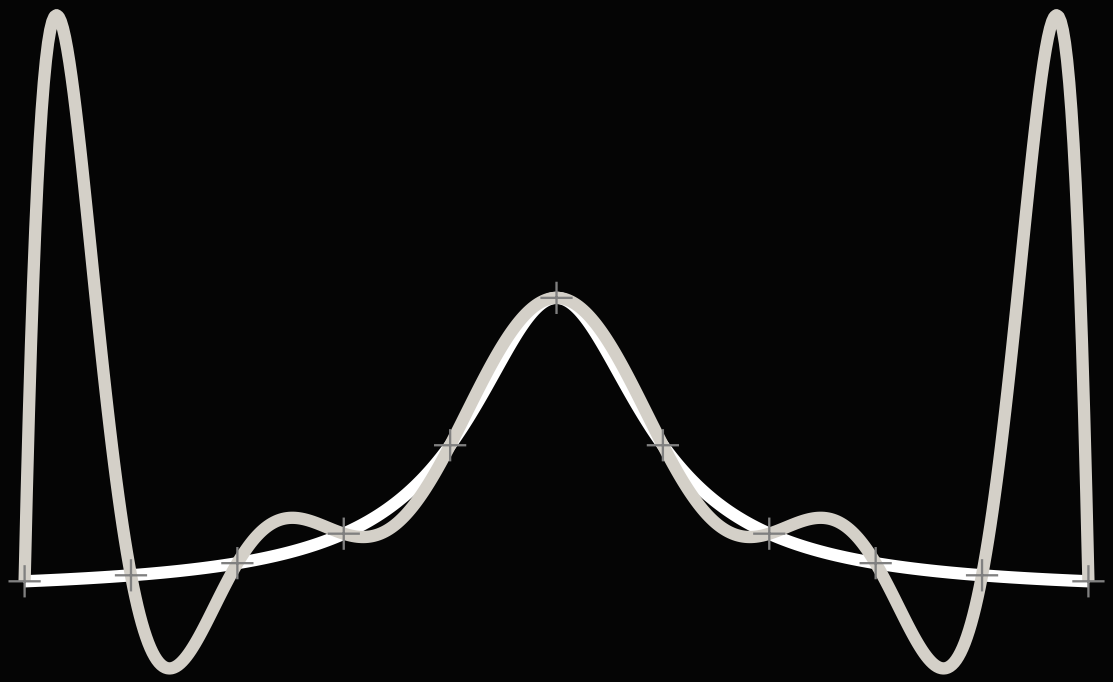
