Semester I

Subject 1

ME010101 Abstract Algebra

1.1 Introduction to Abstract Algebra

Definitions 1.1.1. Set-theoretical foundation of Abstract Algebra

- A **cartesian product**, $A \times B = \{(a, b) \ : \ a \in A, \ b \in B\}$
- A **relation** on a set A is a subset of $A \times A$.
- A **function** f from A into B, $f : A \to B$ is a relation such that *every element of* A *is related to some unique element of* B (well defined).

Definitions 1.1.2. Functions : A **binary operation** on a set A is a function $* : A \times A \to A$.

"A binary operation on a set A gives an algebra on A."

Abstract Algebra It is the study of algebraic structures. We are interested in a few algebraic structures : 1. Group 2. Ring 3. Integral Domain 4. Field

Definitions 1.1.3. A **binary algebraic structure** $\langle G, * \rangle$ is a set G together with a binary operation $*$.

Definitions 1.1.4 (Group)**.** A set G, closed under a binary relation $*$ satisfying the following three axioms -G1, G2, & G3 is a group.

G1 Associativity
For any three elements $a, b, c \in G$, $a * (b * c) = (a * b) * c$.

G2 Identity element
There exists a unique element $e \in G$ such that $a * e = a = e * a$ for any element $a \in G$.

G3 Inverse elements
For any element $a \in G$ there exists a unique element a^{-1} such that $a * a^{-1} = e = a^{-1} * a$.

Definitions 1.1.5. Group terminologies :

- A group is **abelian** if the binary operation is commutative. $a * b = b * a$
- The **order** of a group $G, *$ is the number of elements in G.

Definitions 1.1.6. $\mathbb{Z}_n = \{0, 1, 2, \cdots, n-1\}$

Remark. Consider $3, 4 \in \mathbb{Z}_5$, $3 * 4 = 2 = 4 * 3$ since $7 \cong 2 \pmod 5$. $\langle \mathbb{Z}_5, +_5 \rangle$ is an abelian group of order 5.

Definitions 1.1.7. Homomorphism & Isomorphism

- A function $\phi : A \to B$ is a **homomorphism** if for any two elements $x, y \in A$, $\phi(xy) = \phi(x)\phi(y)$
- A function $\phi : A \to B$ is an **isomorphism** if ϕ is a bijective, homomorphism. If two binary structures are isomorphic, then they have the same (algebraic) structure.

Definitions 1.1.8. Subgroup

- A subset H of a group $\langle G, * \rangle$ is a **subgroup** of G if H is group with the same binary operation $*$. And is denoted by $H \leq G$.
- G is the **improper** subgroup of G and every other subgroup is **proper**.S5.5
- $\{e\}$ is the **trivial** subgroup of G and every other subgroup is **non-trivial**.
- The **subgroup generated by** $g \in G$ is the subgroup $\{g^n : n \in \mathbb{Z}\}$.
- The **order** of an element g is order of the subgroup generated by g.
- An element $g \in G$ is a **generator** of G if g generates G.
- A group is **cyclic** if it has a generator.

Remark. Cyclic Groups :

- Cyclic groups are abelian.
- Subgroups of cyclic groups are cyclic.

1.1.1 Some Proof Techniques

Equality of two Sets $A = B \iff A \subset B$ and $B \subset A$
If $x \in A \implies x \in B$, then $A \subset B$

Uniqueness Suppose there are two elements that qualify our conditions. We show (using the conditions) that they are the same, that is, unique.

For example, $3 + a = \pi$. Suppose $a = x, y$. Then $3 + x = 3 + y \implies x = y$, provided that the values of a comes from a set in which left cancelation law can be applied. Then a is unique.

Remember : We usually don't care to show what this unique element is. It may be also be the case that there is no such element, that is, proof of uniqueness doesn't imply existence.

Existence There are constructive and non-constructive proofs for existence problems. Suppose we want to prove that $a*b$ has an inverse element. We know that a, b has inverse elements a^{-1}, b^{-1}. From those elements, we construct an element $b^{-1}*a^{-1}$ which is an inverse of $a*b$ by construction.

And we may also prove existence without actually giving an object. Suppose we want to prove that $x^y \in \mathbb{Q}$ for some irrational numbers x and y. We know that $\sqrt{2} \notin \mathbb{Q}$. Then, $\sqrt{2}^{\sqrt{2}}$ is either rational or irrational. Suppose it is irrational, then $\left(\sqrt{2}^{\sqrt{2}}\right)^{\sqrt{2}} = 2$ is rational. Thus the proof is complete, but we are yet to know whether $\sqrt{2}^{\sqrt{2}}$ is an irrational or rational.

1.2 Direct Products and Finitely Generated Abelian Groups

Definitions 1.2.1 (Cartesian product of sets)**.** Let $S_1,\ S_2, \cdots,\ S_n$ be a sets. Their cartesian product,

$$S_1 \times S_2 \times \cdots \times S_n = \prod_{i=1}^{n} S_n = \{(a_1, a_2, \cdots, a_n) : a_i \in S_i\} \tag{1.1}$$

For example, $\{A, B, C\} \times \{1, 2\} = \{(A, 1), (A, 2), (B, 1), (B, 2), (C, 1), (C, 2)\}$.

Question 1. How many elements in $\mathbb{Z}_3 \times \mathbb{Z}_{10} \times \mathbb{Z}_9$?

Theorem 1.2.1 (Direct product of Groups)**.** *Let $G_1,\ G_2,\ \cdots,\ G_n$ be groups. Then their cartesian product is a group with the binary operation $*$,*

$$(a_1,\ a_2,\ \cdots,\ a_n) * (b_1,\ b_2,\ \cdots,\ b_n) = (a_1 * b_1,\ a_2 * b_2,\ \cdots,\ a_n * b_n) \tag{1.2}$$

where the binary operation in $a_i * b_i$ is the binary operation of the group G_i.

Proof. $\prod_{i=1}^{n} G_i$ is a group if it satisfies the group axioms.

G1 Associativity

$$\begin{aligned}
&(a_1,\ a_2,\ \cdots,\ a_n)\big((b_1,\ b_2,\ \cdots,\ b_n)(c_1,\ c_2,\ \cdots,\ c_n)\big)\\
&\quad=(a_1,\ a_2,\ \cdots,\ a_n)(b_1c_1,\ b_2c_2,\ \cdots,\ b_nc_n)\\
&\quad=\big(a_1(b_1c_1),\ a_2(b_2c_2),\ \cdots,\ a_n(b_nc_n)\big)\\
&\quad=\big((a_1b_1)c_1,\ (a_2b_2)c_2,\ \cdots,\ (a_nb_n)c_n\big)\\
&\quad=(a_1b_1,\ a_2b_2,\ \cdots,\ a_nb_n)(c_1,\ c_2,\ \cdots,\ c_n)\\
&\quad=\big((a_1,\ a_2,\ \cdots,\ a_n)(b_1,\ b_2,\ \cdots,\ b_n)\big)(c_1,\ c_2,\ \cdots,\ c_n)
\end{aligned}$$

G2 Existence of a unique identity element in $\prod_{i=1}^{n} G_i$

Let e_i be the identity element in G_i. Then $(e_1,\ e_2,\ \cdots,\ e_n)$ is the identity element in $\prod_{i=1}^{n} G_i$.

$$\begin{aligned}(a_1,\ a_2,\ \cdots,\ a_n)&(e_1,\ e_2,\ \cdots, e_n)\\ &= (a_1e_1,\ a_2e_2,\ \cdots,\ a_ne_n)\\ &= (a_1,\ a_2,\ \cdots,\ a_n)\end{aligned}$$

G3 Existence of unique inverse element for each element in $\prod_{i=1}^{n} G_i$

Let $(a_1,\ a_2,\ \cdots, a_n)$ be in $\prod_{i=1}^{n} G_i$. Then it has the inverse element $(a_1^{-1},\ a_2^{-1},\ \cdots,\ a_n^{-1})$ in $\prod_{i=1}^{n} G_i$.

$$\begin{aligned}(a_1,\ a_2,\ \cdots a_n)&(a_1^{-1},\ a_2^{-1},\ \cdots, a_n^{-1})\\ &= (a_1a_1^{-1},\ a_2a_2^{-1},\ \cdots,\ a_na_n^{-1})\\ &= (e_1,\ e_2,\ \cdots,\ e_n)\end{aligned}$$

□

Remark. We usually write ab instead of $a * b$ and relevent binary operations are used in different contexts. Student should be able to recongnise the difference from the context.

Remark. $\mathbb{Z}_n = \{0, 1, \cdots, (n-1)\}$ is a group with $+_n$. (addition modulo n)

For example, Consider $(1, 2) \in \mathbb{Z}_2 \times \mathbb{Z}_3$. We have, $(1, 2) + (1, 2) = (0, 1)$ since $1 +_2 1 = 0$ and $2 +_3 2 = 1$.

Definitions 1.2.2. Suppose all the groups G_i are abelian. Then $\prod_{i=1}^{n} G_i$ is the direct sum of the groups G_i. And is represented by $\oplus_{i=1}^{n} G_i$.

Theorem 1.2.2. *$\mathbb{Z}_m \times \mathbb{Z}_n$ is cyclic and is isomorphic to $\mathbb{Z}_{mn}$ if and only if m and n are relatively prime.*

Proof. Sufficient part : Consider the cyclic subgroup[1] H generated by $(1, 1) \in \mathbb{Z}_m \times \mathbb{Z}_n$. It is enough to prove that the order of this cyclic subgroup H is mn. The order of H is the smallest power of $(1, 1)$ that gives the identity $(0, 0)$. The first component gives 0 for multiples of m. And the second component gives 0 for multiples of n. Since m, n are relatively prime, mn is the smallest power of $(1, 1)$ that will give $(0, 0)$. Thus $\mathbb{Z}_m \times \mathbb{Z}_n = H$ and is cyclic. Every cyclic group of order mn is isomorphic to $\mathbb{Z}_{mn}$. Therefore, $\mathbb{Z}_m \times \mathbb{Z}_n \approx \mathbb{Z}_{mn}$.

