rangle = \int_a^b P(x)P_n(x)w(x)\,\mathrm{d}x = 0$, which means that $P(x)P_n(x)w(x) = 0$. Since $w(x)$ is a weight function it is strictly positive over (a, b), and $P(x)P_n(x) = 0$ for any $x \in (a, b)$.↯

This shows that $P_n(x) = \prod_{j=0}^{r}(x - x_j)^{\alpha_j}$.

Notes

Orthogonal polynomials

Now lets assume that some of the α_i are strictly larger than 1. Then reordering the roots we can write

$$P_n(x) = \underbrace{\prod_{j=0}^{m}(x - x_j)^{\alpha_j - 2} \cdot \prod_{j=m}^{r}(x - x_j)^{\alpha_j}}_{Q(x)} \cdot \prod_{j=0}^{m}(x - x_j)^2.$$

Since Q has degree strictly less than n, it is orthogonal to P_n and

$$\int_a^b Q^2(x)\prod_{j=0}^{m}(x - x_j)^2 w(x)\,\mathrm{d}x = \int_a^b Q(x)P_n(x)w(x)\,\mathrm{d}x = 0.$$

Hence $Q^2(x)\prod_{j=0}^{m}(x - x_j)^2 = 0$ and $P_n(x) = 0$.↯

As we have proven that α_j, $0 \le j \le r$, is never larger than 1, it means P_n has n distinct real roots in (a, b). □

Notes

Chebyshev polynomials

We now study Chebyshev polynomials, a family of orthogonal polynomials.

Proposition

For $x \in [-1, 1]$, the functions $T_n(x) = \cos(n \arccos x)$ is called Chebyshev polynomials and verify the following properties.

1. For all $x \in [-1, 1]$, $T_{n+1}(x) + T_{n-1}(x) = 2xT_n(x)$;
2. For all $n \in \mathbb{N}$, T_n is a polynomial over $[-1, 1]$;
3. Over $(-1, 1)$, the $(T_n)_{n\in\mathbb{N}}$ define a collection of orthogonal polynomials, with respect to the weight function $w(x) = \frac{1}{\sqrt{1-x^2}}$.
4. The roots of T_n are given by $\cos\left(\frac{2k-1}{2n}\pi\right)$, $1 \le k \le n$;

Proof. 1. First we calculate $T_0(x) = 1$, $T_1(x) = \cos(\arccos x) = x$, and $T_2(x) = \cos(2\arccos x) = -1 + 2\cos^2(\arccos x) = -1 + 2x^2$.
Then recall that $\cos[(n + 1)\theta] + \cos[(n - 1)\theta] = 2\cos(\theta)\cos(n\theta)$.

Notes

125

Chebyshev polynomials

As a result we directly obtain for any $n \geq 1$

$$\begin{aligned} T_{n+1}(x) + T_{n-1}(x) &= \cos[(n+1)\arccos x] + \cos[(n-1)\arccos x] \\ &= 2\cos(\arccos x)\cos(n\arccos x) \\ &= 2x\cos(n\arccos x) = 2xT_n(x). \end{aligned}$$

2. Using the recurrence relation from 1. it is clear that the T_n are polynomials, since T_0 and T_1 are polynomials. Note that it is also easy to show by induction that the coefficient of x^n is 2^{n-1}.

3. For the weight function $w(x) = \frac{1}{\sqrt{1-x^2}}$ over $(-1, 1)$, and $n \neq m$

$$\begin{aligned} \langle T_n, T_m \rangle &= \int_{-1}^{1} \cos(n\arccos x)\cos(m\arccos x)\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x \\ &= \frac{1}{2}\int_0^{\pi} \cos[(n+m)x] + \cos[(n-m)x]\,\mathrm{d}x = 0. \end{aligned}$$

Notes

126

Chebyshev polynomials

Note that Chebyshev polynomials are not orthonormal as they are not monic. It however suffices to define $\overline{T}_n(x) = 1/2^{n-1}T_n(x)$ to obtain a collection of monic orthogonal polynomials.

4. Let x be root of $T_n(x)$, $n \in \mathbb{N}$. Thus $\cos(n\arccos x) = 0$. This means that over $[0, \pi]$ we have $\arccos x = -\frac{\pi}{2n} + \frac{k\pi}{n}$, $k \in \mathbb{Z}$.

So x is a root if and only if $x = \cos\left(\frac{2k-1}{2n}\pi\right)$, for $1 \leq k \leq n$. □

Notes

127

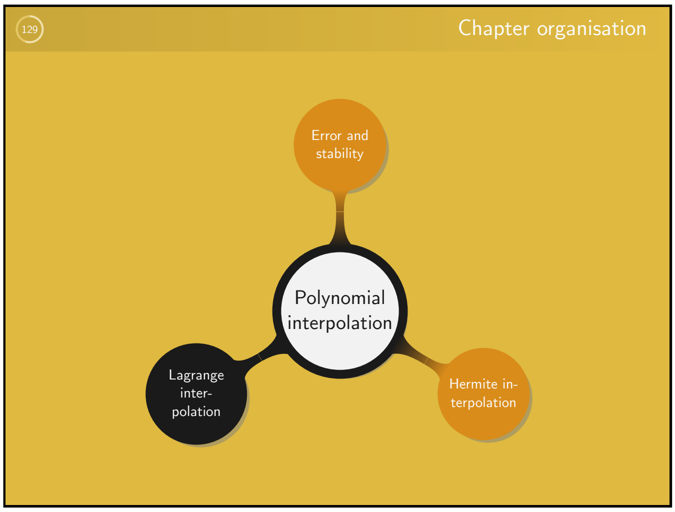
Key points

- Why is topology fundamental in this course?
- What is the difference between a topology and a σ-algebra?
- Explain the Riemann and Lebesgue integrals
- What is a weight function?
- What are orthogonal polynomials?

Notes

3. Polynomial interpolation

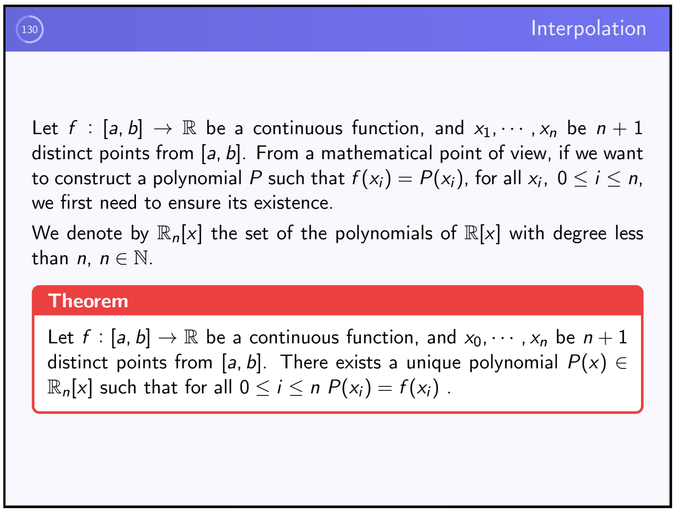
Notes



129

Chapter organisation

Notes



130

Interpolation

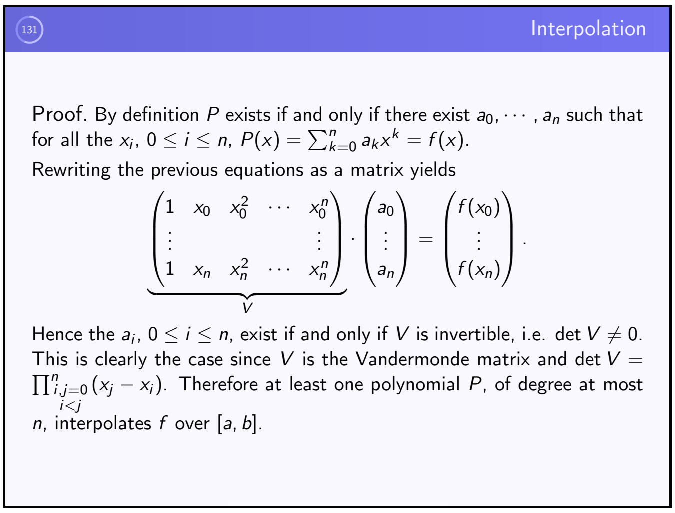
Let $f : [a, b] \to \mathbb{R}$ be a continuous function, and $x_1, \cdots, x_n$ be $n + 1$ distinct points from $[a, b]$. From a mathematical point of view, if we want to construct a polynomial P such that $f(x_i) = P(x_i)$, for all x_i, $0 \le i \le n$, we first need to ensure its existence.

We denote by $\mathbb{R}_n[x]$ the set of the polynomials of $\mathbb{R}[x]$ with degree less than n, $n \in \mathbb{N}$.

Theorem

Let $f : [a, b] \to \mathbb{R}$ be a continuous function, and $x_0, \cdots, x_n$ be $n + 1$ distinct points from $[a, b]$. There exists a unique polynomial $P(x) \in \mathbb{R}_n[x]$ such that for all $0 \le i \le n$ $P(x_i) = f(x_i)$.

Notes



131

Interpolation

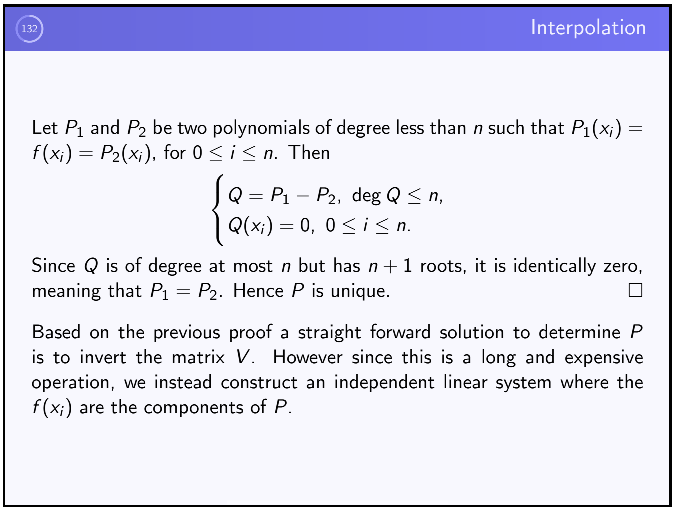
Proof. By definition P exists if and only if there exist $a_0, \cdots, a_n$ such that for all the x_i, $0 \le i \le n$, $P(x) = \sum_{k=0}^{n} a_k x^k = f(x)$.

Rewriting the previous equations as a matrix yields

$$\underbrace{\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}}_{V} \cdot \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f(x_0) \\ \vdots \\ f(x_n) \end{pmatrix}.$$

Hence the a_i, $0 \le i \le n$, exist if and only if V is invertible, i.e. $\det V \neq 0$. This is clearly the case since V is the Vandermonde matrix and $\det V = \prod_{\substack{i,j=0 \\ i<j}}^{n} (x_j - x_i)$. Therefore at least one polynomial P, of degree at most n, interpolates f over $[a, b]$.

Notes



132

Interpolation

Let P_1 and P_2 be two polynomials of degree less than n such that $P_1(x_i) = f(x_i) = P_2(x_i)$, for $0 \le i \le n$. Then

$$\begin{cases} Q = P_1 - P_2, \ \deg Q \le n, \\ Q(x_i) = 0, \ 0 \le i \le n. \end{cases}$$

Since Q is of degree at most n but has $n + 1$ roots, it is identically zero, meaning that $P_1 = P_2$. Hence P is unique. □

Based on the previous proof a straight forward solution to determine P is to invert the matrix V. However since this is a long and expensive operation, we instead construct an independent linear system where the $f(x_i)$ are the components of P.