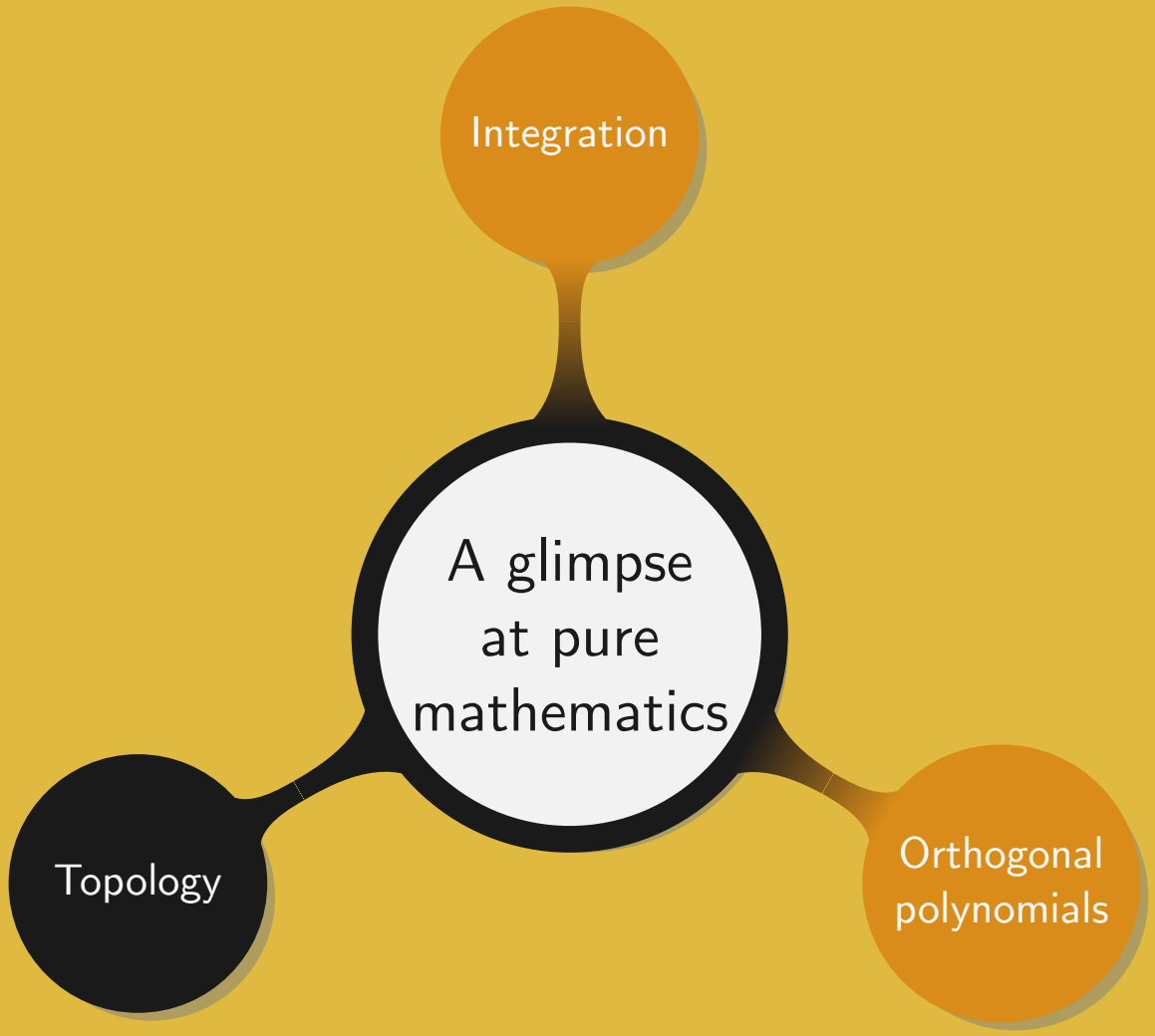