Necessary part : Suppose $\gcd(m, n) = d > 1$. Then mn/d is the smallest integer divisible by both m and n. Consider $(r, s) \in \mathbb{Z}_m \times \mathbb{Z}_n$. r gives 0 in mn/d since it is a multiple of m. Similarly, s gives 0 in mn/d since it is a multiple of n. Thus, $\frac{mn}{d}(r, s) = (0, 0)$. And the cyclic group generated by any element of $\mathbb{Z}_m \times \mathbb{Z}_n$ is a proper subgroup. Therefore $\mathbb{Z}_m \times \mathbb{Z}_n$ has no generators and it is not cyclic. □

[1] $(1, 1) \in Z_m \times Z_n$. The cyclic group generated by $(1, 1)$ has all its elements in $Z_m \times Z_n$. And therefore, it is a subgroup of $Z_m \times Z_n$

Corollary 1.2.2.1. *$\prod_{i=1}^{n} \mathbb{Z}_{m_i}$ is cyclic and is isomorphic to $Z_{m_1 m_2 \cdots m_n}$ if and only if any two of the numbers m_i are relatively prime.*

Question 2. Prove : For any non-negative integer n, there exists a cyclic group of order n, which is unique upto isomorphism.

Theorem 1.2.3. *Let $(a_1,\ a_2,\ \cdots,\ a_n) \in \prod_{i=1}^{n} G_i$. And a_i are of finite order r_i in G_i. Then the order of $\prod_{i=1}^{n} G_i$ is the least common multiple of r_is.*

Proof. Least common multiple of r_is is the smallest positive integer d which is a multiple of all r_is. For each i, the r_ith multiple of a_i gives 0 (identity). Thus, the order of the cyclic subgroup generated by $(a_1,\ a_2,\ \cdots,\ a_n)$ is the least common multiple of all the r_is. $\square$

Remark. Consider $(3, 6, 12, 16) \in \mathbb{Z}_4 \times \mathbb{Z}_{12} \times \mathbb{Z}_{20} \times \mathbb{Z}_{24}$. Order of $3 \in \mathbb{Z}_4$ is $4/\gcd(3, 4) = 4$ ie, $\langle 3 \rangle = \{3, 2, 1, 0\}$
Order of $6 \in \mathbb{Z}_{12}$ is $12/\gcd(6, 12) = 2$ ie, $\langle 6 \rangle = \{6, 0\}$
Order of $12 \in \mathbb{Z}_{20}$ is $20/\gcd(12, 20) = 5$ ie, $\langle 12 \rangle = \{12, 4, 16, 8, 0\}$
Order of $16 \in \mathbb{Z}_{24}$ is $24/\gcd(16, 24) = 3$ ie, $\langle 16 \rangle = \{16, 8, 0\}$
Order of $(3, 6, 12, 16)$ is $lcm(4, 2, 5, 3) = 2^2\ 3\ 5 = 60$.

Remark. Define $\overline{G_i} = \{(e_1,\ e_2,\ \cdots,\ e_{i-1},\ a_i,\ e_{i+1},\ \cdots,\ e_n) : a_i \in G_i\}$. Then $G_i \approx \overline{G_i}$. And $\prod_{i=1}^{n} G_i$ is the internal direct product of $\overline{G_i}$s.

For example, $\mathbb{Z}_2 \times \mathbb{Z}_3 \approx (\mathbb{Z}_2 \times \{0\}) \otimes (\{0\} \times \mathbb{Z}_3)$

Question 3. Internal direct product form of $\mathbb{Z}_{12} \times \mathbb{Z}_{60} \times \mathbb{Z}_{24}$?

1.3 Fundamental Theorem

Definitions 1.3.1. A group G is **finitely generated** if G has a finite subset that generates G.

Theorem 1.3.1 (fundamental theorem of finitely generated abelian groups)**.** *Every finitely generated abelian group G is isomorphic to a direct product of cyclic groups in the form*

$$\mathbb{Z}_{{p_1}^{r_1}} \times \mathbb{Z}_{{p_2}^{r_2}} \times \cdots \times \mathbb{Z}_{{p_n}^{r_n}} \times \mathbb{Z} \times \mathbb{Z} \times \cdots \times \mathbb{Z} \tag{1.3}$$

where p_i are primes, not necessarily distict and r_i are positive integers. The direct product is unique, except for the possible rearrangement of the factors.

Proof. —proof is omitted— $\square$

For example, $G = \mathbb{Z}_{20} \times \mathbb{Z} \times \mathbb{Z}_{15} \times \mathbb{Z} \approx \mathbb{Z}_{2^2} \times \mathbb{Z}_3 \times \mathbb{Z}_5 \times \mathbb{Z}_5 \times \mathbb{Z} \times \mathbb{Z}$. In the above case, Betti number of G is 2 (number of $\mathbb{Z}$ factors). For any finite abelian group, Betti number is 0.

Remark (finite abelian groups)*.* Every finite group is finitely generated. And thus we can enumerate finite abelian group of any order.

Remark. There are precisely 6 different abelian groups of order $360 = 2^3 3^2 5$.

1. $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_5$
2. $\mathbb{Z}_2 \times \mathbb{Z}_{2^2} \times \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_5$
3. $\mathbb{Z}_{2^3} \times \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_5$
4. $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_{3^2} \times \mathbb{Z}_5$
5. $\mathbb{Z}_2 \times \mathbb{Z}_{2^2} \times \mathbb{Z}_{3^2} \times \mathbb{Z}_5$
6. $\mathbb{Z}_{2^3} \times \mathbb{Z}_{3^2} \times \mathbb{Z}_5$

Question 4. Group of order 360 with at least an element of order 8

Definitions 1.3.2. A group G is **decomposible** if it is isomorphic to a direct product of two proper, non-trivial subgroups. Otherwise G is **indecomposible**.

For example, $\mathbb{Z}_{2^3} \times \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_5 \approx \mathbb{Z}_{24} \times \mathbb{Z}_{15}$ is decomposible.

Remark. $G = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Z}_3 \times \mathbb{Z}_5 \approx \mathbb{Z}_2 \times \mathbb{Z}_6 \times \mathbb{Z}_{30}$ is also decomposible. Since $G \approx (\mathbb{Z}_2 \times \mathbb{Z}_6) \times \mathbb{Z}_{30}$.

Theorem 1.3.2. *The finite indecomposible abelian groups ar exactly the cyclic groups with order a power of a prime.*

Proof. Necessary part: Let G be a finite, indecomposible abelian group. By fundamental theorem of finitely generated abelian groups, G is isomorphic to a direct product of cyclic groups of prime power order.

$$G \approx \mathbb{Z}_{{p_1}^{r_1}} \times \mathbb{Z}_{{p_2}^{r_2}} \times \cdots \times \mathbb{Z}_{{p_n}^{r_n}}$$

Thus for G to be indecomposible the direct product should be a cyclic group of prime power order. $G \approx \mathbb{Z}_{{p_1}^{r_1}}$.

Sufficient part : Let p be a prime and r a non-negative integer. Cyclic group of order p^r is isomorphic to $\mathbb{Z}_{p^r}$. Since every cyclic groups are abelian, $\mathbb{Z}_{p^r}$ is an abelian group of finite order p^r. It is enough to prove that $\mathbb{Z}_{p^r}$ is indecomposible.

A proper, non-trivial subgroup of $\mathbb{Z}_{p^r}$ is of the form $\mathbb{Z}_{p^i}$ where $0 < i < r$. Suppose $\mathbb{Z}_{p^r}$ is decomposible. Then $\exists i, j \in \mathbb{Z}^+$ such that $\mathbb{Z}_{p^r} \approx \mathbb{Z}_{p^i} \times \mathbb{Z}_{p^j}$ and $i + j = r$. Clearly, p^i and p^j are not relatively prime, thus $\mathbb{Z}_{p^r} \not\approx \mathbb{Z}_{p^i} \times \mathbb{Z}_{p^j}$. Therefore, cyclic groups of order prime power are indecomposible. □

Theorem 1.3.3. *If m divides the order of a finite abelian group G, then G has a subgroup of order m.*

Proof. Let $G \approx \mathbb{Z}_{{p_1}^{r_1}} \times \mathbb{Z}_{{p_2}^{r_2}} \times \cdots \times \mathbb{Z}_{{p_n}^{r_n}}$. Then $|G| = p_1^{r_1} p_2^{r_2} \cdots p_n^{r_n}$. Suppose m divides $|G|$, then $m = p_1^{s_1} p_2^{s_2} \cdots p_n^{s_n}$ where $0 \leq s_i \leq r_i$. Define $H = \mathbb{Z}_{{p_1}^{s_1}} \times \mathbb{Z}_{{p_2}^{s_2}} \times \cdots \times \mathbb{Z}_{{p_n}^{s_n}}$. Then H is subgroup of order m. □

Theorem 1.3.4. *If m is square-free integer, then every abelian group of order m is cyclic.*

Proof. Let m be a square-free integer and G be an abelian group of order m. By fundamental theorem of finitely generated abelian groups

$$G \approx \mathbb{Z}_{{p_1}^{r_1}} \times \mathbb{Z}_{{p_2}^{r_2}} \times \cdots \times \mathbb{Z}_{{p_n}^{r_n}}$$

We have, m is square-free. Thus $r_i = 1$ and p_i are distinct. Therefore, $G \approx \mathbb{Z}_{p_1 p_2 \cdots p_n}$ is a cyclic group of order m. □

1.4 Exercises §11

Question 5. Enumerate subgroups of $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_4$

1.4.1 Abelian Groups

Remark. Direct product of abelian groups is abelian.

Question 6. Enumerate abelian groups of order 72 ?