Notes

133

Lagrange interpolation

In other words we write the polynomial P as

$$P(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x), \text{ where } \ell_i \in \mathbb{R}_n[X] \text{ and } \ell_i(x_j) = \begin{cases} 0 \text{ if } j \neq i \\ 1 \text{ if } j = i \end{cases}.$$

By construction, $\ell_i(x_j) = 0$ if $j \neq i$, meaning that x_j is a root of ℓ_i for $j = 0, \cdots, i-1, i+1, \cdots, n$. Therefore ℓ_i is divisible by $\prod_{\substack{j=0 \\ j\neq i}}^{n}(x - x_j)$ and there exists c_i such that $\ell_i(x) = c_i \prod_{\substack{j=0 \\ j\neq i}}^{n}(x - x_j)$. Since we also know $\ell_i(x_i) = 1$ we can deduce $c_i = \dfrac{1}{\prod_{\substack{j=0 \\ j\neq i}}^{n}(x_i - x_j)}$. Hence, $\ell_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \dfrac{x - x_j}{x_i - x_j}$, for $0 \leq i \leq n$.

Notes

134

Lagrange interpolation

Definition (Lagrange polynomials)

Let $f : [a, b] \to \mathbb{R}$ be a continuous function, and $x_0, \cdots, x_n$ be $n+1$ distinct points from $[a, b]$. The polynomial

$$L_n(f) = \sum_{i=0}^{n} f(x_i)\ell_i(x), \text{ with } \ell_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \quad \text{for } 0 \leq i \leq n,$$

is called the *Lagrange form of the interpolation polynomial* and the $\ell_i(x)$, *Lagrange polynomials*.

Example. Let $f(t_0), \cdots, f(t_n)$ represent $n+1$ measurements at a regular interval, what is the value of f at $(t_{i+1} - t_i)/2$?

Notes

135

Lagrange interpolation

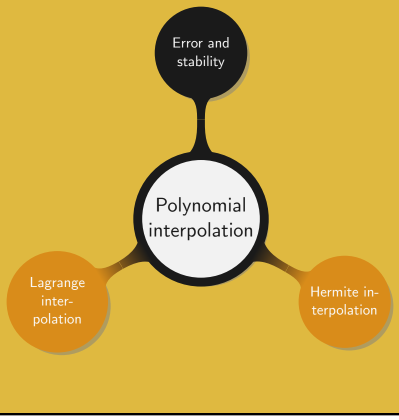
Algorithm. (*Lagrange interpolation*)

```
Input  : n+1 pairs $(x_0, f(x_0)), \cdots, (x_n, f(x_n))$, and $c \in [\min_i x_i, \max_i x_i]$
Output: $f(c)$
$res \leftarrow 0$;
for $i \leftarrow 0$ to $n$ do
    $\ell_i \leftarrow f(x_i)$;
    for $j \leftarrow 0$ to $n$ do
        if $j \neq i$ then $\ell_i \leftarrow \ell_i \frac{c - x_j}{x_i - x_j}$;
    end for
    $res \leftarrow res + \ell_i$
end for
return $res$
```

Notes

136

Chapter organisation



Notes

From theorem 3.130 we know that a unique polynomial, depending on the x_i, interpolates f. We now want to study the error of interpolation, i.e. the difference between $f(x)$ and $L_n(f)$ when $x \neq x_i$.

Theorem

Let $[a, b]$ be a real interval, $(x_i)_{i=0,\cdots,n}$ be $n+1$ distinct points of $[a, b]$, and f be in $C^{n+1}[a, b]$. If $L_n(f)$ is the Lagrange interpolation polynomial of f at the points x_i, then for all x in $[a, b]$ there exists $\xi_x \in [\min(x, \min_{0\le i\le n} x_i), \max(x, \max_{0\le i\le n} x_i)] \subset [a, b]$ such that

$$Ef(x) = f(x) - L_n(f)(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{j=0}^{n}(x - x_j).$$

Notes

Proof. Let $x \in [a, b]$. For the sake of clarity we denote $L_n(f)$ by P.

If $x = x_j$, then $f(x_j) - P(x_j) = 0$ while $\frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{n}(x_j - x_i)$ is also 0 for any ξ_x. So the result holds.

Therefore we now assume $x \notin \{x_0, \cdots, x_n\}$ and define the polynomial $\Pi_n(t) = \prod_{j=0}^{n}(t - x_j)$ for any $t \in [a, b]$. We then construct the polynomial

$$Q(t) = P(t) + \frac{f(x) - P(x)}{\Pi_n(x)}\Pi_n(t) \qquad (3.1)$$

and the function $g(t) = f(t) - Q(t)$.

We now investigate some properties of g. First note that $g \in C^{n+1}[a, b]$ for $f \in C^{n+1}[a, b]$ and $Q \in C^\infty[a, b]$.

Notes

Then observe that $g(x_i) = 0$ since $f(x_i) = P(x_i)$ and $\Pi_n(x_i) = 0$.

Moreover as $g(x) = f(x) - P(x) - \frac{f(x)-P(x)}{\Pi_n(x)}\Pi_n(x) = 0$, we can conclude that g has $n+2$ distinct roots in $[a, b]$.

As a result we know that $g \in C^{n+1}[a, b]$ and has $n+2$ distinct roots in $[a, b]$. Therefore Rolle's theorem (2.83) can be applied, and there exists ξ_x in the smallest interval containing the roots of g, such that $g^{(n+1)}(\xi_x) = 0$. Clearly the smallest interval containing the roots of $g = f - Q$ is $[\min(x, \min_{0\le i\le n} x_i), \max(x, \max_{0\le i\le n} x_i)]$ and

$$f^{(n+1)}(\xi_x) - Q^{(n+1)}(\xi_x) = 0.$$

Then from (**??**), it is easy to get $Q^{(n+1)}(t) = \frac{f(x)-P(x)}{\Pi_n(x)}\Pi_n^{(n+1)}(t)$.

Notes

Finally as $\Pi_n^{(n+1)}(t) = (n+1)!$ and $Q^{(n+1)}(t) = \frac{f(x)-P(x)}{\Pi_n(x)}(n+1)!$, we obtain

$$Ef(x) = f(x) - P(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{j=0}^{n}(x - x_j).$$

□

Corollary

Using the previous notations and setting $M_n = \max_{t\in[a,b]}\left|f^{(n+1)}(t)\right|$,

$$\left|f(x) - L_n(f)\right| \le \frac{M_n}{(n+1)!}\left|\Pi_n(x)\right| \le \frac{M_n}{(n+1)!}(b-a)^{n+1}.$$

Notes

141

Choosing the nodes

As $|\Pi_n(x)| \le \max_{t\in[a,b]} |\Pi_n(t)|$ the previous corollary (3.140) yields

$$|f(x) - L_n(f)| \le \frac{M_n}{(n+1)!} \max_{t\in[a,b]} |\Pi_n(t)|.$$

In other words the error decreases as $\max_{t\in[a,b]} |\Pi_n(t)|$ decreases, meaning that the error depends on the choice of the nodes x_i. Therefore we now focus on how to choose the x_i in order to minimize $\max_{t\in[a,b]}$.

Observing that Π_n is a monic polynomial of degree $n+1$, we want to choose the x_i such that

$$\max_{x\in[a,b]} \left| \prod_{j=0}^{n} (x - x_j) \right| \le \max_{x\in[a,b]} \left| q(x) \right|, \quad \forall q \in E_{n+1},$$

where $E_{n+1} = \left\{ P \in \mathbb{R}_{n+1}[x], P \text{ is monic} \right\}$.

Notes

142

Choosing the nodes

For the sake of simplicity we assume $a = -1$ and $b = 1$ and then apply the bijective change of variable

$$\begin{aligned} \varphi : [-1,1] &\longrightarrow [a,b] \\ X &\longmapsto \frac{b-a}{2} X + \frac{b+a}{2} \end{aligned} \tag{3.2}$$

to recover the original case.

Theorem

For any polynomial q in E_n, $\max_{t\in[-1,1]} |q(t)| \ge 2^{1-n}$.

Proof. Assume there exists $r \in E_n$ such that $\max_{t\in[-1,1]} |r(t)| < 2^{1-n}$.

Notes

143

Choosing the nodes

From proposition 2.124 we know that the nth Chebyshev polynomial of the first kind has leading coefficient 2^{n-1}, and extrema -1 and 1 at the points x_i', $0 \le i \le n$.

Therefore $\overline{T}_n(x) = 2^{1-n} T_n(x)$ is in E_n, and $m(t) = r(t) - \overline{T}_n(t)$ is a polynomial of degree at most $n-1$ over $[-1,1]$. Moreover at the x_i', m evaluates as $m(x_i') = r(x_i') - (-1)^k 2^{1-n}$.

As m changes sign n times, the intermediate values theorem (2.76) implies that m has n distinct roots. This is impossible since $\deg m < n$.↯ □

Now note that as $\overline{T}_n(x) = 2^{1-n} T_n(x)$, we have

$$\max_{x\in[-1,1]} |\overline{T}_n(x)| = 2^{1-n} \max_{x\in[-1,1]} |T_n(x)| = 2^{1-n}.$$

Notes

144

Choosing the nodes

Together with theorem 3.142 it means that choosing the x_i as the roots of T_n minimizes the error. In this case it is given by

$$\left| f(x) - L_n(f) \right| \le \frac{1}{(n+1)!} \cdot \max_{x\in[-1,1]} \left| \prod_{k=0}^{n} (x - x_k) \right| \cdot \max_{x\in[-1,1]} \left| f^{(n+1)}(x) \right|.$$

But as $\prod_{k=0}^{n} (x - x_k)$ is exactly $\overline{T}_{n+1}(x)$ we get from corollary 3.140

$$\left| f(x) - L_n(f) \right| \le \frac{M_n}{2^n (n+1)!}.$$

All the previous discussion being only valid over $[-1,1]$ we now use the map $\varphi(X)$ (**??**) in order to determine the best nodes over any interval $[a,b]$ of $\mathbb{R}$.

Notes

Choosing the nodes

From proposition 2.124 the $(n+1)$th Chebyshev polynomial of the first kind has roots $x_k = \cos\left(\frac{2k+1}{2n+2}\pi\right)$, $0 \le k \le n$. Hence after applying the map, the roots are given by

$$x_k = \frac{b+a}{2} + \frac{b-a}{2}\cos\left(\frac{2k+1}{2n+2}\pi\right), \quad \forall 0 \le k \le n.$$

The only task left is the estimation of $\max_{x\in[a,b]}\left|\prod_{k=0}^{n}(x-x_k)\right|$. First we rewrite

$$\begin{aligned}\left|\prod_{k=0}^{n}(x-x_k)\right| &= \left|\prod_{k=0}^{n}\left(\frac{b+a}{2}+\frac{b-a}{2}x-x_k\right)\right|\\ &= \left|\prod_{k=0}^{n}\left(\frac{b-a}{2}x-\frac{b-a}{2}\cos\left(\frac{2k+1}{2n+2}\pi\right)\right)\right|.\end{aligned}$$

Notes

Choosing the nodes

Then factoring by $(b-a)/2$ we obtain

$$\left|\prod_{k=0}^{n}(x-x_k)\right| = \left(\frac{b-a}{2}\right)^{n+1}\left|\prod_{k=0}^{n}\left(x-\cos\left(\frac{2k+1}{2n+2}\pi\right)\right)\right|.$$

Finally the max over $[a,b]$ is given by

$$\begin{aligned}\max_{x\in[a,b]}\left|\prod_{k=0}^{n}(x-x_k)\right| &= \left(\frac{b-a}{2}\right)^{n+1}\max_{x\in[-1,1]}\left|\prod_{k=0}^{n}(x-x_k)\right|\\ &= \left(\frac{b-a}{2}\right)^{n+1}\frac{1}{2^n}\\ &= 2\left(\frac{b-a}{4}\right)^{n+1}.\end{aligned}$$

Notes

Choosing the nodes

Hence the error in the Lagrange interpolation of $f \in C^{n+1}[a,b]$ is minimized when choosing the nodes as the roots of the $(n+1)$th Chebyshev polynomials of the first kind. In this case the error is given by

$$|f(x)-L_n(f)| \le \frac{2M_n}{(n+1)!}\left(\frac{b-a}{4}\right)^{n+1}.$$

We now turn our attention to the “quality” of the interpolation as the number of nodes increases.

Notes

Stability of a numerical method

Definition (Stability of a numerical method)

In general, a numerical method is said to *stable* if a small perturbation on the inputs induces a small variation on the result.