Introduction to Numerical Methods

2. A glimpse at pure mathematics

Ailin & Manuel – Fall 2025

Chapter organisation



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.

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.

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.

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 2.4 apply to metric 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 2.6).

(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.

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. $\square$

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

(i) $\Rightarrow$ (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[1]. Hence U is a neighborhood for all its points, i.e. U is open.

(ii) $\Rightarrow$ (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.[1] $\square$

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

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. $\square$

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 $U = \{(-n, n), n \geq 1\}$ 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.

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 $f(A)$. □

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 2.13 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 $f(A)$ $\square$

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 (2.14) and A is bounded. □

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

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. □

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 2.15 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 $[a, b]$, $a, b \in \mathbb{R}$.

(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 $f^{(n)}(c) = 0$.

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

(i) Case $n = 1$: apply the extreme value theorem (2.14) 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.

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. $\square$

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
   i. $\mathcal{H}$ is a vector-subspace of $\mathcal{F}(\mathcal{E}, \mathbb{K})$;
   ii. The constant applications are in $\mathcal{H}$;
   iii. 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 $\mathcal{F}(\mathcal{E}, X)$.

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.

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 $\mathcal{H}$.

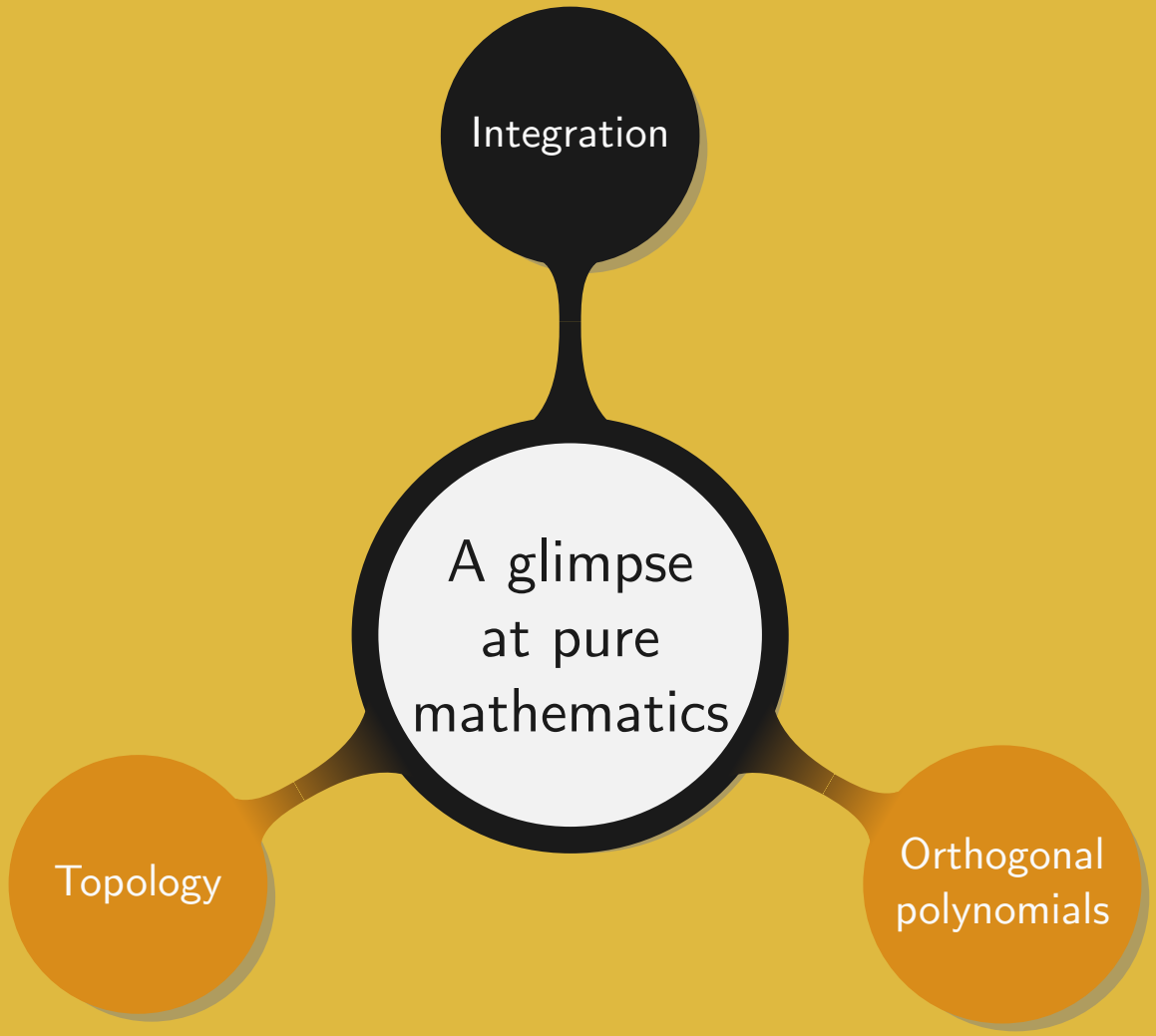
Finally note that the two most important conditions for the theorem to be correct are the compactness of X and $C(X, \mathbb{R})$ being separable. □