Remark. Let G be an abelian group. The subset H of G with identity element and all elements of order n is subgroup of G if and only if n is a prime.

Proof. Suppose $a \in H$. Then every power of a has order n. Suppose n is not prime. Then d divides n and a^d has order n/d. □

1.4.2 Torsion Group and Torsion Coefficients

Remark. If group G is abelian, then its elements of finite order forms a subgroup.(hint : $a, b \in G$ has finite order, then ab has finite order. And $a \in G$ has finite order, then a^{-1} has finite order)

Definitions 1.4.1. The **torsion group** of an abelian group G is the subgroup of G containing only those elements of finite order.An abelian group is **torsion free** if identity element is the only element of finite order. G is Torsion free if Torsion group of G is trivial, $\{e\}$.

Definitions 1.4.2. The integers $m_1, m_2, \cdots, m_n$ are torsion coefficients of G such that $G \approx \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2} \times \cdots \times \mathbb{Z}_{m_n}$ where m_i divides m_{i+1}.

For example, $\mathbb{Z}_6 \times \mathbb{Z}_{12} \times \mathbb{Z}_{20}$ has torsion coefficients $2, 12, 60$

Remark (Algorithm to find torsion coefficients of a group). Suppose G has a direct product form.

Step 1 Find power of each prime in the direct product form

Step 2 List power of each prime

Step 3 Append 1s on left to make all lists to equal length

Step 4 Product of ith number on each list gives m_i

For example, $G \approx \mathbb{Z}_6 \times \mathbb{Z}_{12} \times \mathbb{Z}_{20}$

Step 1 $\mathbb{Z}_6 \times \mathbb{Z}_{12} \times \mathbb{Z}_{20} \approx \mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Z}_{2^2} \times \mathbb{Z}_3 \times \mathbb{Z}_{2^2} \times \mathbb{Z}_5$

Step 2 (2,4,4), (3,3), (5)

Step 3 (2,4,4), (1,3,3), (1,1,5)

Step 4 (2,12,60)

1.4.3 Torsion p-subgroup

"Caution : Torsion p-subgroup is a name suggested by 'Jacob'. And is not among the standard terminology in group theory."

Remark. Let G be a group. Let p be an integer, $(p > 1)$. Then the set of all element of G of order p together with the identity element is a group, if p is a prime.

For example : Consider symmetric group, S_3. It has three elements of order 2, namely $\mu_1 = (2,3)$, $\mu_2 = (1,3)$ and $\mu_3 = (1,2)$. Clearly, the set of all element of order 2 together with identity ( ) is not a subgroup of S_3 as $(1,2)(1,3) = (3,1,2)$ is an element of order 3. Thus by counter-example, for a prime p, the set of all elements of a non-abelian group G together with identity is not necessarily a subgroup of G.

Let G be an abelian group. Let p be a prime. Let H be the set of all element of G of order p together with the identity e. Let $g, h \in H$. Clearly, $g^p = e$ and $h^p = e$. For every $g, h \in G$, $gh \in G$, since G is abelian $(gh)^p = g^p h^p = e$. Also we have, $g^{-1} = g^{p-1}$ is a element of order p. Therefore, H is a subgroup of G.

Let g be an element of order 4. Then g^2 is a element of order 2. Thus, the subgroup generated by g or the smallest subgroup containing g has a element of order 2. Therefore, elements of order 4 together with identity cannot be a subgroup of G. Thus, 'Torsion p-subgroup' exists only if p is a square-free integer.

Let g be an element of order 6. Then g^2 is an element of order 3. Again, elements of order 6 together with identity cannot be a subgroup of G. Thus, 'Torsion p-subgroup' exists only if p is a power of a prime. Thus, 'Torsion p-subgroup' of G exists only if G is abelian and p is a prime.

1.4.4 Normal Factors of G

This is a warm-up exercise for §37.5, where the theory is discussed by Fraleigh. However, we are able to conclude the following :

Remark. Let $G = H \times K$. Let $g \in G$. Then $g = (h,k) \in H \times K$.Clearly, H is a subset of G, and $H \times \{e\} \leq G$. This subgroup is isomorphic to H. Thus $h \in H$ suggests the existence of $(h, e) \in G$.

Similarly $k \in K$ suggests $(e,k) \in G$. Therefore, $hk \in G$ suggests $(h,e)(e,k) = (h,k) = (e,k)(h,e)$. We know that, $kh \in G$ suggest $(e,k)(h,e) \in G$. Thus, $hk = kh$ for every $h \in H$ and every $k \in K$.

In other words, if $G = H \times K$, then H, K are isomorphic to normal subgroups H', K' of G such that $H' \cap K' = \{e\}$ and $G \simeq K \times H$.

1.5 Cosets and Homomorphism

Definitions 1.5.1. A **permutation group** S_n is the set of all permutations on the set $\{1, 2, \cdots, n\}$.

Remark. Consider, $(1,2,3)(4,5), (1,2)(3,4) \in S_5$. I was wrong about the order in which the permuations are carried out. It follows the same order as function composition. That is, $f \circ g(x)$ implies $f(g(x))$. Similarly, $\sigma\rho$ implies

$\sigma(\rho(1, 2, \ldots, n))$.

$$\begin{pmatrix}1 & 2\end{pmatrix}\begin{pmatrix}3 & 4\end{pmatrix} * \begin{pmatrix}1 & 2 & 3\end{pmatrix}\begin{pmatrix}4 & 5\end{pmatrix} = \begin{pmatrix}1 & 2 & 3 & 4 & 5\\ 2 & 3 & 1 & 5 & 4\\ 1 & 4 & 2 & 5 & 3\end{pmatrix} = \begin{pmatrix}2 & 4 & 5 & 3\end{pmatrix}$$

$$\begin{pmatrix}1 & 2 & 3\end{pmatrix}\begin{pmatrix}4 & 5\end{pmatrix} * \begin{pmatrix}1 & 2\end{pmatrix}\begin{pmatrix}3 & 4\end{pmatrix} = \begin{pmatrix}1 & 2 & 3 & 4 & 5\\ 2 & 1 & 4 & 3 & 5\\ 3 & 2 & 5 & 1 & 4\end{pmatrix} = \begin{pmatrix}1 & 3 & 5 & 4\end{pmatrix}$$

Clearly, S_5 is a non-abelian group of order 120.

Definitions 1.5.2. Kernel of a function $\phi : G \to G'$ is the inverse image of the identity element in G'.

Definitions 1.5.3. Cosets and Normal Subgroup,

- A **left coset** gH is the subset $\{gh \in G : h \in H\}$ where $g \in G$ and $H \leq G$.
- A **right coset** Hg is the subset $\{hg \in G : h \in H\}$ where $g \in G$ and $H \leq G$.
- A subgroup H of group G is **normal** if $gH = Hg,\ \forall g \in G$.
- All subgroups of abelian groups are normal.

Remark. For example, $H = \{1, \rho^2, \mu, \mu\rho^2\}$ is a normal subgroup of D_4. And $K = \{1, \mu\}$ is a subgroup of D_4 which is not normal. Note that, $\rho\mu \neq \mu\rho$. Clearly, $\rho K \neq K\rho$. However, $\rho H = \{\rho, \rho^3, \mu\rho^3, \mu\rho\} = H\rho$.

Remark (Lagrange's Theorem). Let G be a finite group. If $H \leq G$, then order of H divides order of G.

- $aH \cap bH \neq \phi \implies aH = bH$.
- $\forall g \in G,\ g \in gH$.
- $\forall g \in G,\ |gH| = |H|$.

Remark (Cayley's Theorem). Every group is isomorphic to a group of isomorphisms.

Definitions 1.5.4. Let G, G' be groups. A **group homomorphism** is a function $\phi : G \to G'$ such that $\phi(x)\phi(y) = \phi(xy)$. Clearly $\phi(e) = e'$. A **trivial homomorphism** is a function $\phi : G \to G'$ such that $\phi(G) = \{e'\}$.

Remark. Group homomorphism $\phi : G \to G'$ preserves identity, inverses and subgroups. And kernel of group homomorphism is a normal subgroup of G.

Remark. For example, $\phi : D_4 \to \mathbb{Z}_2$ defined by $\phi(\rho) = 1$ and $\phi(\mu) = 0$ is a group homomorphism with $\ker(\phi) = \{1, \rho^2, \mu, \mu\rho^2\}$.

1.6 Factor Groups

Remark (factor group). Let H be a normal subgroup of a group G. Then the **factor group** of G over H, G/H is the group of cosets of H in G.

Remark. For example, $H = \{1, \rho^2, \mu, \mu\rho^2\}$ is normal subgroup of D_4. And the factor group $D_4/H = \{1H, \rho H\}$.

Theorem 1.6.1. *Let $\phi : G \to G'$ be a group homomorphism with kernel H. Then cosets of H form a factor group, G/H where $(aH)(bH) = (abH)$ Also, $\mu : G/H \to \phi[G]$ defined by $\mu(aH) = \phi(a)$ is an isomorphism.*

Proof. Let $\phi : G \to G'$ be a group homomorphism with $\ker(\phi) = H$. We have, $\phi^{-1}(\phi(a)) = \{g \in G : \phi(g) = \phi(a)\}$.

Let $x \in aH$. Then $x = ah$ for some $h \in H$. And $\phi(x) = \phi(ah) = \phi(a)\phi(h) = \phi(a)$, since $\phi(h) = e'$. Thus, $x \in \phi^{-1}(\phi(a))$ and $aH \subset \phi^{-1}(\phi(a))$.

Let $x \in \phi^{-1}(\phi(a))$. Then $\phi(x) = \phi(a)$. And $\phi(a)^{-1}\phi(x) = e' \implies \phi(a^{-1}x) = e'$. Clearly, $a^{-1}x \in \ker(\phi)$. Thus, there exists $h \in H$ such that $a^{-1}x = h$. Therefore, $x = ah$ for some $h \in H$,. Thus, $x \in aH$ and $\phi^{-1}(\phi(a)) \subset aH$. Therefore, $\phi^{-1}(\phi(a)) = aH$.