In the case of Lagrange interpolation let $\widetilde{f}$ be the perturbed value of a continuous function f, and define $\varepsilon = \widetilde{f} - f$.

The Lagrange form of the interpolation polynomial (def. 3.134) of f and $\widetilde{f}$ is given by $L_n(f) = \sum_{i=0}^{n} f(x_i)\ell_i(x)$ and $L_n(\widetilde{f}) = \sum_{i=0}^{n}\widetilde{f}(x_i)\ell_i(x)$, respectively.

Notes

Stability of Lagrange method

Noting that

$$\begin{aligned} L_n(\widetilde{f}) - L_n(f) &= \sum_{i=0}^{n} \widetilde{f}(x_i)\ell_i(x) - \sum_{i=0}^{n} f(x_i)\ell_i(x) \\ &= \sum_{i=0}^{n} \left(\widetilde{f}(x_i) - f(x_i)\right)\ell_i(x) \\ &= L_n(\widetilde{f} - f), \end{aligned}$$

we get

$$\left|L_n(\varepsilon)\right| \leq \max\left|\widetilde{f}(x_i) - f(x_i)\right| \cdot \max_{x\in[a,b]} \sum_{i=0}^{n} \left|\ell_i(x)\right|. \tag{3.3}$$

This means that a perturbation on the inputs is "amplified" by a factor $\max_{x\in[a,b]} \sum_{i=0}^{n} \left|\ell_i(x)\right|$ on the output.

Notes

The Lebesgue constant

Definition

In Lagrange interpolation the number $\Lambda_n = \max_{x\in[a,b]} \sum_{i=0}^{n} \left|\ell_i(x)\right|$ is called the *Lebesgue constant*.

As the number n of nodes increases in Lagrange interpolation, Λ_n increases such that it tends to infinity. Since Lebesgue constant only depends on the choice of the nodes it means the more nodes the less stable.

It can be shown that any set of interpolation nodes leads to $\Lambda_n > \dfrac{2}{\pi}\log(n+1) + \dfrac{1}{2}$. In particular the Chebyshev nodes are nearly optimal, Λ_n growing at a rate $\frac{2}{\pi}\log n + \frac{2}{\pi}\left(\gamma + \log\frac{8}{\pi}\right) + o(1)$ with γ the Euler constant given by $\gamma = \int_0^\infty \frac{1}{\lfloor x\rfloor} - \frac{1}{x}\,\mathrm{d}x$.[3].

These results being more advanced and complex to prove we now focus our attention on the case of equidistant nodes.

[3] A slightly weaker result will be proven in homework 5

Notes

Stability for equidistant nodes

Proposition

When applying Lagrange interpolation with equidistant nodes, Λ_n becomes larger than $\dfrac{2^n}{4n^2}$ as n tends to infinity.

Proof. First observe that the result holds when $n = 1$ as $\Lambda_1 = 1$.

Let $k \geq 2$. Since the nodes are equidistant $x_k = x_0 + kh$ for all $1 < k \leq n$ and some constant h. Similarly for any $x \in [a, b]$ there exists s such that $x = x_0 + sh$. Then

$$\prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} = \prod_{\substack{j=0\\ j\neq i}}^{n} (s-j) \Big/ \prod_{\substack{j=0\\ j\neq i}}^{n} (i-j)$$

First we rewrite the denominator as $\left|\prod_{\substack{j=0\\ j\neq i}}^{n}(i-j)\right| = i!(n-i)!$.

Notes

Stability for equidistant nodes

The goal being to minimize $\max_{x\in[a,b]} \sum_{i=0}^{n} \left|\ell_i(x)\right|$ looking at the behaviour of the numerator for any point that is not a node will yield a result. Note however that this result will not be optimal, i.e. it might be far below the actual maximum.

For the sake of simplicity we now assume $s = 1/2$. In that case

$$\Lambda_n \geq \sum_{i=0}^{n} \left|\ell_i\left(x_0 + \frac{h}{2}\right)\right| = \sum_{i=0}^{n} \frac{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{3}{2}\cdots(i+\frac{3}{2})\cdot(i+\frac{1}{2})\cdots(n-\frac{1}{2})}{i!(n-i)!}.$$

As $k \geq 2$ it is easy to see that $k - 1/2 \geq k - 1$ and get

$$\left|\ell_i\left(x_0 + \frac{h}{2}\right)\right| \geq \frac{1}{4n(i-1)} \cdot \frac{1\cdot 2\cdots(n-1)n}{i!(n-i)!}.$$

Notes

153

Stability for equidistant nodes

Rearranging the terms and noting that $i-1 \leq n$ results in

$$\left|\ell_i(x_0+\frac{1}{2})\right| \geq \frac{n!}{4n^2 i!(n-i)!}.$$

Hence we finally obtain the expected result

$$\Lambda_n \geq \left(\sum_{i=0}^{n}\binom{n}{i}\right)\frac{1}{4n^2} = \frac{2^n}{4n^2}.$$

□

As we know how Λ_n behaves we now want to more precisely investigate how it affects L_n viewed as a function.

Notes

154

Lebesgue constant and Lagrange method

Theorem

Let $E = C[a,b]$ for $[a,b]$ a real interval and $F = \mathbb{R}_n[x]$. Then L_n is in $\mathcal{L}(E,F)$ and $\|L_n\| = \Lambda_n$.

Proof. It is simple to see that L_n is a linear application as

$$\begin{aligned} L_n(af+bg)(x) &= \sum_{i=0}^{n}\left(\ell_i(x)(af+bg)(x_i)\right) \\ &= \sum_{i=0}^{n}\left(\ell_i(x)(af(x_i)+bg(x_i))\right) \\ &= a\sum_{i=0}^{n}\left(\ell_i(x)f(x_i)\right) + b\sum_{i=0}^{n}\left(\ell_i(x)g(x_i)\right) \\ &= aL_n(f)(x) + bL_n(g)(x) \end{aligned}$$

Notes

155

Lebesgue constant and Lagrange method

Moreover as L_n is a continuous function over the real interval $[a,b]$ the extreme value theorem (2.79) states that L_n attains its maximal value. So L_n belongs to $\mathcal{L}(E,F)$.

Over $\mathcal{L}(E,F)$ we have $\|L_n\| = \sup\limits_{\substack{f\in[a,b]\\ f\neq 0}} \dfrac{\|L_n(f)\|_\infty}{\|f\|_\infty}$ (definition 1.42). But based on equation (**??**), we also have $\|L_n(f)\|_\infty \leq \|f\|_\infty\Lambda_n$, and we can conclude that

$$\|L_n\| \leq \Lambda_n. \tag{3.4}$$

The only thing left to prove is that $\|L_n\| \geq \Lambda_n$. Let $\xi \in [a,b]$ be such that

$$\Lambda_n = \sum_{i=0}^{n}|\ell_i(\xi)| = \max_{x\in[a,b]}\sum_{i=0}^{n}|\ell_i(x)|.$$

Notes

156

Lebesgue constant and Lagrange method

Defining $\text{sign}(x)$ over $\mathbb{R}$ by

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases},$$

we can rewrite Λ_n as

$$\Lambda_n = \sum_{i=0}^{n}\ell_i(\xi)\cdot\text{sign}(\ell_i(\xi)). \tag{3.5}$$

In order to prove the final part of the result we now construct a continuous, piecewise affine function such that $\|g\|_\infty = 1$, and $g(x_i) = \text{sign}(\ell_i(\xi))$, $i = 1,\cdots,n$.

Notes

157

Lebesgue constant and Lagrange method

Let $\sigma : \{0, \cdots, n\} \to \{0, \cdots, n\}$ such that for all $0 \leq i \leq n-1$, $x_{\sigma(i)} < x_{\sigma(i+1)}$. Then for each interval $[x_{\sigma(i)}, x_{\sigma(i+1)}]$ there exists a polynomial P_i of degree at most one such that $P_i(x_{\sigma(i)}) = \text{sign}(\ell_{\sigma(i)}(\xi))$ and $P_i(x_{\sigma(i+1)}) = \text{sign}(\ell_{\sigma(i+1)}(\xi))$.

"Connecting" all the "pieces" together we construct the function

$$g : [a, b] \longrightarrow \mathbb{R}$$

$$x \longmapsto \begin{cases} \text{sign}(\ell_{\sigma(0)}(\xi)) & a \leq x \leq x_{\sigma(0)} \\ P_i(x) & x_{\sigma(i)} \leq x \leq x_{\sigma(i+1)}, \quad 0 \leq i \leq n-1 \\ \text{sign}(\ell_{\sigma(n)}(\xi)) & x_{\sigma(n)} \leq x \leq b \end{cases}$$

By construction it is clear that g is continuous, $\|g\|_\infty = 1$, and $g(x_i) = \text{sign}(\ell_i(\xi))$.

Notes

158

Lebesgue constant and Lagrange method

Therefore equation (**??**) can be applied to g

$$\Lambda_n = \sum_{i=0}^{n} \ell_i(\xi) g(x_i) = L_n(g)(\xi).$$

In term on norms this means $\Lambda_n \leq \|L_n(g)\|_\infty \leq \|L_n\| \cdot \|g\|_\infty$. But as $\|g\|_\infty = 1$ we have proven that $\Lambda_n \leq \|L_n\|$.

Together with equation (**??**) this completes the proof.

□

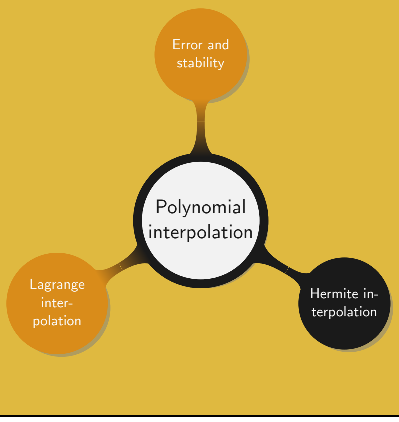
Corollary

The Lagrange interpolation method with equidistant nodes becomes unstable as the number of nodes increases.

Notes

159

Chapter organisation



Notes

160

General idea

Let f be a continuous function defined from a real interval $[a, b]$ into $\mathbb{R}$ and $x_0, \cdots, x_k$ be $k+1$ distinct points from $[a, b]$.

Assuming there is a finite family of integers $\alpha_0, \cdots, \alpha_k$ such that for all i, $f^{(l)}(x_i)$ exists for $0 \leq l \leq \alpha_i$, we want to find a polynomial P verifying

$$P^{(l)}(x_i) = f^{(l)}(x_i), \quad \text{for } 0 \leq l \leq \alpha_i, \text{ and } i = 0, \cdots, k.$$

From a more informal perspective we are given some values from an unknown function as well as some values of its derivatives. The goal is then to find a polynomial that interpolates this function at all the points and their defined derivatives.

This type of interpolation is called *Hermite interpolation*.

Notes

Theorem

Let $x_0, \cdots, x_k$ be $k+1$ distinct points from $[a, b]$, and f be a continuous function from $[a, b]$ into $\mathbb{R}$ such that there is a finite sequence of integers $\alpha_0, \cdots, \alpha_k$ for which $f^{(l)}(x_i)$ exists, $0 \leq l \leq \alpha_i$, and $i \in \{0, \cdots, k\}$.

Setting $n = k + \sum_{i=0}^{k} \alpha_i$, there exists a unique polynomial P of degree at most n, called *Hermite interpolation polynomial*, such that

$$P^{(l)}(x_i) = f^{(l)}(x_i), \quad \text{for } 0 \leq l \leq \alpha_i, \ i = 0, \cdots, k.$$

Proof. We define the application L by

$$L : \mathbb{R}_n[x] \longrightarrow \mathbb{R}^{n+1}$$
$$P \longmapsto \left(P(x_0), P'(x_0), \cdots, P^{(\alpha_0)}(x_0), \cdots, P(x_k), \cdots, P^{(\alpha_k)}(x_k)\right).$$

Notes

Since L is clearly a linear application different from 0, proving that it is bijective will directly yield the existence and unicity of P.