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 $C([a, b])$. □

Chapter organisation



Definition

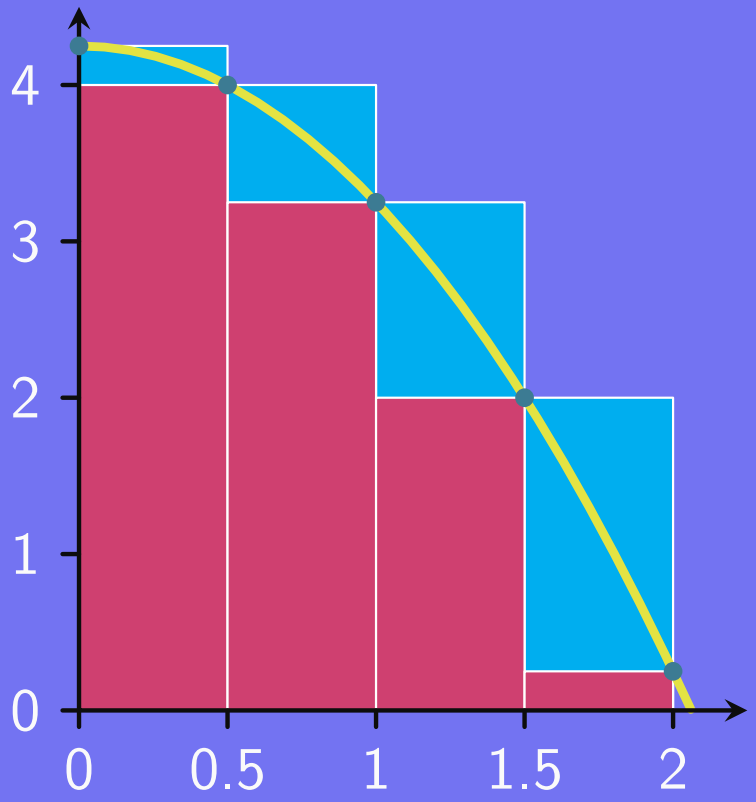
Let $I = [a, b]$ be an interval in $\mathbb{R}$ and $f : I \to \mathbb{R}$, be continuous. Partitioning I using $n + 1$ nodes $a = x_0 < x_1 < \cdots < x_n = b$ the *Riemann sum* of f is defined as

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

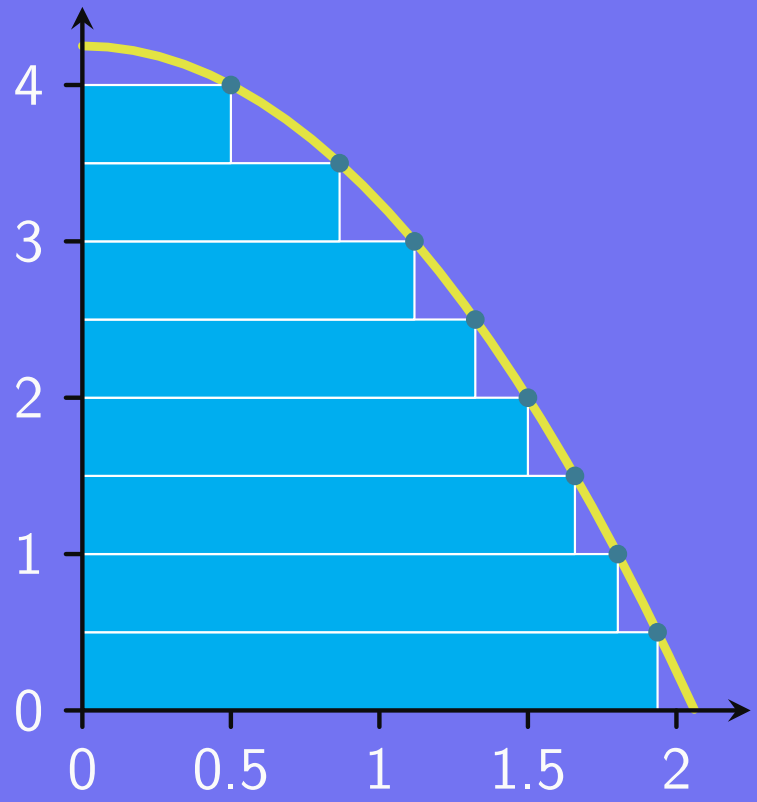
for $\xi_i \in [x_{i-1}, x_i]$. Each choice of ξ_i defines a different Riemann sum.