Similarly, $\phi^{-1}(\phi(a)) = Ha$. Thus $aH = Ha$ and H is a normal subgroup of G. Therefore, we have the factor group G/H.

To prove : $\mu : G/H \to \phi[G]$ is a one-one correspondence. ie, $aH \overset{\mu}{\longleftrightarrow} \phi(a)$. To prove : μ is injective. Suppose $\mu(aH) = \mu(bH)$. Then $\phi(a) = \phi(b)$. And $b \in \phi^{-1}(\phi(a)) = aH$. Therefore, $bH = aH$.

To prove : μ is surjective. Let $\phi(a) \in \phi[G]$. Then, there exists aH such that $\mu(aH) = \phi(a)$.

We have, $\mu(aH) = \phi(a)$, $\mu(bH) = \phi(b)$, and $\mu((ab)H) = \phi(ab)$.
Therefore, $\mu((aH)(bH)) = \mu((ab)H) = \phi(ab) = \phi(a)\phi(b) = \mu(aH)\mu(bH)$. Thus μ is a homomorphism. Therefore μ is an isomorphism. □

Theorem 1.6.2. *Let G be a group and $H \leq G$. Then left coset multiplication is well-defined by $(aH)(bH) = (ab)H$ if and only if H is a normal subgroup of G.*

Proof. Necessary part : Suppose $(aH)(bH) = (ab)H$ is well-defined. Let $a \in G$. It is enough to prove that $aH = Ha$. Let $x \in aH$. Then $(xH)(a^{-1}H) = (xa^{-1})H$. Also $(aH)(a^{-1}H) = eH = H$. We have, coset multiplication is well-defined. Thus $xa^{-1} = h \in H \implies x = ha \in Ha$. Then, $aH \subset Ha$. Similarly, $Ha \subset aH$ and $aH = Ha$. Therefore, H is a normal subgroup of G.

Sufficient part: Suppose H is a normal subgroup of G, and let $x \in aH$ and $y \in bH$. $x \in aH \implies x = ah_1$ for some $h_1 \in H$ $y \in bH \implies y = bh_2$ for some $h_2 \in H$. Therefore $xy = (ah_1)(bh_2) = (a(h_1(bh_2)) = a((h_1 b)h_2) = a((bh_3)h_2) = a(b(h_3h_2)) = a(bh_4)$. Since H is a group, $h_3h_2 = h_4 \in H$ Thus, $xy = a(bh_4) = (ab)h_4 \in (ab)H$ for all $x \in aH$ and $y \in bH$ Thus $(aH)(bH) = (ab)H$. □

Corollary 1.6.2.1. *Let H be a normal subgroup of G. Then the cosets of H form a group G/H under the binary operation $(aH)(bH) = (ab)H$.*

Proof. Let H be a normal subgroup and aH, bH, cH are cosets of H in G.

G1 Associativity
$(aH)[(bH)(cH)] = (aH)[(bc)H] = [a(bc)]H = [(ab)c]H = [(ab)H](cH) = [(aH)(bH)](cH)$

G2 Existence of identity, eH
$(aH)(eH) = (ae)H = aH$ and $(eH)(aH) = (ea)H = aH$.

G3 Existence of inverse $(a^{-1}H)$
$(aH)(a^{-1}H) = (aa^{-1})H = eH$ and $(a^{-1}H)(aH) = (a^{-1}a)H = eH$.

□

Remark. $n\mathbb{Z}$ is a normal subgroup of $\mathbb{Z}$. And $\mathbb{Z}/n\mathbb{Z} \approx \mathbb{Z}_n$. $\mathbb{Z}_n$ is a torsion group isomorphic to a factor group of torsion free group $\mathbb{Z}$.

Remark. Let $c \in \mathbb{R}^*$. Then the cyclic group generated by c is a normal subgroup of $\mathbb{R}$ and $\mathbb{R}/\langle c \rangle \approx \mathbb{R}_c$.

Question 7. Let $c = 0.31$. Find the coset $x \ + \langle 0.31 \rangle$ containing 2.

1.7 Fundamental Homomorphism & Automorphisms

1.7.1 Fundamental Homomorphism Theorem

Theorem 1.7.1. *Let H be a normal subgroup of G. Then $\gamma : G \to G/H$ is defined by $\gamma(x) = xH$ is a homomorphism with kernel H.*

Proof. $\gamma(x)\gamma(y) = (xH)(yH)$ Let $h_1, h_2 \in H$, $(xh_1)(yh_2) = xyh_3h_2 = xyh_4$ for some $h_3, h_4 \in H$. Therefore $(xH)(yH) = (xy)H$. $\gamma(xy) = (xy)H = \gamma(x)\gamma(y)$. $\gamma(x) = xH = H \iff x \in H$. Therefore, $\ker(\gamma) = H$. □

Remark. Suppose H is a normal subgroup of a group G. Then, a homomorphism $\gamma : G \to G/H$ is a natural homomorphism.

Theorem 1.7.2 (Fundamental Homomorphism)**.** *Let $\phi : G \to G'$ be a group homomorphism with kernel H. Then $\phi[G]$ is a group, and $\mu : G/H \to \phi[G]$ given by $\mu(gH) = \phi(g)$ is an isomorphism. If $\gamma : G \to G/H$ is the homomorphism given by $\gamma(g) = gH$, then $\phi(g) = \mu\gamma(g)$ for each $g \in G$.*

Proof. μ is an isomorphism $G/H \overset{\mu}{\leftrightarrow} \phi[G]$. $\mu(\gamma(g)) = \mu(gH) = \phi(g)$. Thus $\mu\gamma = \phi$. □

"Every group homomorphism $\phi : G \to G'$ with kernel N has a unique natural group homomorphism $\gamma : G \to G/N$ and a unique isomorphism $\mu : G/N \to G'$ such that $\phi = \mu\gamma$. That is, $\phi(g) = \mu(\gamma(g)) = \mu(gN)$."

1.7.2 Inner Automorphism

Theorem 1.7.3. *Let H be a subgroup of G, then the following statements are equivalent:*

1. $ghg^{-1} \in H, \ \forall g \in G, \ h \in H$
2. $gHg^{-1} = H, \ \forall g \in G$
3. $gH = Hg, \ \forall g \in G$

Proof. Let G be a group and $H \leq G$. By right multiplication, $gHg^{-1} = H \iff gH = Hg$ Trivially, $\forall h \in H,\ ghg^{-1} \in H \iff gHg^{-1} \subset H$ Therefore, it is enough to prove that $H \subset gHg^{-1}$ Let $h \in H$ and $x \in G$, then $xhx^{-1} = h'$ for some $h' \in H$. Then $h = x^{-1}h'x = x^{-1}h'(x^{-1})^{-1} \in gHg^{-1}$ where $x^{-1} = g \in G$. Thus $h \in gHg^{-1}$ and $H \subset gHg^{-1}$. Therefore $gHg^{-1} = H$. □

Definitions 1.7.1. Automorphisms :

- An **automorphism** of G is an isomorphism $\phi : G \to G$
- The **inner automorphism** of G by $g \in G$ is the isomorphism $i_g : G \to G$ defined by $i_g(x) = gxg^{-1}$ for all $x \in G$
- The **conjugate** of x by g is the element $gxg^{-1} \in G$
- The **conjugate subgroup** of subgroup H, $i_g[H] = \{ghg^{-1} : h \in H\}$

Remark. For example, $i_\rho : D_4 \to D_4$ defined by $i_\rho(x) = \rho x \rho^{-1}$ is an inner automorphism. We have, $H = \{1, \mu\}$ is a subgroup of D_4. The conjugate subgroup $i_\rho[H] = \{1, \mu\rho^2\}$.

Remark. Normal subgroups are invariant under any inner automorphism.

1.8 Exercise §14

1.8.1 Normal subgroups

Question 8. Prove that the notion of normality is stronger than abelian.

Remark. Let G be a group. Then intersection of normal subgroups of G is a normal subgroup of G.

Proof. Let $\mathcal{N}$ be a nonempty subfamily of normal subgroups of G. Let N be the intersection of subgroups in $\mathcal{N}$. Clearly, intersection of subgroups of G is also subgroup of G.

Suppose $x \in N$. Then $x \in H$ for every $H \in \mathcal{N}$. Suppose $x \in N$. Then $gxg^{-1} \in H$, for every normal subgroup $H \in \mathcal{N}$. Thus $gxg^{-1} \in N$. Therefore, N is a normal subgroup of G. □

Challenge 1. Let $\phi : G \to G'$ and $\psi : G \to G'$ be two group homomorphism with kernel H and K respectively. Show that $\phi\psi : G \to G'$ defined by $\phi\psi(g) = \phi(g)\psi(g)$ is also a group homomorphism, but $\ker(\phi\psi) \neq H \cap K$.
(hint : Construct ψ such that $\phi(g)\psi(g) = e$ for some $g \notin H$)

Remark. Let $S \subset G$, then G has a smallest, normal subgroup containing S.

Proof. Let N be the intersection of all normal subgroups of G containing S. Then $S \subset N$ and N is a normal subgroup of G. Let H be the smallest, normal subgroup of G containing S. Then $N \subset H$. Therefore, $H = N$. □

Remark. If a finite group G has exactly one subgroup H of order m, then H is normal.

Proof. Let G be a finite group. Suppose G has only one subgroup of order m, say H. We know that, conjugate of a subgroup is a subgroup of same order. Thus, conjugates of H are H itself. Thus, $xHx^{-1} = H$ for every $x \in G$. Therefore, $xH = Hx$ for every $x \in G$ and H is a normal subgroup of G. $\square$

Remark. If G has a subgroup of order s, then the intersection of all subgroups of order s is a normal subgroup of G.