Let P be in the kernel of L. Then $L(P) = 0$, i.e. $P^{(l)}(x_j) = 0$, for all $0 \leq l \leq \alpha_j$, $0 \leq j \leq k$. In other words x_j is a root of order $\alpha_j + 1$ of P. Hence P is divisible by $\prod_{j=0}^{k}(x - x_j)^{\alpha_j+1}$ and there exists a polynomial q such that

$$P(x) = q(x) \prod_{j=0}^{k} (x - x_j)^{\alpha_j+1}.$$

If q is not identically equal to 0 then

$$\deg P = \deg q + \deg \prod_{j=0}^{k} (x - x_j)^{\alpha_j+1} = \deg q + \sum_{j=0}^{k} (\alpha_j + 1)$$
$$= \deg q + k + 1 + \sum_{j=0}^{k} \alpha_j = \deg q + n + 1.$$

Notes

We just proved that P is of degree larger than $n+1$, which is not since it belongs to $\mathbb{R}_n[x]$.↯

Hence q is identically 0, and $\ker L = \{0\}$. So by lemma 1.36, L is a bijection. □

Remark. The previous theorem does not provide any information on how to construct P. As a simple overview the idea consists in setting $P^{(l)}(x_i)$ to $f^{(l)}(x_i)$, for $0 \leq l \leq \alpha_i$, $i = 0, \cdots, k$. Then express all these elements in the canonical basis $(e_1, \cdots, e_n)$ of $\mathbb{R}^{n+1}$ and determine $L^{-1}(e_j)$ for all $j \in \{0, \cdots, n\}$. Finally write P in term of $L^{-1}(e_j)$, for all $j \in \{0, \cdots, n\}$.

Notes

Theorem

Let $x_0, \cdots, x_k$ be $k+1$ distinct points from $[a, b]$, and $f : [a, b] \to \mathbb{R}$ be in $C^{n+1}[a, b]$, such that there is a finite sequence of integers $\alpha_0, \cdots, \alpha_k$ for which $f^{(l)}(x_i)$ exists, $0 \leq l \leq \alpha_i$, and $i \in \{0, \cdots, k\}$.

If P is the Hermite interpolation polynomial of f at $x_0, \cdots, x_k$, then for all $x \in [a, b]$ there exists $\xi_x \in \left[\min(x, \min_i x_i), \max(x, \max_i x_i)\right]$ such that

$$f(x) - P(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^{k} (x - x_i)^{\alpha_i+1}.$$

Proof. Since the formula is clearly true when $x = x_i$, we assume $x \neq x_i$, $0 \leq i \leq k$.

Notes

165

Error

We now proceed following a similar strategy as in the case of Lagrange interpolation.

Calling $\Pi(t)$ the product $\prod_{i=0}^{k}(t-x_i)^{\alpha_i}+1$ we want to prove the existence of ξ_x such that

$$(n-1)!\frac{f(x)-P(x)}{\Pi(x)} - f^{(n+1)}(\xi_x) = 0,$$

where P is Hermite's polynomial.

Let $Q(t) = \frac{f(x)-P(x)}{\Pi(x)}\Pi(t) - f(t) + P(t)$. Then Q is a function with $n+2$ zeros: all the x_i counted with their multiplicity, and x.

Therefore, by Rolle's theorem (2.83) there exists ξ_x in the smallest interval containing the roots of Q, such that $Q^{(n+1)}(\xi_x) = 0$.

Notes

166

Error

Moreover as $P \in \mathbb{R}_n[x]$ we know that $P^{(n+1)} = 0$ and obtain

$$Q^{(n+1)}(t) = \frac{f(x)-P(x)}{\Pi(x)}\Pi^{(n+1)}(t) - f^{(n+1)}(t).$$

Finally, noting that Π is a monic polynomial from $\mathbb{R}_{n+1}[x]$ yields

$$\frac{f(x)-P(x)}{\Pi(x)}(n+1)! = f^{(n+1)}(\xi_x).$$

Hence

$$f(x) - P(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{k}(x-x_i)^{\alpha_{i+1}}.$$

□

Notes

167

Key points

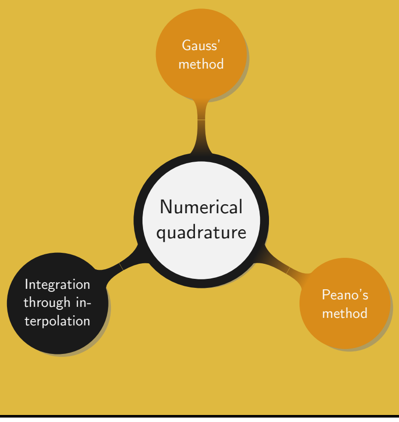
- Explain Lagrange interpolation
- What is a good choice of nodes for Lagrange interpolation?
- What is the stability of a numerical method?
- What does the Lebesgue constant measure?
- Explain Hermite interpolation

Notes

4. Numerical quadrature

Notes

Chapter organisation



Notes

Initial problem

Let $[a, b]$ be a compact from $\mathbb{R}$, f be a continuous function from $[a, b]$ into $\mathbb{R}$, and w be a weight function over (a, b).

For $\alpha, \beta \in [a, b]$ we want to approximate

$$I(f) = \int_{\alpha}^{\beta} f(x)w(x)\,\mathrm{d}x,$$

by a formula of type

$$I_k(f) = \sum_{i=0}^{k} \lambda_{i,k} f(x_i),$$

where the x_i, $0 \leq i \leq n$ are nodes from $[a, b]$ (either given or to be determined), and the λ_i are to be calculated.

Remark. The maps $I_k : C[a, b] \to \mathbb{R}$, $f \mapsto I_k(f)$ are linear forms.

Notes

General setting

Definitions

Let $\left(I_k(f)\right)_{k\geq 0}$ be a sequence of approximation of $I(f)$.

1. The sequence $I_k(f)$ is *convergent* if $\lim_{k\to\infty} I_k(f) = I(f)$;
2. The *error of approximation* is given by $E_k(f) = I(f) - I_k(f)$;
3. An integration method is said to be of *order* N, $N \in \mathbb{N}$, if for any polynomial $P \in \mathbb{R}_N[x]$, $E_k(P) = 0$, $k \geq N$;
4. An integration method is said to be of *order exactly* N if it is of order N and there exists $P \in \mathbb{R}_{N+1}[x]$ such that $E_k(P) \neq 0$ for some $k \geq N$;

Notes

Toward an integration method

Remark. Since $I_k(f)$ is a linear application the order of an integration method can be easily determined by finding the largest N such that $E_k(x^N) = 0$ and $E_k(x^{N+1}) \neq 0$.
The order of a method measures how precise it is. A "good method" is expected to be convergent, precise, stable, and easily implementable.

When facing a "complex" integral the most straightforward solution would be to approximate the function by its interpolation polynomial and integrate it. Unfortunately the result quickly becomes inconsistent as the size of the interval of integration increases. To patch up this idea we divide up the initial interval $[\alpha, \beta]$ and *mesh* the domain with sub-intervals $[\alpha_i, \alpha_{i+1}]$. On each of those, the approximation of the function by a polynomial is expected to be precise enough to yield a consistent results.

Notes

173

Toward an integration method

In other words, for a function f continuous on $[a, b]$ and P its interpolation polynomial over $[\alpha_i, \alpha_{i+1}]$ we want to approximate

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)w(x)\,\mathrm{d}x \approx \int_{\alpha_i}^{\alpha_{i+1}} P(x)w(x)\,\mathrm{d}x.$$

Therefore this strategy will only work for "nice" weight functions.

Example. For the constant weight function $w(x) = 1$ we rewrite

$$\int_{\alpha}^{\beta} f(x)\,\mathrm{d}x = \sum_{i=0}^{k-1} \int_{\alpha_i}^{\alpha_{i+1}} f(x)\,\mathrm{d}x.$$

Assuming $f(x)$ is interpolated at $\xi_i \in [\alpha_i, \alpha_{i+1}]$ by a polynomial of degree 0 we obtain the approximation formula

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)\,\mathrm{d}x \approx \int_{\alpha_i}^{\alpha_{i+1}} f(\xi_i)\,\mathrm{d}x = (\alpha_{i+1} - \alpha_i)f(\xi_i). \tag{4.1}$$

Notes

174

Toward an integration method

We now evaluate how precise this formula is. Clearly if $f(x) = 1$ then the formula is exact. If $f(x) = x$ then we get

$$\int_{\alpha_i}^{\alpha_{i+1}} x\,\mathrm{d}x = \frac{(\alpha_{i+1} - \alpha_i)(\alpha_{i+1} + \alpha_i)}{2}.$$

From equation (**??**) we have to distinguish two cases: if $\xi_i \neq (\alpha_{i+1}+\alpha_i)/2$ then the formula is of order exactly 0; otherwise the formula is of order at least 1. But as for $f(x) = x^2$

$$\int_{\alpha_i}^{\alpha_{i+1}} x^2\,\mathrm{d}x \neq \left(\frac{\alpha_{i+1} + \alpha_i}{2}\right)^2 (\alpha_{i+1} - \alpha_i),$$

it is of order exactly 1 and the *composite quadrature formula* over $[\alpha, \beta]$ is given by

$$\int_{\alpha}^{\beta} f(x) \approx \sum_{i=0}^{k} f(\xi_i)(\alpha_{i+1} - \alpha_i) = I_k.$$

Since this is a Riemann sum of f, $\lim_{k\to\infty} I_k = \int_\alpha^\beta f(x)\,\mathrm{d}x$. Hence the method converges.

Notes

175

A method based on interpolation

Following the previous example we want to investigate the more general case where the function f is approximated by a polynomial of arbitrary degree using Lagrange interpolation.

Let f be a continuous function over $[\alpha, \beta] \subset [a, b]$. We want to derive a general formula in the case of a weight function $w(x) = 1$.

For the sake of simplicity we set $\alpha_{i+\frac{1}{2}} = (\alpha_i + \alpha_{i+1})/2$ and remap each of the $[\alpha_i, \alpha_{i+1}]$ into $[-1, 1]$, a "reference interval", using the change of variable

$$\begin{aligned} \varphi_i^{-1} : [\alpha_i, \alpha_{i+1}] &\longrightarrow [-1, 1] & \text{and} \quad \varphi_i : [-1, 1] &\longrightarrow [\alpha_i, \alpha_{i+1}] \\ x &\longmapsto 2\frac{x - \alpha_{i+\frac{1}{2}}}{\alpha_{i+1} - \alpha_i}, & s &\longmapsto \frac{\alpha_{i+1} - \alpha_i}{2} s + \alpha_{i+\frac{1}{2}}. \end{aligned} \tag{4.2}$$

Hence, once the change of variable applied, the initial problem boils down to finding a quadrature formula for $\int_{-1}^{1} \varphi(s)\,\mathrm{d}s$, for $\varphi \in C[-1, 1]$.

Notes

176

A method based on interpolation

We start by approximating φ by P_φ, its Lagrange interpolation polynomial at the l pairwise distinct nodes $\tau_1, \cdots, \tau_l \in [-1, 1]$. Hence

$$P_\varphi(s) = \sum_{j=0}^{l} \varphi(\tau_j)\ell_j(s) \quad \text{with, } \ell_j(s) = \prod_{\substack{i=0 \\ i\neq j}}^{l} \frac{s - \tau_i}{\tau_j - \tau_i}.$$

The quadrature formula over $[-1, 1]$ is then given by

$$\begin{aligned} \int_{-1}^{1} \varphi(s)\,\mathrm{d}s &\approx \int_{-1}^{1} P_\varphi(s)\,\mathrm{d}s \\ &= 2\sum_{j=0}^{l} \varphi(\tau_j)w_j, \quad \text{with } w_j = \frac{1}{2}\int_{-1}^{1} \ell_j(s)\,\mathrm{d}s. \end{aligned} \tag{4.3}$$

Using the change of variable (**??**) we obtain the quadrature formula over $[\alpha_i, \alpha_{i+1}]$.

Notes

A method based on interpolation

Renaming $\alpha_{i+1}-\alpha_i$ into h_i we have

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)\,\mathrm{d}x = \frac{h_i}{2}\int_{-1}^{1}\varphi_i(s)\,\mathrm{d}s \approx h_i\sum_{j=0}^{l}\varphi_i(\tau_j)w_j = h_i\sum_{j=0}^{l} w_j f(\alpha_{ij}), \qquad (4.4)$$

where $\alpha_{ij}=\tau_j\frac{h_i}{2}+\alpha_{i+\frac{1}{2}}$. By construction α_{ij} is in $[\alpha_i,\alpha_{i+1}]$ since τ_j is in $[-1,1]$.