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.

Riemann vs. Lebesgue integral



Riemann integral



Lebesgue integral

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*

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;

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

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

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 $A_i \in \mathcal{A}$, $\mu(\bigcup_i A_i) = \sum_{i=0}^{\infty} \mu(A_i)$;

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.

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 $\int \mathbb{1}_A \, \mathrm{d}\mu = \mu(A)$, 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 $a_i \in \mathbb{R}$
$$\int \left(\sum_i a_i \mathbb{1}_{A_i} \, \mathrm{d}\mu \right) = \sum_i a_i \mu(A_i).$$
3. A monotonic increasing sequence of positive simple functions converges pointwise, or uniformly if f is bounded, to f. Thus
$$\int_\Omega f \, \mathrm{d}\mu = \sup \left\{ \int_\Omega f_n \, \mathrm{d}\mu : 0 \leq f_n \leq f \text{ and } f_n \text{ simple} \right\}.$$

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 2.26 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.

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.

(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

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

Proof. By the extreme value theorem (2.14) 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

$$m \leq \underbrace{\frac{1}{\int_a^b g(x)\,\mathrm{d}x}\int_a^b f(x)g(x)\,\mathrm{d}x}_{A} \leq M.$$

Then by the intermediate value theorem (2.11), 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 $[a, x] \subset \mathbb{R}$,

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

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;