Proof. Let G be a group. Let H be a subgroup of G of order s. Let N be the intersection of all subgroups of G of order s. ?? $\square$

1.8.2 Linear Groups

Definitions 1.8.1. The set of all $n \times n$, non-singular matrices with real entries is a group under matrix multiplication. This group is the **General Linear Group** and is denoted by $GL(n, \mathbb{R})$.

Definitions 1.8.2. The set of all $n \times n$ matrices with real entries and determinant ± 1 is a group under matrix multiplication. This group is the **Special Linear Group** and is denoted by $SL(n, \mathbb{R})$.

Remark. $SL(n, \mathbb{R})$ is a normal subgroup of $GL(n, \mathbb{R})$

Proof. Let $A \in GL(n, \mathbb{R})$ and $B \in SL(n, \mathbb{R})$. Let $r = |A|$ and we have $|B| = \pm 1$. Then $|AB| = |A|\ |B| = \pm r$.

Let M_r be the set of all matrices with determinant $\pm r$ where $r \in \mathbb{R}$. Clearly, for every matrix A with determinant r, $A \in M_r$. And there exists a matrix $C = B^{-1}AB$. Then, $|C| = |B^{-1}|\ |A|\ |B| = \pm r$. Thus, $C \in M_r$ and $AB = BC$. Clearly, M_r is a left coset of $GL(n, \mathbb{R})$. Therefore, $SL(n, \mathbb{R})$ is a normal subgroup of $GL(n, \mathbb{R})$. $\square$

1.8.3 Factor Group

Remark. A_n is a normal subgroup of S_n. And $S_n/A_n \simeq \mathbb{Z}_2$.

Proof. Let function $\phi : S_n \to \mathbb{Z}_2$ be defined by $\phi(\sigma) = 0$ if σ is an even permutation and $\phi(\sigma) = 1$ if σ is an odd permutation. Then ϕ is a homomorphism with kernel A_n, the set of all even permutations in S_n. We know that, the kernel of a homomorphism is a normal subgroup of the domain. Therefore, A_n is a normal subgroup of S_n. $\square$

Remark. If H is normal subgroup of G and $(G : H) = m$, then $a^m \in H$ for all $a \in G$.

Proof. Let H be a normal subgroup of G such that $(G : H) = m$. Then $|G/H| = m$. Let $a \in G$, then $aH \in G/H$. Then by Lagrange's theorem, order of aH divides m. Therefore, $(aH)^m = eH$. We have, $(aH)^m = a^m H$. Therefore, $a^m \in H$. $\square$

Remark. Every factor group of an abelian group is also abelian.

Proof. Let G be an abelian group and H be a normal subgroup of G. Then G/H is a factor group of G. Let $aH, bH \in G/H$. Then, $(aH)(bH) = (ab)H = (ba)H = (bH)(aH)$. Clearly, factor group of an abelian group is abelian. $\square$

Remark. Let G be a group and T be the torsion subgroup of G, then the factor group, G/T is torsion free.

Proof. Let G be a group T its torsion subgroup. Let $g \in G$. If the order of g is finite, then $g \in T$. Suppose G/T has a element of finite order $m > 1$, say xT where $x \notin T$. Then x is an element of infinite order.

We have, order of xT is m. That is, $(xT)^m = eT$. Thus, $(xT)^m = (x^m)T = eT$. Therefore, $x^m = y \in T$. Since $y \in T$, y is an element of finite order, say r. Then $(x^m)^r = x^{mr} = e$. This is a contradiction as x is an element of infinite order. Therefore, every element of G/T except eT are of infintie order. Clearly, G/T is torsion free. □

1.8.4 Commutator subgroup

Definitions 1.8.3. A **commutator** c in group G is an element in the form $c = aba^{-1}b^{-1}$ for some $a, b \in G$. The **commutator subgroup** is the smallest normal subgroup containing all commutators in G.

Remark. Let C be the commutator subgroup of G, then G/C is abelian.

Proof. Let C be the commutator subgroup of G. Let $x, y \in G$. Then $x^{-1}yxy^{-1} \in C$. Therefore, $x(x^{-1}yxy^{-1}) \in xC$ and $y(x^{-1}yxy^{-1}) \in yC$. But, $x(x^{-1}yxy^{-1})y(x^{-1}yxy^{-1}) = yx(x^{-1}yxy^{-1}) \in (yx)C$. Therefore, $(xC)(yC) = (yx)C = (yC)(xC)$. Clearly, G/C is abelian. □

Remark. The factor group G/C is the abelianised version of G.

Remark. Let G be a group and C be the commutator group of G. Let H be a normal subgroup of G. If G/H is abelian, then C is a subgroup of H.

Question 9. Find commutator subgroup of the dihedral group D_4 ?

1.8.5 Automorphism

Remark. Every inner automorphism is an identity map for an abelian group.

Proof. Let G be an abelian group and $g \in G$. Then $i_g : G \to G$ is defined by $i_g(x) = gxg^{-1} = gg^{-1}x = x$. Clearly, i_g is an identity map. □

Remark. Set of all $g \in G$ such that the inner automorpism i_g is an identity map is normal.

Proof. Let H be the set of all $g \in G$ such that i_g is an identity map.

$$\begin{aligned} H &= \{g \in G : i_g(x) = x,\ \forall x \in G\} \\ &= \{g \in G : gxg^{-1} = x,\ \forall x \in G\} \\ &= \{g \in G : gx = xg,\ \forall x \in G\} \end{aligned}$$

Therefore, H is a normal subgroup of G. □

Remark. Set of automorphisms Γ of a group G is a group under composition. And the set of inner automorphisms is a normal subgroup of Γ.

Proof. Let G be a group. We know that, the composition of two automorphisms is also an automorphism. Also, the composition of functions is associative. Let i be the identity map and μ be any automorphism. Then $i\mu = \mu = \mu i$. Since automorphisms are bijective, there exists a unique inverse μ^{-1} for each automorphism μ such that $\mu\mu^{-1} = i$. Clearly, the set of all automorphisms of a group G is also a group. □

Remark. Subgroup conjugacy is an equivalance relation on the set of subgroups.

Proof. Let H be a subgroup of a group G. Conjugation is reflexive, since $eHe^{-1} = H$ and $H \sim H$.

Let K be a conjugate subgroup of H. That is, $H \sim K$. Then $K = gHg^{-1}$. Clearly, $H = g^{-1}K(g^{-1})^{-1}$. Thus, H is a conjugate of K. ie, $K \sim H$.

Let $H \sim K$ and $K \sim L$. Then we have $x, y \in G$ such that $K = xHx^{-1}$ and $L = yKy^{-1}$. Now, $L = y(xHx^{-1})y^{-1} = (yx)H(yx)^{-1}$. Thus, $H \sim L$. □

Question 10. Find the automorphism group $\Gamma(\mathbb{Z}_2 \times \mathbb{Z}_4)$?

1.9 Simple Groups

Remark. Factor Group Computations :

- The converse of Lagrange's theorem is false.
  For example, A_4 has order 12, but doesn't have a subgroup of order 6.
- Factor group of a cyclic group is cyclic.
  (hint : if g is a generator of G, then gH is a generator of G/H.)

Question 11. Show that $\mathbb{Z}_4 \times \mathbb{Z}_6 / \langle (2, 3) \rangle \approx \mathbb{Z}_4 \times \mathbb{Z}_3$.

Definitions 1.9.1. A group is simple if it is non-trivial and has no proper, non-trivial normal subgroups.

Remark. Abelian, simple groups are $\mathbb{Z}_p$, the cyclic groups of prime order.

Remark. Symmetric group, S_3 is not simple.
The subgroup, $\{\rho_0, \rho_1, \rho_2\}$ is a normal subgroup of S_3.

Remark. Smallest nonabelian, simple group is A_5.
Every nonabelian, simple group of order 60 is isomorphic to A_5.

Remark. Simple groups :

- Alternating groups A_n are simple for $n \geq 5$.
- Every finite group can be factorised into simple groups.
- Every finite, non-abelian, simple group is of even order.
- Group homomorphism preserves normal subgroups.
- M is maximal normal subgroup of G if and only if G/M is simple.
- Center $Z(G) = \{z \in G : zg = gz, \ \forall g \in G\}$ is normal.
- Center of non-abelian groups of order pq are trivial if p, q are primes.
- Factor group, G/N is abelian if and only if C is a subgroup of N.

Question 12. G is simple and H is subgroup of G, then H is simple ?

1.10 Group Action on a Set

Definitions 1.10.1. An action of a group G on a set X is a map. $* : G\times X \to X$ such that

1. $ex = x,\ \forall x \in X$
2. $(g_1g_2)(x) = g_1(g_2x),\ \forall x \in X,\ \forall g_1, g_2 \in G$

Then X is a G-set.

Theorem 1.10.1. *Let X be a G-set. $\forall g \in G$, $\sigma_g : X \to X$ defined by $\sigma_g(x) = gx$ is a permutation of X. Also, the map $\phi : G \to S_X$ defined by $\phi(g) = \sigma_g$ is a homomorphism with the property that $\phi(g)(x) = gx$.*

Proof. Suppose X is a G-set. Let $g \in G$, and $x_1, x_2 \in X$.

Suppose $\sigma_g(x_1) = \sigma_g(x_2)$. $\implies gx_1 = gx_2 \implies g^{-1}(gx_1) = g^{-1}(gx_2) \implies (g^{-1}g)x_1 = (g^{-1}g)x_2$. $\implies ex_1 = ex_2 \implies x_1 = x_2$. Thus, σ_g is injective.

Let $x \in X$. Then $\sigma_g(g^{-1}x) = g(g^{-1}x) = (gg^{-1})x = ex = x$. Thus, σ_g is surjective. Therefore, σ_g is a permutation of X, $\sigma_g \in S_X$.