Therefore this yields the *elementary quadrature formula*

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)\,\mathrm{d}x \approx h_i\sum_{j=0}^{l} f(\alpha_{ij})w_j,$$

while the associated composite formula is given by

$$\int_{\alpha}^{\beta} f(x)\,\mathrm{d}x \approx \sum_{i=0}^{k-1} h_i\sum_{j=0}^{l} f(\alpha_{ij})w_j.$$

Notes

Convergence of the method

The previous composite quadrature formula formally expresses our initial idea described on slide 4.172. After deriving the quadrature formula we now investigate its "quality", starting with the study of its convergence.

Theorem

Let f be a continuous function over $[\alpha,\beta]\subset[a,b]$. Using the previous notations, let $(I_k)_{k\in\mathbb{N}}$ be the sequence defined by

$$I_k=\sum_{i=0}^{k-1} h_i\sum_{j=0}^{l} f(\alpha_{ij})w_j.$$

Then $(I_k)_{k\in\mathbb{N}}$ converges toward $\int_\alpha^\beta f(x)\,\mathrm{d}x$ as k tends to infinity. Moreover the method of integration through interpolation over $l+1$ nodes has order at least l.

Notes

Convergence of the method

Proof. When rearranging the terms of I_k into

$$\sum_{i=0}^{k-1} h_i\sum_{j=0}^{l} f(\alpha_{ij})w_j=\sum_{j=0}^{l}\left[\sum_{i=0}^{k-1} h_i f(\alpha_{ij})\right]w_j,$$

one can observe that $\sum_{i=0}^{k-1} h_i f(\alpha_{ij})$ is in fact a Riemann sum.

Thus when k tends to infinity the step of the mesh h_i tends to zero and

$$\int_\alpha^\beta f(x)\,\mathrm{d}x=\lim_{k\to\infty}\sum_{i=0}^{k-1} h_i f(\alpha_{ij}).$$

Hence

$$\lim_{k\to\infty} I_k=\sum_{j=0}^{l}\left(w_j\int_\alpha^\beta f(x)\,\mathrm{d}x\right). \qquad (4.5)$$

Notes

Convergence of the method

From equation (**??**) it remains to prove that $\sum_{j=0}^{l} w_j=1$. Recalling that $w_j=\frac{1}{2}\int_{-1}^{1}\ell_j(s)\,\mathrm{d}s$, it is clear that w_j does not depend upon the choice of the function φ.

In particular in the case where $\varphi(s)=1$ over $[-1,1]$, P_φ is also equal to 1 and by the quadrature formula (**??**)

$$\int_{-1}^{1} P_\varphi(s)\,\mathrm{d}s=2\sum_{j=0}^{l} w_j=2.$$

This proves that $\lim_{k\to\infty} I_k=\int_\alpha^\beta f(x)\,\mathrm{d}x.$

The order of the method is trivially seen as soon as one notices that $\varphi=P_\varphi$ when $\varphi\in\mathbb{R}_l[x]$. □

Notes

181

Stability of the method

Theorem

If w_j is strictly positive for any j in $\{0, \cdots, l\}$, then the integration method is stable with respect to the variations of the continuous function f to be integrated.

Proof. Assuming $I_k(f)$ is given by

$$I_k(f) = \sum_{i=0}^{k-1}(\alpha_{i+1} - \alpha_i)\sum_{j=0}^{l} f(\alpha_{ij})w_j,$$

a small perturbation on f changes $f(\alpha_{ij})$ into $\widetilde{f}_{ij}$ yielding

$$I_k(\widetilde{f}) = \sum_{i=0}^{k-1}(\alpha_{i+1} - \alpha_i)\sum_{j=0}^{l} \widetilde{f}_{ij}w_j.$$

Notes

182

Stability of the method

Looking at the difference between $I_k(\widetilde{f})$ and $I_k(f)$ we get

$$\left|I_k(\widetilde{f}) - I_k(f)\right| \leq \sum_{i=0}^{k-1}(\alpha_{i+1} - \alpha_i)\sum_{j=0}^{l} \left|f(\alpha_{ij}) - \widetilde{f}_{ij}\right| w_j.$$

Then denoting by ε the maximum of $|f(\alpha_{ij}) - \widetilde{f}_{ij}|$, and recalling that w_j is strictly positive we see that the method is stable since

$$\left|I_k(\widetilde{f}) - I_k(f)\right| \leq \varepsilon\sum_{i=0}^{k-1}(\alpha_{i+1} - \alpha_i)\sum_{j=0}^{l} w_j = \varepsilon(\beta - \alpha).$$

□

Notes

183

Error of the method

For φ in $C^{l+1}[-1, 1]$, one can directly determine

$$E(\varphi) = \int_0^1 \varphi(s)\,\mathrm{d}s - \int_0^1 P_\varphi(s)\,\mathrm{d}s.$$

That is calculate

$$\varphi(s) - P_\varphi(s) = \frac{\varphi^{(l+1)}(\xi_x)}{(l+1)!}\prod_{j=0}^{l}(s - \tau_j),$$

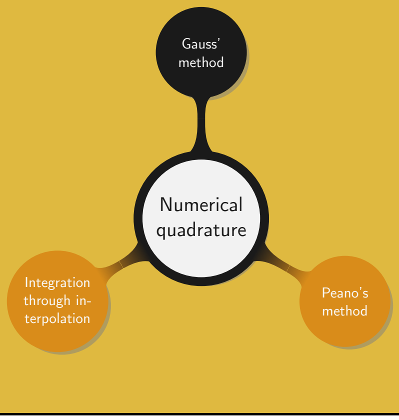
and then study the result in order to find the most optimal choice of nodes and "control" $E(\varphi)$.

In section 3 we will investigate a more general approach which covers integration through interpolation and Gauss' method. In particular results on the study of the error will be proven there.

Notes

184

Chapter organisation



Notes

185

Basic idea

Given f a continuous function over $[a,b]$, and w a weight function over $(\alpha,\beta) \subset [a,b]$, we want to approximate the integral $\int_\alpha^\beta f(x)w(x)\,\mathrm{d}x$ by a quadrature formula of the form $\sum_{i=0}^{k} \lambda_i f(x_i)$, for some numbers λ_i, $0 \leq i \leq k$, where k is a positive integer.

For l smaller than k, Gauss method chooses $k-l+1$ points, among the $k+1$ nodes, outside (α,β). The goal is then to find all the nodes and λ_i such that the integration formula

$$\int_\alpha^\beta f(x)w(x)\,\mathrm{d}x \approx \sum_{i=0}^{k} \lambda_i f(x_i)$$

is of order at least $k+l$.

Notes

186

Existence and unicity

Theorem

Let k and l be two integers such that $k \geq l-1$ and $x_0,\cdots,x_{l-1}$ be in $[\alpha,\beta]$ while $x_l,\cdots,x_k$ are not in this interval. For x in (α,β), let $\Pi_{k,l}(x)$ denote the product $\prod_{j=l}^{k}(x-x_j)$. Then there exists a quadrature formula that is uniquely determined by $x_l,\cdots,x_k$. It is of the form

$$\sum_{i=0}^{k} \lambda_i f(x_i) \approx \int_\alpha^\beta f(x)w(x)\,\mathrm{d}x$$

and has order $k+l$. Moreover the x_i, $0 \leq i \leq l-1$, are the roots of the $(l+1)$th orthogonal polynomial P_l associated to the weight

$$\theta(x) = \operatorname{sign}\left(\Pi_{k,l}(x)\right)\Pi_{k,l}(x)w(x), \quad x \in (\alpha,\beta).$$

Notes

187

Existence and unicity

Proof. Before proving the existence and unicity of the quadrature formula we first need to verify that

$$\forall x \in (\alpha,\beta) \quad \theta(x) = \operatorname{sign}\left(\Pi_{k,l}(x)\right)\Pi_{k,l}(x)w(x)$$

is a weight function.

Since $x_l,\cdots,x_k \notin (\alpha,\beta)$ we see that $\Pi_{k,l}(x) = \prod_{j=l}^{k}(x-x_j)$ does not change sign over (α,β). So its sign function is either equal to 1 or -1. Either way, $\theta(x)$ is strictly positive over (α,β). Moreover

$$\int_\alpha^\beta \theta(x)\,\mathrm{d}x \leq \max_t \left|\Pi_{k,l}(t)\right| \int_\alpha^\beta w(x)\,\mathrm{d}x.$$

As $w(x)$ is a weight function over $(a,b) \supset (\alpha,\beta)$ it is clear that $\int_\alpha^\beta \theta(x)\,\mathrm{d}x < \infty$. Therefore $\theta(x)$ is a weight function over (α,β).

Notes

188

Existence and unicity

We now focus on the unicity of the quadrature formula, i.e. we prove that if the formula exists and is of order $k+l$, then the x_i, $0 \leq i < l$ and λ_i, $0 \leq i \leq k$ are unique.

Let $\widetilde{P}_l(x) = \prod_{j=0}^{l-1}(x-x_j)$. We want to prove that $\widetilde{P}_l = P_l$, the $(l+1)$st orthogonal polynomial associated to the weight $\theta(x)$, i.e.

$$\int_\alpha^\beta \widetilde{P}_l(x)Q(x)\theta(x)\,\mathrm{d}x = 0, \quad \forall Q \in \mathbb{R}_{l-1}[x].$$

Let Q be a polynomial of degree less than $l-1$. As the formula is of order $k+l$ and of the form $\int_\alpha^\beta f(x)w(x)\,\mathrm{d}x \approx \sum_{i=0}^{k} \lambda_i f(x_i)$ then

$$\int_\alpha^\beta \widetilde{P}_l(x)Q(x)\theta(x)\,\mathrm{d}x = \operatorname{sign}\left(\Pi_{k,l}(x)\right) \int_\alpha^\beta \prod_{j=0}^{k}(x-x_j)Q(x)w(x)\,\mathrm{d}x$$

$$= \sum_{j=0}^{k} \lambda_j \cdot 0 = 0.$$

Notes

189

Existence and unicity

This proves that for any polynomial Q of degree at most $l-1$

$$\int_\alpha^\beta \widetilde{P}_l(x)Q(x)\theta(x)\,\mathrm{d}x = 0.$$

In other words $\widetilde{P}_l$ is an orthogonal polynomial, and as it is monic it is equal to P_l (proposition 2.118). In particular this means that the roots of $\widetilde{P}_l$, $x_0, \cdots, x_{l-1}$, are uniquely defined with respect to the weight function θ.

For the unicity of the λ_i observe that we have just demonstrated that the choice of the nodes $x_l, \cdots, x_k$ fixes $x_0, \cdots, x_{l-1}$. Thus all the x_i, $0 \le i \le k$ are fixed and known, such that we can consider

$$\ell_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{k} \frac{x-x_j}{x_i-x_j}.$$

Notes

190

Existence and unicity

Since ℓ_i is a polynomial of degree $k \le k+l$, the formula is exact and we obtain

$$\begin{aligned}\int_\alpha^\beta \ell_i(x)w(x)\,\mathrm{d}x &= \sum_{j=0}^{k} \lambda_i \ell_i(x_j) \\ &= \lambda_i \ell_i(x_i) = \lambda_i.\end{aligned} \tag{4.6}$$

As a result all the λ_i are uniquely determined by the x_i, $0 \le i \le k$.

From the previous parts of the demonstration we know that P_l is the $(l+1)st$ orthogonal polynomial associated to the weight function $\theta(x)$ over (α, β). Furthermore P_l has degree l and its roots $x_0, \cdots, x_{l-1}$ are all distinct.