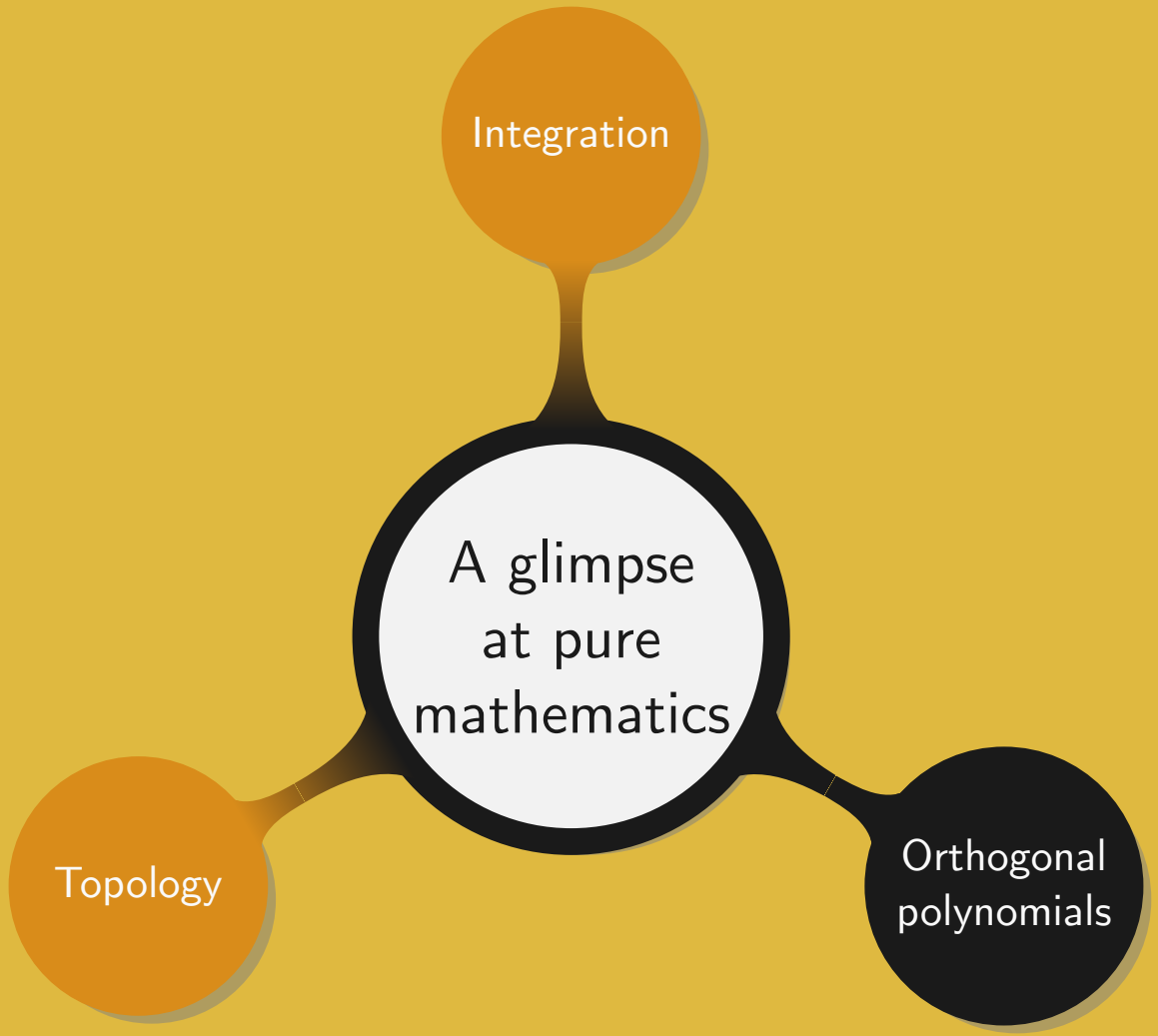
Weight function

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.

Chapter organisation



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 1.17 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 $v \in (\mathbb{R}^d, \|\cdot\|_1)$

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

Projection on an inner product space

As this shows that v is continuous, we can apply theorem 1.28 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 (2.17). As v is continuous, $B'(a,r)$ is compact (theorem 2.13) and by Bolzano-Weierstrass (2.16) 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.

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 (2.16).
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

$$\left| \|f_n\| - \|e\| \right| \leq \|f_n - e\|$$

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 (2.16) to extract a subsequence $(f_{n_k})_{k \geq 0}$ that converges to $\widetilde{f}$.
Then we have

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

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. 1.22)

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

Rearranging the terms and using m yields

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

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

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

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 $f \in F$, $\langle e - P_F(e), f \rangle = 0$.

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 $\lambda \in \mathbb{R}$,

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

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

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

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

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

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

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

□

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

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

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 1.18. Again this is clear as by definition of a weight function, $w(x)$ is strictly positive over (a, b).

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 2.40 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*.

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

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

By the characterisation of the projection (theorem 2.40), 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.

Proof. Let F_k be $\text{span}(e_1, \cdots, e_k)$. For $k = 1$ we simply take $\varphi_1 = e_1$.

Assume $(\varphi_1, \cdots, \varphi_k)$ has been constructed and $F_k = \text{span}(\varphi_1, \cdots, \varphi_k)$, 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 $F_{k+1} = \text{span}(\varphi_1, \cdots, \varphi_k, e_{k+1})$, 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.

Construction of orthogonal polynomials

Let $E = C[a, b]$. From example 2.47, 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 $P_i \in F$, $0 \leq i \leq n$, such that

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

Definition

The polynomials P_i, $0 \leq i \leq n$, such that (**??**) is verified are called *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 $n \geq 1$;
4. For any $n \geq 1$ 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);

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, $0 \leq i \leq n-1$. This yields the result.

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 $k \leq n-2$, $c_k = 0$. 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 $n+1$.

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 $\lambda_nP_n + \mu_{n-1}P_{n-1}$ 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 $\mu_{n-1} > 0$.