Let $g_1, g_2 \in G$. And $\phi(g_1)(x) = \sigma_{g_1}(x) = g_1x$, $\phi(g_2)(x) = \sigma_{g_2}(x) = g_2x$. $\phi(g_1g_2)(x) = \sigma_{g_1g_2}(x) = (g_1g_2)x = g_1(g_2x) = \sigma_{g_1}(g_2x) = \phi(g_1)(g_2x) = \phi(g_1)(\sigma_{g_2}(x)) = \phi(g_1)(\phi(g_2)(x)) = \phi(g_1)\phi(g_2)(x)$. Therefore, $\phi(g_1g_2) = \phi(g_1)\phi(g_2)$ and ϕ is a homomorphism. $\square$

Definitions 1.10.2. Group Action,

- G acts **faithfully** on X, if e is the only element that leaves every $x \in X$ fixed.
- G is **transitive** on X if for every $x_1, x_2 \in X$, $\exists g \in G$ such that $gx_1 = x_2$. G is transitive on X iff the subgroup $\phi[G]$ of S_X is transitive.

1.11 Isotropy subgroups & Orbits

Definitions 1.11.1. Let X be a G-set.

- The subset fixed by g, $X_g = \{x \in X : gx = x\}$
- The isotropy subgroup of x, $G_x = \{g \in G : gx = x\}$
  Let $Y \subset X$, then $G_Y = \{g \in G : gy = y, \forall y \in Y\}$ is a subgroup of G.
- The orbit of x in X under G, $Gx = \{gx \in X : g \in G\}$

Theorem 1.11.1. *Let X be a G-set. Then G_x is a subgroup of G, $\forall x \in X$.*

Proof. Let $x \in X$. And $g_1, g_2 \in G_x$. Then $g_1x = x$ and $g_2x = x$.

Clearly, $(g_1g_2)x = g_1(g_2x) = g_1x = x$. Therefore, $g_1g_x \in G_x$. Also $ex = x \implies e \in G_x$. Let $g \in G_x$. Then $gx = x \implies g^{-1}(gx) = g^{-1}x \implies (g^{-1}g)x = g^{-1}x \implies x = g^{-1}x$. Thus, for any $g \in G_x$, $g^{-1} \in G_x$. Therefore, G_x is a subgroup of G for any $x \in X$. $\square$

Theorem 1.11.2. *Let X be a G-set and $x_1, x_2 \in X$. Then the relation $\sim$ defined by $x_1 \sim x_2$ iff $gx_1 = x_2$ is an equivalence relation.*

Proof. Let $x \in X$. Then $ex = x \implies x \sim x$. Let $x_1, x_2 \in X$ and $x_1 \sim x_2$. Then there exists some $g \in G$ such that $gx_1 = x_2$. We have, $g^{-1}x_2 = g^{-1}(gx_1) = (g^{-1}g)x_1 = ex_1 = x_1$. Therefore, $x_2 \sim x_1$. Let $x_1, x_2, x_3 \in X$ and $x_1 \sim x_2$ and $x_2 \sim x_3$. Then there are $g_1, g_2 \in G$ such that $g_1x_1 = x_2$ and $g_2x_2 = x_3$. Clearly, $g_2g_1 \in G$ and $(g_2g_1)x_1 = g_2(g_1x_1) = g_2x_2 = x_3$. Therefore, $x_1 \sim x_3$. □

Theorem 1.11.3. *Let X be a G-set and $x \in X$. Then $|Gx| = (G : G_x)$. If $|G|$ is finite, then $|Gx|$ is a divisor of $|G|$.*

Proof. We have Gx is the orbit of x in X under G and L_{G_x} is the left cosets of G_x in G. Let $x_1 \in Gx$. Then there exists $g_1 \in G$ such that $x_1 = g_1x$. Define $\psi : Gx \to L_{G_x}$ by $\psi(x_1) = g_1G_x$.
Step 1 : ψ is well-defined.

Let $x_1 \in Gx$. Suppose there exists $g_1, g_1' \in G$ such that $g_1x = x_1$ and $g_1'x = x_1$. Then we have, $g_1x = g_1'x \implies x = g_1^{-1}(g_1'x) = (g_1^{-1}g_1')x$. Thus, $g_1^{-1}g_1' \in G_x$. Therefore, $g_1(g_1^{-1}g_1') \in g_1G_x$. Clearly, $g_1(g_1^{-1}g_1') = (g_1g_1^{-1})g_1' = g_1' \in g_1G_x$. Therefore, $g_1G_x = g_1'G_x$. And $\psi(x_1) = g_1G_x$ is well-defined.
Step 2 : ψ is one-to-one.

Suppose $\psi(x_1) = \psi(x_2)$. Let $x_1, x_2 \in Gx$ such that $x_1 = g_1x$ and $x_2 = g_2x$. Then we have $\psi(x_1) = \psi(x_2) \implies g_1G_x = g_2G_x$. Thus, $g_2 = g_1g$ for some $g \in G_x$. Clearly, $x_2 = g_2x = (g_1g)x = g_1(gx) = g_1x = x_1$.
Step 3 : ψ is onto.

Let g_1G_x be a left coset of G_x in G. Then we have, $g_1 \in G$ and $g_1x \in Gx$, say x_1. Therefore, there exists $x_1 \in Gx$ such that $\psi(x_1) = g_1G_x$. □

1.12 Exercise §16

Definitions 1.12.1. Let X be a G-set. And $S \subset X$. Then S is a **sub-G-set** if the orbit Gs of each $s \in S$ is contained in S.

$$S = \bigcup_{s \in S} Gs \tag{1.4}$$

Remark.
- Every G-set is a union of its orbits.
- Every G-set is isomorphic to the disjoint union of left coset G-sets.

Definitions 1.12.2. Let G be a group. Let X, Y be two G-sets. Then function $\phi : X \to Y$ is a G-set isomorphism if

1. ϕ is a bijection and
2. ϕ is a G-set homomorphism
   ie, $g\phi(x) = \phi(gx)$, for every $x \in X$ and every $g \in G$.

1.13 Application of G-Sets to Counting

Theorem 1.13.1 (Burnside)**.** *Let G be a finite group and X a finite G-set. If r is the number of orbits in X under G, then*

$$r|G| = \sum_{g \in G} |X_g| \tag{1.5}$$

Proof. Let $N = \{(g, x) \in G \times X : gx = x\}$.
Step 1 : $|N| = \sum_{g \in G} |X_g|$.

Let $g \in G$. We have, X_g is the set of all $(g, x) \in N$ with g as first member. Enumerating elements of N for each $g \in G$, we get $\sum_{g \in G} |X_g| = N$.

Step 2 : $|N| = r|G|$.

Let $x \in X$. We have, G_x is the set of all $g \in G$ such that $gx = x$. In other words, G_x is the set of all $g \in G$ such that $(g, x) \in N$ with x as second member. However, $|Gx| = (G : G_x)$.

$$\implies |Gx| = \frac{|G|}{|G_x|} \implies |G_x| = \frac{|G|}{|Gx|}$$

Enumerating element of N for each $x \in X$, we get

$$|N| = \sum_{x \in X} |G_x| = \sum_{x \in X} \frac{|G|}{|Gx|} = |G| \sum_{x \in X} \frac{1}{|Gx|}$$

Let $Gx = \mathscr{O}$ be an orbit containing x with length k.

$$\implies \sum_{x \in \mathscr{O}} \frac{1}{|Gx|} = \sum_{i \in 1}^{k} \frac{1}{k} = 1$$

Let r be the number of orbits in X under the group action of G. Clearly, the orbits of X are disjoint. Thus,

$$|N| = |G| \sum_{i=1}^{r} \sum_{x \in \mathscr{O}_i} \frac{1}{|Gx|} = |G| \sum_{1=1}^{r} 1 = r|G|$$

Therefore, $\sum_{g \in G} |X_g| = |N| = r|G|$. □

Corollary 1.13.1.1. *If G is a finite group and X is a finite G-set, then*

$$\textit{number of orbits in } X \textit{ under } G = \frac{1}{|G|} \sum_{g \in G} |X_g| \tag{1.6}$$

Proof. From Burnside forumla, we have

$$\sum_{g \in G} |X_g| = r|G| \implies r = \frac{1}{|G|} \sum_{g \in G} |X_g|$$

□

1.14 Isomorphism Theorems 1-2

1.14.1 First Isomorphism Theorem

Theorem 1.14.1 (first isomorphism)**.** *Let $\phi : G \to G'$ be a homomorphism with kernel K, and let $\gamma_K : G \to G/K$ be the canonical homomorphism. There is a unique isomorphism $\mu : G/K \to \phi[G]$ such that $\phi(x) = \mu(\gamma_K(x))$ for each $x \in G$.*

Proof. By, fundamental homomorphism theorem.[Fraleigh, 2013, §14.1] □

$$\begin{tikzcd} G \arrow[rr, "\phi"] \arrow[dr, "\gamma_K"'] & & \phi[G] \\ & G/K \arrow[ur, dotted, "\mu"'] & \end{tikzcd}$$

Figure 1.1: First Isomorphism Theorem

1.14.2 Second Isomorphism Theorem

Definitions 1.14.1. Join $H \vee N$ is the smallest subgroup of G containing HN where $HN = \{hn : h \in H, n \in N\}$.

Lemma 1.14.2. *Let N be a normal subgroup of G and let $\gamma : G \to G/N$ be the canonical homomorphism. Then the map ϕ from the set of normal subgroups of G containing N to the set of normal subgroup of G/N given by $\phi(L) = \gamma[L]$ is one-to-one and onto.*

Proof. Let G be a group and N be a normal subgroup of G. Given that $\gamma : G \to G/N$ the canonical homomorphism. That is, $\gamma(g) = gN$ for every $g \in G$. Let L, M be normal subgroups of G containing N. Since homomorphism preserves normality, $\gamma(L)$ is a normal subgroup of $\gamma[G]$.