Notes

191

Existence and unicity

For λ_i defined as in (**??**) it is left to prove that the quadrature formula

$$\int_\alpha^\beta f(x)w(x)\,\mathrm{d}x \approx \sum_{j=0}^{k} \lambda_i f(x_j)$$

has order at least $k+l$.
By construction it is clear that the method is of order at least k, since it suffices to consider the basis composed of the Lagrange polynomials ℓ_i. So let f be a polynomial of degree strictly larger than k, but less than $k+l$. It is then possible to proceed to the euclidean division of f by $\prod_{j=0}^{k}(x-x_j)$ and there exist q and R two polynomials such that

$$f(x) = q(x)\prod_{j=0}^{k}(x-x_j) + R(x), \quad \text{with} \begin{cases} \deg R \le k \\ \deg q \le l-1. \end{cases}$$

Notes

192

Existence and unicity

Then the quadrature formula can be rewritten in terms of q and R

$$\int_\alpha^\beta f(x)w(x)\,\mathrm{d}x = \underbrace{\int_\alpha^\beta q(x)\prod_{j=0}^{k}(x-x_j)w(x)\,\mathrm{d}x}_{A} + \int_\alpha^\beta R(x)w(x)\,\mathrm{d}x.$$

Note that R has degree less than k. As in this case we already know the quadrature formula to be exact this yields

$$\int_\alpha^\beta R(x)w(x)\,\mathrm{d}x = \sum_{i=0}^{k} \lambda_i R(x_i) = \sum_{i=0}^{k} \lambda_i f(x_i).$$

Therefore for the quadrature formula to be of order $k+l$ we need to show that A equals 0.

Notes

193

Existence and unicity

Recalling that θ does not change sign over (α, β) (slide 4.187)

$$\begin{aligned}\prod_{j=0}^{k}(x-x_j)w(x) &= \prod_{j=0}^{l-1}(x-x_j)\prod_{j=l}^{k}(x-x_j)w(x)\\ &= P_l(x)\operatorname{sign}\left(\Pi_{k,l}(x)\right)\theta(x),\end{aligned}$$

such that we get

$$A = \int_\alpha^\beta q(x)P_l(x)\operatorname{sign}\left(\Pi_{k,l}(x)\right)\theta(x)\,\mathrm{d}x.$$

Since q is of degree less than $l-1$, P_l is orthogonal to q for the weight θ. Hence $A = 0$ and this concludes the proof. □

Notes

194

Existence and unicity

Corollary

When $k = l - 1$, i.e. no node is taken out of (α, β), there exists a unique quadrature formula of the form

$$\int_\alpha^\beta f(x)w(x)\,\mathrm{d}x \approx \sum_{i=0}^{k}\lambda_i f(x_i),$$

and with order $2k+1$. The $x_i \in (\alpha, \beta)$, $0 \leq i \leq k$, are the roots of the $(k+2)$nd orthogonal polynomial associated to the weight w. Moreover all the λ_i are strictly positive.

Proof. The only part of the corollary that remains to be proven is that all the λ_i, $0 \leq i \leq k$, are strictly larger than 0.

Notes

195

Existence and unicity

From the first part of the result we know that the quadrature formula is of order at least $2k+1$, which is larger than the degree of ℓ_i^2. Thus, as the formula is exact we have

$$\int_\alpha^\beta \ell_i^2(x)w(x)\,\mathrm{d}x = \sum_{j=0}^{k}\lambda_j\ell_i^2(x_j) = \lambda_i.$$

Hence noting that $\ell_i^2(x)w(x)$ is strictly positive over (α, β), we have proven that λ_i is strictly positive. □

Following the existence and unicity we now focus on the quality of Gauss' integration method by estimating its error.

Notes

196

Error

Theorem

Let $x_0, \cdots, x_{l-1}$ be the nodes from (α, β) and $x_l, \cdots, x_k$ be the points of $[a, b]$ exterior to (α, β). Then for any function $f \in C^{k+l+1}[a, b]$ there exists ξ in $[a, b]$ such that

$$E(f) = \frac{f^{(k+l+1)}(\xi)}{(k+l+1)!}\int_\alpha^\beta \prod_{i=0}^{l-1}(x-x_i)^2\prod_{j=l}^{k}(x-x_j)w(x)\,\mathrm{d}x.$$

Proof. Let P be the Hermite interpolation polynomial of f over $[a, b]$ at the nodes $x_0, \cdots, x_{l-1}, x_l, \cdots, x_k$, with

$$\begin{aligned}P(x_i) &= f(x_i), \quad \text{for } 0 \leq i \leq k,\\ P'(x_i) &= f'(x_i), \quad \text{for } 0 \leq i \leq l-1.\end{aligned}$$

From theorem 3.161, P has degree $n = k + \sum_{i=0}^{k}\alpha_i = k + l$, since $\alpha_i = 0$ for $i \in l, \cdots, k$ and 1 otherwise.

Notes

Error

Since Gauss' method has order at least $k+l$ we have

$$\int_{\alpha}^{\beta} P(x)w(x)\,\mathrm{d}x = \sum_{i=0}^{k} \lambda_i P(x_i) = \sum_{i=0}^{k} \lambda_i f(x_i),$$

and we can write

$$\begin{aligned} E(f) &= \int_{\alpha}^{\beta} f(x)w(x)\,\mathrm{d}x - \sum_{i=0}^{k} \lambda_i f(x_i) \\ &= \int_{\alpha}^{\beta} f(x)w(x)\,\mathrm{d}x - \int_{\alpha}^{\beta} P(x)w(x)\,\mathrm{d}x \\ &= \int_{\alpha}^{\beta} \left(f(x) - P(x)\right) w(x)\,\mathrm{d}x. \end{aligned}$$

Notes

Error

From the error formula for Hermite interpolation polynomial (theorem 3.164) we know that for all $x \in [a,b]$ there exists $\xi_x \in [a,b]$ such that

$$f(x) - P(x) = \frac{f^{(k+l+1)}(\xi_x)}{(k+l+1)!}\Pi(x),$$

with $\Pi(x) = \prod_{i=0}^{k}(x-x_i)^{\alpha_i+1}$.
Note that in our case α_i is 0 for $i \in \{l, \cdots, k\}$ and 1 for $i \in \{0, \cdots, l-1\}$.
Without loss of generality we now assume $\int_{\alpha}^{\beta} \Pi(x)w(x)\,\mathrm{d}x$ to be strictly positive. This is possible since Π has a constant sign over (α, β).

Notes

Error

Then, as $f \in C^{(k+l+1)}[a,b]$, the extreme values theorem (2.79) can be applied and

$$\frac{m}{(k+l+1)!}\int_{\alpha}^{\beta} \Pi(x)w(x)\,\mathrm{d}x \le E(f) \le \frac{M}{(k+l+1)!}\int_{\alpha}^{\beta} \Pi(x)w(x)\,\mathrm{d}x,$$

with $m = \min_{x\in[a,b]} f^{(k+l+1)}(x)$ and $M = \max_{x\in[a,b]} f^{(k+l+1)}(x)$.

Hence

$$m \le (k+l+1)!\frac{E(f)}{\int_{\alpha}^{\beta} \Pi(x)w(x)\,\mathrm{d}x} \le M. \tag{4.7}$$

Applying the intermediate values theorem (2.76) to $f^{(k+l+1)}$, for any q in $[m,M]$, there exists ξ such that $f^{(k+l+1)}(\xi) = q$.

Notes

Error

In particular from (**??**) we know that $(k+l+1)!\frac{E(f)}{\int_{\alpha}^{\beta} \Pi(x)w(x)\,\mathrm{d}x}$ is in $[m,M]$, meaning there exists $\xi \in [a,b]$ such that

$$E(f) = \frac{f^{(k+l+1)}(\xi)}{(k+l+1)!}\int_{\alpha}^{\beta} \prod_{i=0}^{l-1}(x-x_i)^2 \prod_{j=l}^{k}(x-x_j)w(x)\,\mathrm{d}x.$$

□

Corollary

Gauss' method has order exactly $k+l$.

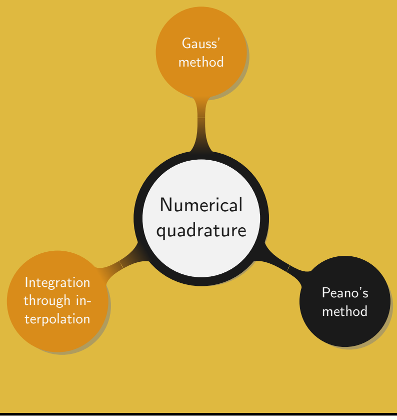
Proof. It is clear since the method is of order at least $k+l$ and

$$E(x \mapsto x^{k+l+1}) = \frac{(k+l+1)!}{(k+l+1)!}\int_{\alpha}^{\beta} \Pi(x)w(x)\,\mathrm{d}x \neq 0.$$

□

Notes

Chapter organisation



Notes

General idea

In this section we investigate a generic integration method where a minimum amount of assumptions is assumed. In particular it covers both integration through interpolation and Gauss' method.

Let $w(x)$ be a weight function over $(\alpha, \beta) \subset [a, b]$ and f be a function defined almost everywhere on $[a, b]$ and which is Lebesgue integrable on (α, β) with respect to the weight w. We are interested in the generic quadrature formula

$$\int_\alpha^\beta f(x)w(x)\,\mathrm{d}x \approx \sum_{i=0}^{k} \lambda_i f(x_i), \quad \lambda_i \in \mathbb{R}, x_i \in [a, b], \forall 0 \leq i \leq k.$$

This method is called *Peano's method*.

Notes

Peano kernel

Definition

The application K_N, called *Peano kernel*, is defined by

$$K_N(t) = E(x \in [\alpha, \beta] \mapsto [(x-t)_+]^N),$$

where

$$[(x-t)_+]^N = \begin{cases} (x-t)^N, & \text{if } t \leq x \\ 0, & \text{if } t > x \end{cases}$$

Theorem (Peano kernel theorem)

Assuming Peano's method is of positive order N, f is a $C^{N+1}[a, b]$ function, and $f^{(N)}$ is absolutely continuous then

$$E(f) = \frac{1}{N!} \int_a^b K_N(t) f^{(N+1)}(t)\,\mathrm{d}t.$$

Notes

Peano kernel theorem

Proof. From Taylor's theorem (2.101) f can be rewritten

$$f(x) = P_N(x) + \frac{1}{N!} \int_a^x (x-t)^N f^{(N+1)}(t)\,\mathrm{d}t,$$

with $P_N \in \mathbb{R}_N[x]$.

Replacing $(x-t)$ by $(x-t)_+$ we make the range of integration independent of x to obtain

$$f(x) = P_N(x) + \frac{1}{N!} \int_a^b [(x-t)_+]^N f^{(N+1)}(t)\,\mathrm{d}t.$$

Using the definition of f we now evaluate the error of Peano's method.