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, $0 \leq i \leq r$. 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 $P_n(x) = \prod_{j=0}^{r}(x - x_j)^{\alpha_j}$.

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). □

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 $\cos\left(\frac{2k-1}{2n}\pi\right)$, $1 \leq k \leq n$;

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 $\cos[(n+1)\theta] + \cos[(n-1)\theta] = 2\cos(\theta)\cos(n\theta)$.

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 $n \neq m$

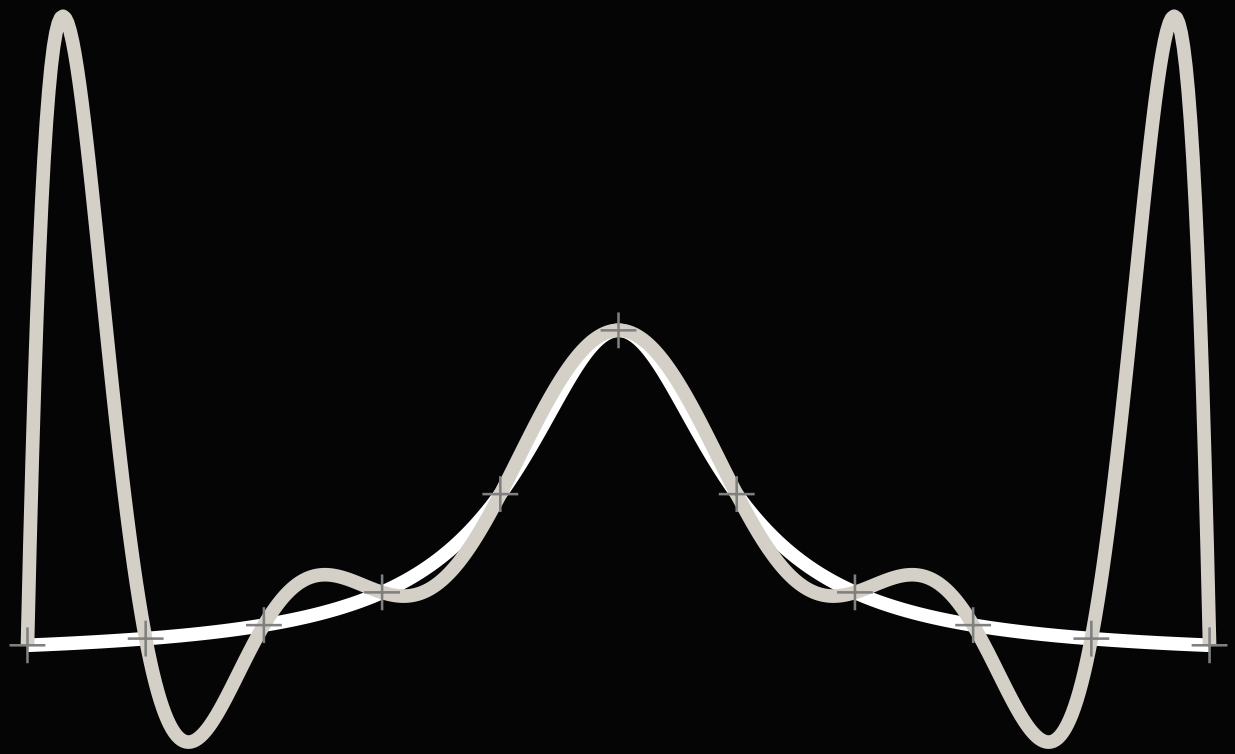
$$\begin{aligned} \langle T_n, T_m \rangle &= \int_{-1}^{1} \cos(n\arccos x)\cos(m\arccos x)\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x \\ &= \frac{1}{2}\int_0^{\pi} \cos[(n+m)x] + \cos[(n-m)x]\,\mathrm{d}x = 0. \end{aligned}$$

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 $1 \leq k \leq n$. □

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?



Thank you!