Suppose $\phi(L) = \phi(M)$. By definition of ϕ,

$$\gamma[L] = \phi(L) = \phi(M) = \gamma[M]$$

Since N is normal, $\gamma^{-1}(\gamma(g)) = gN$ for every $g \in G$. And

$$\gamma^{-1}(\gamma[L]) = \gamma^{-1}\left(\bigcup_{g\in L}\gamma(g)\right) = \bigcup_{g\in L}\gamma^{-1}(\gamma(g)) = \bigcup_{g\in L} gN = L$$

for every point $g \in L$, the left coset gN is contained in L. ($\because N \leq L$) Therefore,

$$\gamma^{-1}(\phi(L)) = \gamma^{-1}(\gamma[L]) = L$$

$$\gamma^{-1}(\phi(M)) = \gamma^{-1}(\gamma[M]) = M$$

Thus, $L = M$. Therefore, ϕ is injective.

Let H be a normal subgroup of G/N. Then $\gamma^{-1}(H)$ is a normal subgroup of G. We have, $eN \in H$ and $\gamma^{-1}(eN) = eN = N$. Thus, $N \subset \gamma^{-1}(H)$. Thus, there exists $\gamma^{-1}(H)$, a normal subgroup of G containing N such that $\phi(\gamma^{-1}(H)) = H$. Therefore, ϕ is surjective. $\square$

Lemma 1.14.3. *If N is a normal subgroup of G, then $H \cap N = HN = NH$. Furthermore, if H is also normal in G, then HN is normal in G.*

Proof. Let G be a group and N be a normal subgroup of G. Also let H be a subgroup of G.
Claim : $HN = \{hn \in G : h \in H,\ n \in N\}$ is a subgroup of G.

G1 Closure : $h_1n_1h_2n_2 = h_1(h_2n_3)n_2 = h_3n_4 \in HN$, where $h_1, h_2 \in H$ and $n_1, n_2, n_4 \in N$. Since N is normal, $n_1h_2 = h_2n_3$ for some $n_3 \in N$.

$$
\begin{tikzcd}
HN \arrow[rr, "\gamma_{|_{HN}}"] \arrow[dr] & & \gamma[H] \arrow[dl, dotted, dash, "\mu_1"'] \arrow[dr, dotted, dash, "\mu_2"] & & H \arrow[ll, "\gamma_{|_{H}}"'] \arrow[dl] \\
& HN/N \arrow[rr, dotted, dash, "\mu_2^{-1} \circ \mu_1"] & & H/H \cap N &
\end{tikzcd}
$$

Figure 1.2: Second Isomorphism Theorem

G2 $HN \subset G$, thus HN satisfies associativity.

G3 Since $H, N \leq G$, $e \in H$ and $e \in N$. Thus, $e = ee \in HN$.

G4 Let $h_1 n_1 \in HN$. $H \leq G$ and $h_1 \in H \implies h_1^{-1} \in H$. Then $n_1^{-1} h_1^{-1} = h_1^{-1} n_2$ for some $n_2 \in N$, since N is normal. Thus, every element in HN, $h_1 n_1$ has an inverse $h_1^{-1} n_2 \in HN$.

Let H be a normal subgroup of G. Let $g \in G$. Then $g(h_1 n_1) = (gh_1)n_1 = (h_2 g)n_1 = h_2(gn_1) = h_2(n_2 g) = (h_2 n_2)g$ for some $h_2 \in H$ and $n_2 \in N$ since both H and N are normal subgrops of G. Thus $gHN = HNg$ for every $g \in G$. Therefore, HN is a normal subgroup of G. □

Theorem 1.14.4 (second isomorphism)**.** *Let H be a subgroup of G and let N be a normal subgroup of G. Then $(HN)/N \simeq H/(H \cap N)$.*

Proof. Consider the canonical homomorphism $\gamma : G \to G/N$ with kernel N. Then γ restricted to HN is a homomorphism from HN onto $\gamma[H]$.

$$\gamma_{|_{HN}} : HN \to \gamma[H] \text{ where } \gamma_{|_{HN}}(hn) = \gamma(hn) = (hn)N = hN = \gamma(h) \tag{1.7}$$

Since $\ker(\gamma) = N$ and $N \subset HN$. We have, $\ker(\gamma_{|_{HN}}) = N$. By first isomorphism therorem there exists a unique isomorphism $\mu_1 : HN/N \to \gamma[H]$ where $\mu_1(hnN) = \gamma(h) = hN$.

Similarly, γ restricted to H is also a homomorphism onto $\gamma[H]$.

$$\gamma_{|_{H}} : H \to \gamma[H] \text{ where } \gamma_{|_{H}}(h) = \gamma(h) = hN. \tag{1.8}$$

Since $\ker(\gamma) = N$ and $H \cap N \neq \phi$. $\ker(\gamma_{|_{H}}) = H \cap N$. By first isomorphism theorem, there exists a unique isomorphism $\mu_2 : H/(H \cap N) \to \gamma[H]$

Then $HN/N \simeq H/(H \cap N)$, since the composition of two isomorphisms, $\mu_2^{-1} \circ \mu_1 : HN/N \to H/(H \cap N)$ is an isomorphism □

1.15 Third Isomorphism Theorem

Theorem 1.15.1 (third isomorphism)**.** *Let H and K be normal subgroup of G with $K \leq H$. Then $G/H \simeq (G/K)/(H/K)$.*

$$\begin{tikzcd}
G \arrow[rr, "\gamma_H"] \arrow[dr, "\gamma_K"'] \arrow[drr, "\phi"] & & G/H \arrow[d, dotted, dash, "\mu"] \\
& G/K \arrow[r, "\gamma_{H/K}"'] & (G/K)/(H/K)
\end{tikzcd}$$

Figure 1.3: Third Isomorphism Theorem

Proof. Consider the function $\phi : G \to (G/K)/(H/K)$ defined by $\phi(g) = gK(H/K)$. Then, ϕ is a homomorphism.

$$\begin{aligned}
\phi(ab) &= (ab)K/(H/K) \\
&= (aK)(bK)\ (H/K),\ \because (ab)K = (aK)(bK) \\
&= aK(H/K)\ bK(H/K),\ \because (xy)H/K = xH/K\ yH/K \\
&= \phi(a)\phi(b)
\end{aligned}$$

The kernel of ϕ is the set $\{a \in G : \phi(a) = H/K\}$. Since the coset H/K was originally the points in H, we have $\ker(\phi) = H$. By first isomorphism theorem, there exists a unique isomorphism $\mu : G/H \to (G/K)/(H/K)$. Thus, $G/H \simeq (G/K)/(H/K)$. □

1.16 Finite, non-abelian Groups

Theorem 1.16.1. *Let p be a prime. Let G be a group of order p^n. Let X be a finite G-set. Then $|X| \equiv |X_G| \pmod{p}$.*

Proof. Suppose there are r orbits in X. Choose an element from each orbit, say $x_1, x_2, \cdots, x_r$. We have,

$$|X| = \sum_{i=1}^{r} |Gx_i| \tag{1.9}$$

Let $X_G = \{x \in X : gx = x,\ \forall g \in G\}$. Then each element in X_G belongs to an orbit of length 1. Let $|X_G| = s$. Then,

$$|X| = |X_G| + \sum_{i=s+1}^{r} |Gx_i| \tag{1.10}$$

We have, $|Gx| = (G : G_x)$. Clearly, both G and G_x are groups of order a multiple of prime p. Thus, for every $x \in X$, $|Gx|$ is multiple of prime p. Therefore, $|X| \cong |X_G| \pmod{p}$. □

Challenge 2. Let p be a prime and G be a finite group G. Design a mechanism to enumerate all elements of order p ?

Definitions 1.16.1. Let G be a group. Let p be a prime. G is a p**-group** if every element of G has order a power of p. A subgroup H of G is a p**-subgroup** of G if every element of H has order a power of p.

Theorem 1.16.2 (Cauchy)**.** *Let p be a prime. Let G be a finite group and p divides $|G|$, then G has a subgroup of order p.*

Proof. Let X be the set of all n-tuples of elements of G such that the product of co-ordinates of each n-tuple is the identity element e of G.

$$X = \{(g_1, g_2, \cdots, g_p) \in G^p : g_1 g_2 \cdots g_p = e\} \tag{1.11}$$

Let $g_1, g_2, \cdots, g_{p-1}$ be any elements in G. Then $g_p = (g_1 g_2 \cdots g_{p-1})^{-1}$ is uniquely determined. That is, the $(p-1)$ co-ordinates of an element in X may be chosen in $|G|$ different ways. Thus, $|X| = |G|^{p-1}$. Since p divides $|G|$, p divides $|X|$.

We have $g_1(g_2 \cdots g_p) = (g_2 g_3 \cdots g_p) g_1$, since $g_2 g_3 \cdots g_p = g_1^{-1}$. Thus, for every $(g_1, g_2, \cdots, g_p) \in X$, the σ permutation of that element is in X. That is, $\sigma(g_1, g_2, \cdots, g_p) = (g_2, g_3, \cdots, g_p, g_1) \in X$. Similarly, $g_2(g_3 g_4 \cdots g_p g_1) = (g_3 g_4 \cdots g_p g_1) g_2$. And $\sigma^2(g_1, g_2, \cdots, g_p) = g_3, g_4, \cdots, g_p, g_1, g_2)$. Clearly, the subgroup generated by σ, is a subgroup of S_p and X is a $\langle\sigma\rangle$-set. We have, $|X| \cong |X_{\langle\sigma\rangle}| \pmod p$. Thus, p divides $|X_{\langle\sigma\rangle}|$.