Notes

205

Peano kernel theorem

The error is given by

$$\begin{aligned}E(f) &= E(P_N) + \frac{1}{N!} E\left(x \mapsto \int_a^b \left[(x-t)_+\right]^N f^{(N+1)}(t)\,\mathrm{d}t \right)\\ &= \frac{1}{N!} \left[\int_\alpha^\beta \left(\int_a^b \left[(x-t)_+\right]^N f^{(N+1)}(t)\,\mathrm{d}t \right) w(x)\,\mathrm{d}x \right.\\ &\qquad \left. - \sum_{i=0}^k \lambda_i \int_a^b \left[(x_i-t)_+\right]^N f^{(N+1)}(t)\,\mathrm{d}t \right].\end{aligned}$$

By the linearity of integration and Fubini's theorem (2.98)

$$\begin{aligned}E(f) = \frac{1}{N!} \int_a^b \Bigg[&\left(\int_\alpha^\beta \left[(x-t)_+\right]^N w(x)\,\mathrm{d}x \right) f^{(N+1)}(t)\\ &- \sum_{i=0}^k \lambda_i \left[(x_i-t)_+\right]^N f^{(N+1)}(t) \Bigg] \mathrm{d}t.\end{aligned}$$

Notes

206

Peano kernel theorem

Rearranging the terms, Peano kernel appears, yielding the result

$$\begin{aligned}E(f) &= \frac{1}{N!} \int_a^b \left[\int_\alpha^\beta \left[(x-t)_+\right]^N w(x)\,\mathrm{d}x - \sum_{i=0}^k \lambda_i \left[(x_i-t)_+\right]^N \right] f^{(N+1)}(t)\,\mathrm{d}t\\ &= \frac{1}{N!} \int_a^b K_N(t) f^{(N+1)}(t)\,\mathrm{d}t.\end{aligned}$$

□

Corollary

Under the same assumptions as Peano kernel theorem we have

$$|E(f)| \le \frac{1}{N!} \sup_{t\in[a,b]} |f^{(N+1)}(t)| \int_a^b |K_N(t)|\,\mathrm{d}t.$$

Notes

207

Peano kernel theorem

Corollary

Under the same assumptions as Peano kernel theorem and assuming Peano kernel does not change sign, there exists $\xi \in (a,b)$ such that

$$E(f) = \frac{f^{(N+1)}(\xi)}{(N+1)!} E(x \mapsto x^{N+1}).$$

Proof. Since K_N is integrable and has a constant sign over $[a,b]$, while $f^{(N+1)}$ is continuous we can apply the first mean value theorem for integrals (theorem 2.99). Then there exists $\xi \in [a,b]$ such that

$$E(f) = \frac{f^{(N+1)}(\xi)}{N!} \int_a^b K_N(t)\,\mathrm{d}t. \tag{4.8}$$

Notes

208

Peano kernel theorem

Determining (**??**) for $f(x) = x^{N+1}$ gives

$$\begin{aligned}E\left(x \mapsto x^{N+1}\right) &= \frac{(N+1)!}{N!} \int_a^b K_N(t)\,\mathrm{d}t\\ &= (N+1) \int_a^b K_N(t)\,\mathrm{d}t.\end{aligned}$$

Thus replacing $\int_a^b K_N(t)\,\mathrm{d}t$ by $E(x \mapsto x^{N+1})/(N+1)$ completes the proof.

□

Example. Determine Peano kernel for the rectangle method

$$\int_\alpha^\beta f(x)\,\mathrm{d}x \approx f(\xi)(\beta-\alpha), \quad \text{for some } \xi \in [\alpha,\beta].$$

Notes

Rectangle method

Reminding that the rectangle method has been investigated in example 4.173 we already know that it is of order 0 if ξ is not $\frac{\alpha+\beta}{2}$ and 1 otherwise.

Case $\xi \neq \frac{\alpha+\beta}{2}$:

$$\begin{aligned} K_0(t) &= E(x \mapsto [(x-t)_+]^0) \\ &= \int_\alpha^\beta [(x-t)_+]^0 \, \mathrm{d}x - [(\xi - t)_+]^0(\beta-\alpha) \\ &= \int_t^\beta \mathrm{d}x - [(\xi-t)_+]^0(\beta-\alpha) \\ &= (\beta - t) - [(\xi-t)_+]^0(\beta-\alpha) \end{aligned}$$

Now if $t \leq \xi$ then $[(\xi-t)_+]^0$ is 1, implying $K_0(t) = \alpha - t$. And when $t > \xi$ then $[(\xi-t)_+]^0$ is 0, such that $K_0(t) = \beta - t$.

Notes

Rectangle method

Case $\xi = \frac{\alpha+\beta}{2}$:

$$\begin{aligned} K_1(t) &= E(x \mapsto [(x-t)_+]^1) \\ &= \int_\alpha^\beta (x-t)_+ \, \mathrm{d}x - \left(\frac{\alpha+\beta}{2} - t\right)_+ (\beta-\alpha) \\ &= \int_t^\beta x - t \, \mathrm{d}x - \left(\frac{\alpha+\beta}{2} - t\right)_+ (\beta-\alpha) \\ &= \frac{(\beta-t)^2}{2} - \left(\frac{\alpha+\beta}{2} - t\right)_+ (\beta-\alpha) \end{aligned}$$

Now if $t \leq \xi$ then $[(\xi-t)_+]^1$ is $\left(\frac{\alpha+\beta}{2} - t\right)$, implying $K_1(t) = \frac{(t-\alpha)^2}{2}$. And when $t > \xi$ then $[(\xi-t)_+]^1$ is 0, such that $K_1(t) = \frac{(\beta-t)^2}{2}$.

Notes

Gauss' method

Theorem

In the case of Gauss' method, Peano kernel keeps a constant sign on $[\alpha, \beta]$.

Proof. Suppose Peano kernel changes sign on $[\alpha,\beta]$. Then, there should exist a continuous function φ such that for all $x \in [a,b]$

$$\int_\alpha^\beta K_N(t)\varphi(t)\,\mathrm{d}t \neq \varphi(x)\int_\alpha^\beta K_N(t)\,\mathrm{d}t. \qquad (4.9)$$

Indeed, observe that

- if $\int_\alpha^\beta K_N(t)\,\mathrm{d}t \leq 0$, then taking $\varphi(x) = K_N(x)_+$ ensures that the left and right hand sides have different signs;
- if $\int_\alpha^\beta K_N(t)\,\mathrm{d}t > 0$, then taking $\varphi(x) = K_N(x)_+ - K_N(x)$ ensures that the left and right hand sides have different signs;

Note that φ cannot be 0 as it would imply that $K_N(x)$ remains strictly negative (1st case) or strictly positive (2nd case).

Notes

Gauss' method

Let $f \in C^{N+1}[a,b]$ such that $f^{(N+1)} = \varphi$ on $[a,b]$.

Thus combining Peano kernel theorem (4.203) together with equation (**??**) we see that for all $x \in [a,b]$

$$E(f) = \frac{1}{N!}\int_a^b K_N(t) f^{(N+1)}(t)\,\mathrm{d}t \neq \frac{\varphi(x)}{N!}\int_\alpha^\beta K_N(t)\,\mathrm{d}t. \qquad (4.10)$$

Then from theorem 4.196, the error of Gauss' method is of the form

$$E(f) = \frac{\varphi(\xi)}{(N+1)!}\int_\alpha^\beta \Pi(x)w(x)\,\mathrm{d}x. \qquad (4.11)$$

Next see that $\displaystyle\int_\alpha^\beta \Pi(x)w(x)\,\mathrm{d}x$ is precisely $E(x \mapsto x^{N+1})$, which is also $(N+1)\displaystyle\int_a^b K_N(t)\,\mathrm{d}t$ (theorem 4.203).

Notes

213

Gauss' method

Hence (**??**) can be rewritten

$$\begin{aligned} E(f) &= \frac{\varphi(\xi)}{N!}\int_a^b K_N(t)\,\mathrm{d}t \\ &= \frac{\varphi(\xi)}{N!}\left(\underbrace{\int_a^\alpha K_N(t)\,\mathrm{d}t}_{A} + \int_\alpha^\beta K_N(t)\,\mathrm{d}t + \underbrace{\int_\beta^b K_N(t)\,\mathrm{d}t}_{B}\right). \end{aligned}$$

From the definition of Peano kernel, B is 0: if $t \geq \beta$ then for all $x \leq \beta$, $(x-t)_+^N = 0$ implying $K_N(t) = 0$.

From the order of Gauss method, A is 0: if $t \leq \alpha$ then $(x-t)_+^N$ is $(x-t)^N$ and as the method has order N, the formula is exact.

Notes

214

Gauss' method

As a result we have shown that

$$E(f) = \frac{1}{N!}\int_a^b K_N(t)\varphi(t)\,\mathrm{d}t = \frac{\varphi(\xi)}{N!}\int_\alpha^\beta K_N(t)\,\mathrm{d}t.$$

This contradicts (**??**) which should have been valid for all $x \in [a, b]$, including ξ.↯ □

We now refocus our attention on the method of integration through interpolation, and determine the expression of the error. A particularly interesting result states that the error depends on the step of the mesh.

Notes

215

Integration through interpolation

From the general idea behind Peano's method it is clear that integration through interpolation is a special case of this more general method. As such Peano kernel theorem (4.203) holds and we can define both the *elementary* and *composite errors*, corresponding the elementary and composite quadrature formulae.

The following discussion reuses the notations introduced in slides 4.175 to 4.177. If the method is of order $N \geq l \geq 0$, then the elementary error for $\varphi \in C^{N+1}[-1, 1]$ is

$$E_{el}(\varphi) = \int_{-1}^1 \varphi(s)\,\mathrm{d}s - 2\sum_{j=0}^{l} w_j\varphi(\tau_j) = \frac{1}{N!}\int_{-1}^1 k_N(t)\varphi^{(N+1)}(t)\,\mathrm{d}t.$$

As for the composite quadrature formula, if the method is of order $N \geq l$, and $f \in C^{(N+1)}[\alpha, \beta]$, then the error is given by

$$E_{comp}(f) = \int_\alpha^\beta f(x)\,\mathrm{d}x - \sum_{i=0}^{k-1} h_i \sum_{j=0}^{l} w_j f(\alpha_{ij}) = \frac{1}{N!}\int_\alpha^\beta K_N(t) f^{(N+1)}(t)\,\mathrm{d}t.$$

Notes

216

Integration through interpolation

Proposition

Let k_n and K_n denote Peano kernel in the case of the elementary error and composite error, respectively.

For all $t \in [\alpha_j, \alpha_{j+1}]$, $0 \leq j \leq k-1$

$$K_N(t) = \left(\frac{h_j}{2}\right) k_N\left(\frac{2}{h_j}\left(t - \alpha_{j+\frac{1}{2}}\right)\right),$$

with $h_j = \alpha_{j+1} - \alpha_j$ and $\alpha_{j+\frac{1}{2}} = \frac{\alpha_{j+1}+\alpha_j}{2}$.

Proof. Let t be fixed in one of the $[\alpha_j, \alpha_{j+1}]$. By definition

$$K_N(t) = E_{comp}(x \mapsto [(x-t)_+]^N).$$

Notes

217

Integration through interpolation

Based on equation (**??**) this expands as

$$K_N(t) = \int_\alpha^\beta [(x-t)_+]^N \, \mathrm{d}t - \sum_{i=0}^{k-1} \frac{h_i}{2} \cdot 2 \sum_{j=0}^{l} w_j \left[\left(\tau_j \frac{h_i}{2} + \alpha_{i+\frac{1}{2}} - t \right)_+ \right]^N .$$

Then defining $f_t(x) = [(x-t)_+]^N$ and applying the bijective map φ_i (**??**) yield

$$\begin{aligned} K_N(t) &= \sum_{i=0}^{k-1} \frac{h_i}{2} \left[\int_{-1}^{1} \varphi(s) \, \mathrm{d}s - 2 \sum_{j=0}^{l} w_j \varphi_i(\tau_j) \right] \\ &= \sum_{i=0}^{k-1} \frac{h_i}{2} E_{el}(\varphi_i). \end{aligned} \tag{4.12}$$

Notes

218

Integration through interpolation

Therefore we now need to determine $E_{el}(\varphi_i)$.

Starting from the definition of φ_i (**??**) we need to find a simpler form to express $[(s\frac{h_i}{2} + \alpha_{i+\frac{1}{2}} - t)_+]^N$. For this we note that for any positive number λ we have $[(\lambda)_+]^N = \lambda^N$. Hence we can rewrite

$$\left[\left(s\frac{h_i}{2} + \alpha_{i+\frac{1}{2}} - t \right)_+ \right]^N = \left(\frac{h_i}{2} \right)^N \left[\left(s - \frac{2}{h_i}(t - \alpha_{i+\frac{1}{2}}) \right)_+ \right]^N .$$

Then setting $\theta_i = \frac{2}{h_i}(t - \alpha_{i+\frac{1}{2}})$ implies

$$\left[\left(s\frac{h_i}{2} + \alpha_{i+\frac{1}{2}} - t \right)_+ \right]^N = \left(\frac{h_i}{2} \right)^N \left[(s - \theta_i)_+ \right]^N .$$

Notes

219

Integration through interpolation

As a result equation (**??**) can be rewritten as

$$K_N(t) = \sum_{i=0}^{k-1} \frac{h_i}{2} \left(\left(\frac{h_i}{2} \right)^N E_{el}(s \mapsto [(s-\theta_i)_+]^N \right) .$$

But as k_N is by definition $E_{el}(s \mapsto [(s-t)_+]^N)$ we obtain

$$K_N(t) = \sum_{i=0}^{k-1} \left(\frac{h_i}{2} \right)^{N+1} k_N(\theta_i). \tag{4.13}$$

To conclude the proof observe that for a fixed t, there is a unique j such that $t \in [\alpha_j, \alpha_{j+1}]$. So in the sum (**??**), $K_N(\theta_i)$ is to be calculated when $i = j$, i.e. $\theta_i \in [-1, 1]$, and when $i \neq j$, i.e. $\theta_i \notin [-1, 1]$.

Notes

220

Integration through interpolation

If $\theta_i \notin [-1, 1]$ then it means it is either smaller than -1 or lager than 1.

In the case where $\theta_i < -1$, $[(s-\theta_i)_+]^N = (s-\theta_i)^N$ is a polynomial of degree N in s. But since the method is of order N $k_N(\theta_i) = 0$. In the case where $\theta_i > 1$ then $[(s-\theta_i)_+]^N = 0$ and $k_N(\theta_i)$ is null.

Hence the only component left from sum (**??**) is when $i = j$ and we have

$$K_N(t) = \left(\frac{h_j}{2} \right) k_N \left(\frac{2}{h_j} \left(t - \alpha_{j+\frac{1}{2}} \right) \right) .$$

□

This result is significant as it highlights the relation between the steps of the mesh and the expression of the error.

Notes

Integration through interpolation

We now provide a more precise result on the expression of the composite error in the general case where the step of the mesh is not necessarily constant.

Proposition

If $f \in C^{N+1}[\alpha,\beta]$ and the method has order N, then

$$\left|E_{comp}(f)\right| \leq \frac{C_{N,f}}{2^{N+2}}(\beta-\alpha)h^{N+1},$$

where

$$\begin{cases} h = \max_{0\leq i\leq k-1} h_i \\ C_{N,f} = \frac{\max_{\alpha,\beta}|f^{(N+1)}|}{N!}\int_{-1}^{1}|k_N(t)|\,\mathrm{d}t. \end{cases}$$

Notes

Integration through interpolation

Proof. Following the same reasoning as the one leading to (**??**) we can easily deduce for any $f \in C^{N+1}[\alpha,\beta]$

$$E_{comp}(f) = \sum_{i=0}^{k-1}\frac{h_i}{2}E_{el}(\varphi_{i,f}), \quad \text{with } \varphi_{i,f}(s) = f\left(s\frac{h_i}{2}+\alpha_{i+\frac{1}{2}}\right). \tag{4.14}$$

Noting that $\varphi_{i,f}^{(N+1)}(s) = \left(\frac{h_i}{2}\right)^{N+1} f^{(N+1)}\left(s\frac{h_i}{2}+\alpha_{i+\frac{1}{2}}\right)$ we get

$$E_{el}(\varphi_{i,f}) = \frac{1}{N!}\int_{-1}^{1}k_N(t)\varphi_{i,f}^{(N+1)}\,\mathrm{d}t,$$

which yields

$$|E_{el}(\varphi_{i,f})| \leq \frac{\max_{[\alpha,\beta]}|f^{(N+1)}|}{N!}\left(\frac{h_i}{2}\right)^{N+1}\int_{-1}^{1}|k_N(t)|k_N\,\mathrm{d}t.$$

Notes

Integration through interpolation

If we now call h the max of all the h_i, $0 \leq i \leq k-1$ and observe that their sum is exactly $\beta-\alpha$, then we obtain the expected result

$$\begin{aligned} \left|E_{comp}(f)\right| &\leq \frac{\max_{[\alpha,\beta]}|f^{(N+1)}|}{N!}\left(\int_{-1}^{1}|k_N(t)|\,\mathrm{d}t\frac{h^{N+1}}{2^{N+2}}\sum_{i=0}^{k-1}h_i\right) \\ &= \frac{h^{N+1}(\beta-\alpha)\max_{[\alpha,\beta]}|f^{(N+1)}|}{2^{N+2}N!}\left(\int_{-1}^{1}|k_N(t)|\,\mathrm{d}t\right) \end{aligned}$$

□

The last result we prove describes the error in the case where Peano kernel does not change sign and the step is constant. As in the previous proofs the constant sign will allow us to apply the mean value theorem to obtain a more precise result than the previous maximum.

Notes

Integration through interpolation

Proposition

Let $f \in C^{N+1}[\alpha,\beta]$. If k_N has a constant sign over $[-1,1]$, and the mesh has constant step $h = (\beta-\alpha)/k$, where k is the number of nodes, then there exists $\xi \in [\alpha,\beta]$ such that

$$E_{comp}(f) = \frac{(\beta-\alpha)h^{N+1}}{2^{N+2}N!}f^{(N+1)}(\xi)\int_{-1}^{1}k_N(t)\,\mathrm{d}t.$$

Proof. Let $f \in C^{N+1}[\alpha,\beta]$. Starting from equation (**??**) we write

$$E_{el}(\varphi_{i,f}) = \frac{1}{N!}\left(\frac{h}{2}\right)^{N+1}\int_{-1}^{1}k_N(t)f^{(N+1)}\left(t\frac{h}{2}+\alpha_{i+\frac{1}{2}}\right)\mathrm{d}t.$$

Notes

225

Integration through interpolation

From the first mean value theorem (2.99) we know the existence of ξ_i in $[\alpha_i, \alpha_{i+1}]$ such that

$$\int_{-1}^{1} k_N(t) f^{(N+1)}\left(t\frac{h}{2} + \alpha_{i+\frac{1}{2}}\right) \mathrm{d}t = f^{(N+1)}(\xi_i) \int_{-1}^{1} k_N(t)\,\mathrm{d}t.$$

Then the composite error (**??**) can be expressed as

$$E_{comp}(f) = \frac{h^{N+1}}{2^{N+2}N!}\,\frac{\beta-\alpha}{k}\left[\sum_{i=0}^{k-1} f^{(N+1)}(\xi_i)\right]\int_{-1}^{1} k_N(t)\,\mathrm{d}t. \tag{4.15}$$

In order to rewrite the sum observe that

$$\sum_{i=0}^{k-1} \min_{[\alpha,\beta]} f^{(N+1)} \leq \sum_{i=0}^{k-1} f^{(N+1)}(\xi_i) \leq \sum_{i=0}^{k-1} \max_{[\alpha,\beta]} f^{(N+1)}.$$

Notes

226

Integration through interpolation

Looking at the left and right sums we directly see that

$$\min_{[\alpha,\beta]} f^{(N+1)} \leq \frac{1}{k}\sum_{i=0}^{k-1} f^{(N+1)}(\xi_i) \leq \max_{[\alpha,\beta]} f^{(N+1)}.$$

Therefore from the intermediate values theorem (2.76) there exists $\xi \in [\alpha, \beta]$ such that $\frac{1}{k}\sum_{i=0}^{k-1} f^{(N+1)}(\xi_i) = f^{(N+1)}(\xi)$.

Finally (**??**) can be rewritten

$$E_{comp}(f) = \frac{h^{N+1} f^{(N+1)}(\xi)}{2^{N+2}N!}(\beta - \alpha)\int_{-1}^{1} k_N(t)\,\mathrm{d}t.$$

□

Notes

227

Convergence of Peano's method

Let $(x_{jk})_{0\leq j\leq k}$ be a sequence of points in an interval $[a,b] \subset \mathbb{R}$. Given $(\alpha,\beta) \subset [a,b]$, and a weight function $w : (\alpha,\beta) \to \mathbb{R}^+$ we consider the quadrature formula, for a function $f \in C[a,b]$,

$$\int_{\alpha}^{\beta} f(x)w(x)\,\mathrm{d}x \approx \sum_{j=0}^{k} \lambda_{jk} f(x_{jk}),$$

and define its error

$$E_k(f) = \int_{\alpha}^{\beta} f(t)w(t)\,\mathrm{d}t - \sum_{j=0}^{k} \lambda_{jk} f(x_{jk}).$$

We now introduce a result which provides necessary and sufficient conditions for the error to converge, i.e. $\lim_{k\to\infty} E_k(f) = 0$.

Notes

228

Polya's convergence theorem

Theorem (Polya's convergence)

Using the previous notations, the error $E_k(f)$ converges to 0 as k tends to infinity if and only if the following two conditions are met

- i For any positive integer n, $\lim_{k\to\infty} E_k(x \mapsto x^n) = 0$;
- ii There exists M, strictly positive, such that for all $k \in \mathbb{N}$ $\sum_{j=0}^{k} |\lambda_{jk}| \leq M$;

Proof. Let f be a continuous function over $[a,b]$.

We first prove that if the two conditions hold then $\lim\limits_{k\to\infty} E_k(f) = 0$.

By Stone-Weierstrass theorem (2.87) there exists a sequence of polynomials $(P_n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} \|f - P_n\|_{\infty} = \lim_{n\to\infty} \max_{x\in[a,b]} |f(x) - P_n(x)| = 0.$$

Notes

229

Polya's convergence theorem

As $C[a, b]$ and $\mathbb{R}$ are vector spaces while E_k is a linear application we would like to apply proposition 1.43. We therefore need to prove that $\|E_k\|$ is finite.

This is straightforward. Since $\int_\alpha^\beta f(t)w(t)\,dt$ is finite, all the $f(x_{jk})$ are finite, and from (ii), there exists $M > 0$ such that for all $k \in \mathbb{N}$, $\sum_{j=0}^{k} |\lambda_{jk}| \leq M$. Thus $E_k(P_n)$ converges to $E_k(f)$.

Now from our first condition we know that $E_k(P_n)$ tends to 0 as k tends to infinity. This yields the first implication.

For the converse, (i) is trivial since it suffices to take $f(x) = x^n$. Therefore we only need to prove (ii). For (ii) first recall that on a compact a convergent sequence is bounded (sketch of proof of Bolzano-Weierstrass (2.81)). Thus we can apply theorem 1.47 to conclude that the linear map E_k is continuous.

Notes

230

Polya's convergence theorem

Hence $(E_k)_{k\geq 0}$ is a sequence of continuous linear maps from the Banach space $C[a, b]$ into the Banach space $\mathbb{R}$. Thus we can apply Banach-Steinhaus theorem (2.85) to obtain the existence of some m such that $\sup_k \|E_k\|$ is less than m.

Let g_k be a continuous function over $[a, b]$, with $\max_{[a,b]} |g_k| = 1$ and $g_k(\lambda_{jk}) = \operatorname{sign}(\lambda_{jk})$. Then for all k we see that

$$\begin{aligned}\sum_{j=0}^{k} |\lambda_{jk}| &= \int_\alpha^\beta g_k(x)w(x)\,dx - E_k(g_k)\\ &\leq \int_\alpha^\beta w(x)\,dx + m.\end{aligned}$$

Choosing $M = \int_\alpha^\beta w(x)\,dx + m$ completes the proof. □

Notes

231

Polya's convergence theorem

Example. Lets consider the Gauss-Chebyshev integration method, i.e. Gauss method together with Chebyshev nodes.

The weight function associated to Chebyshev polynomials is defined on $(-1, 1)$ by $w(x) = \frac{1}{\sqrt{1-x^2}}$, and for $0 \leq j \leq k$, $k \in \mathbb{N}$,

$$x_{jk} = \cos\frac{(2j+1)\pi}{2(k+1)}, \qquad \lambda_{jk} = \frac{\pi}{k+1}.$$

All the λ_{jk} being positive, $\sum_{j=0}^{k} |\lambda_{jk}| = \sum_{j=0}^{k} \lambda_{jk} = \pi$. So the second condition of Polya's convergence theorem is met. The first one is also verified since Gauss method is of order $2k + 1$ (corollary 4.194): for all $1 \leq n \leq 2k+1$ the formula is exact, and when k tends to infinity $E_k(x \mapsto x^n)$ tends to 0.

By Polya's convergence theorem (4.228), the Gauss-Chebyshev method is convergent.

Notes

232

Key points

- What is the order of an integration method?
- Is interpolation useful for numerical integration?
- Describe Peano's method
- Why does Peano's method covers other methods?
- What does it mean for a method to be convergent?

Notes

Thank you, enjoy the Winter break!

Notes