Clearly, $(e, e, \cdots, e) \in X$ and $(e, e, \cdots, e) \in X_{\langle\sigma\rangle}$. Thus, $X_{\langle\sigma\rangle}$ has atleast p elements. Let $(g_1, g_2, \cdots, g_p) \in X_{\langle\sigma\rangle}$, then σ fixes $(g_1, g_2, \cdots, g_p)$.

$$\begin{aligned} \sigma(g_1, g_2, \cdots, g_p) &= (g_1, g_2, \cdots, g_p) \\ \implies (g_2, \cdots, g_p, g_1) &= (g_1, g_2, \cdots, g_p) \end{aligned}$$

Thus, $g_1 = g_2 = \cdots = g_p$, say $g \in G$. That $(g, g, \cdots, g) \in X$ and $g^p = e$ by the definition of X. Therefore, G has an element g of order p. □

Corollary 1.16.2.1. *Let G be a finite group. Then G is a p-group if and only if $|G|$ is a power of p.*

Proof. Let G be a finite group. Suppose G is a p-group. Suppose there exists another prime q, $q \neq p$ such that q divides $|G|$. Then by Cauchy's theorem, G has an element of order q. This contradicts the assumption that G is a p-group. Thus, the only prime that divides $|G|$ is p. Therefore, the order of G is a power of prime p.

Let $|G| = p^n$. Then the factors of p^n are powers of p. By Lagrange's theorem, order of subgroups of G must divide $|G|$. Thus, every subgroup of G has order a power of p. Thus, every element of G has order a power of p. Therefore, G is a p-group. □

Definitions 1.16.2. Let G be a group and H be a subgroup of G. Consider the inner automorphisms $i_g : G \to G$ such that $i_g(x) = gxg^{-1}$. We have, $i_g(H) = gHg^{-1}$ is the conjugate of the subgroup H. The set of all elements in G which has H itself as the conjugate of H is the normaliser of H in G.

$$N[H] = \{g \in G : gHg^{-1} = H\} \tag{1.12}$$

Remark. The normaliser of H in G is a subgroup of G. And $N[H]$ is the largest subgroup of G with H as its normal subgroup.

Lemma 1.16.3. *Let H be a p-subgroup of a finite group G. Then*

$$(N[H] : H) \cong (G : H) \pmod p \tag{1.13}$$

Proof. Let G be a finite group. Let H be a p-subgroup of G. Let $\mathscr{L}$ be the set of all left cosets of H in G. Then, $|\mathscr{L}| = (G : H)$.

Claim : $\mathscr{L}$ is an H-set with group action $h(xH) = (hx)H$. We have, $e(xH) = (ex)H = xH$ and $(g_1g_2)(xH) = (g_1g_2xH) = g_1(g_2xH) = g_1(g_2(xH))$.

Let $\mathscr{L}_H$ be the set of all left cosets that are fixed under action by all element of H.

$$\begin{aligned}\mathscr{L}_H &= \{xH \in \mathscr{L} : h(xH) = xH,\ \forall h \in H\}\\ &= \{xH \in \mathscr{L} : x^{-1}h(xH) = H,\ \forall h \in H\}\\ &= \{xH \in \mathscr{L} : (x^{-1}hx)H = H,\ \forall h \in H\}\\ &= \{xH \in \mathscr{L} : (x^{-1}hx) \in H,\ \forall h \in H\}\\ &= \{xH \in \mathscr{L} : x^{-1} \in N[H]\}\end{aligned}$$

Clearly, left cosets of $\mathscr{L}_H$ has all its elements contained in $N[H]$. Thus,

$$|\mathscr{L}_H| = (N[H] : H) \tag{1.14}$$

We have, $\mathscr{L}$ is an H-set. And H is a p-subgroup. Thus, H has order a power of prime p. Therefore,

$$|\mathscr{L}| \cong |\mathscr{L}_H| \pmod p \tag{1.15}$$

□

Corollary 1.16.3.1. *Let H be a p-subgroup of finite group G. If p divides $(G : H)$, then $N[H] \neq H$.*

Proof. We have, $(G : H) \cong (N[H] : H) \pmod p$. And p divides $(G : H)$. Thus, p divides $(N[H] : H)$. Therefore, $H \neq N[H]$. □

1.17 Sylow Theorems

Theorem 1.17.1 (First Sylow Theorem)**.** *Let G be a finite group of order $p^n m$ where $n \geq 1$ and p does not divide m. Then*

1. *G contains a subgroup of order p^i where $1 \leq i \leq n$.*
2. *Every subgroup H of G of order p^i is normal subgroup of the subgroup of order p^{i+1}, for $1 \leq i \leq n$.*

Proof. We have, G is finite group and p divides $|G|$. By Cauchy's theorem, G has a subgroup of order p.

Suppose G has a subgroup H of order p^i, where $(i < n)$. Then, H is a p-subgroup. Thus, $(G : H) \cong (N[H] : H) \pmod p$ where $N[H]$ is the normaliser of H in G. Clearly, p divides $(G : H)$. Thus, p divides $(N[H] : H)$.

And H is a normal subgroup of $N[H]$. Thus, we have factor group $N[H]/H$ and p divides the order of $N[H]/H$. By Cauchy's theorem, $N[H]/H$ has a subgroup K of order p. Consider the canonical homomorphism $\gamma : N[H] \to N[H]/H$ defined by $\gamma(x) = xH$. Then $\gamma^{-1}(K)$ is a subgroup of $N[H]$ of order p^{i+1} and contains H. Thus, H is a normal subgroup of $\gamma^{-1}(K)$. By mathematical induction, G has subgroups of order p^i for $i = 2, 3, \cdots, n$. □

Definitions 1.17.1. A Sylow p-subgroup of G is a maximal p-subgroup of G.

Theorem 1.17.2 (Second Sylow Theorem)**.** *Let P_1 and P_2 be two Sylow p-subgroups of G. Then P_1 and P_2 are conjugate subgroups of G.*

Proof. Let P_1 and P_2 be two Sylow p-subgroups of G. Let $\mathscr{L}$ be the set of all left cosets of P_1. Then, group P_2 act on $\mathscr{L}$ by $y(xP_1) = (yx)P_1$. Thus, $\mathscr{L}$ is a P_2-set. Therefore, $|\mathscr{L}| \cong |\mathscr{L}_{P_2}| \pmod p$.

Clearly $|\mathscr{L}| = (G : P_1)$. And p doesn't divide $|\mathscr{L}|$. Thus, p does not divide $|\mathscr{L}_{P_2}|$. Therefore, $|\mathscr{L}_{P_2}| > 0$.

Thus, $\mathscr{L}$ has at least an element xP_1 which is fixed in the action of every element in P_2. That is, $yxP_1 = xP_1$ for every $y \in P_2$. Therefore, $x^{-1}yxP_1 = P_1$ for every $y \in P_2$.

In other words, $x^{-1}yx \in P_1$ for every $y \in P_2$. Therefore, $x^{-1}P_2x \leq P_1$. But, $|P_1| = |P_2|$. Thus, $x^{-1}P_2x = P_1$. Therefore, P_1 and P_2 are conjugate subgroups of G. □

Theorem 1.17.3 (Third Sylow Theorem)**.** *If G is a finite group and p divides $|G|$, then the number of Sylow p-subgroups is congruent to* $1 \pmod p$ *and divides $|G|$.*

Proof. Let $\mathscr{S}$ be the set of all Sylow p-subgroups of G. Let P be a Sylow p-subgroup of G. The elements of P act on $\mathscr{S}$ by conjugation. Let $x \in P$ and $T \in \mathscr{S}$, then x carries T into xTx^{-1}. Clearly, $\mathscr{S}$ is a P-set.

We have, $|\mathscr{S}| \cong |\mathscr{S}_P| \pmod p$. Suppose $T \in \mathscr{S}_P$. Then, $xTx^{-1} = T$ for every $x \in P$. Thus, T is a normal subgroup of P. But, T and P are of the same order, since both are Sylow p-subgroups of G. Therefore, $T = P$. Thus, $\mathscr{S}_P = \{P\}$. And $|\mathscr{S}_P| = 1$. Thus, $|\mathscr{S}| \cong 1 \pmod p$.

Let G act of $\mathscr{S}$ by conjugation. Let $x \in G$ and $P \in \mathscr{S}$, then x carries P into xPx^{-1}. Clearly, $\mathscr{S}$ is a G-set. However, every Sylow p-subgroup of G are conjugates. Thus, thus every Sylow p-subgroup of G belong to the same orbit under conjugation action. We have, $|Gx| = (G : G_x)$ and thus length of orbits divides the order of G. Since the number of Sylow p-subgroups is same as the length of the orbit of P, the number of Sylow p-subgroup of G divides the order of G. □

Remark. Let G be a group of order 15. Let $p = 3$. By Third Sylow Theorem, the number of Sylow 3-subgroup of G is congruent to $1 \pmod 3$ and divides 15. We have, congruence class $\hat{1} = \{1, 4, 7, 10, 13\}$. Only 1 divides 15. Thus, there is only one Sylow 3-subgroup of G.

Let $p = 5$. By Third Sylow Theorem, the number of Sylow 5-subgroup of G is congruent to $1 \pmod 5$ and divides 15. We have, congruence class $\hat{1} = \{1, 6, 11\}$. Only 1 divides 15. Thus, there is only one Sylow 3-subgroup of G.

Remark. Let G be a group of order 255. And $1, 3, 5, 15, 17, 51, 85, 255$ are the divisors of 255. Let $p = 3$. By third Sylow theorem, either there is one or eighty-five Sylow 3-subgroups of G. Suppose there 85 Sylow 3-subgroups. Then there are 170 elements of order 3.

Let $p = 5$. By third Sylow theorem, either there are one or fifty-one Sylow 5-subgroups of G. Suppose there 51 Sylow 5-subgroups. Then there are 204 elements of order 5.

Let $p = 17$. By third Sylow theorem, there is exactly one Sylow 17-subgroup of G. Thus, there are exactly 16 elements of order 17 in G.

1.18 Sylow Theorem : Applications

Definitions 1.18.1. A group G is solvable if there is sequence of subgroups $\{e\} = H_0 \leq H_1 \leq \cdots H_n = G$ such that for $i = 0, 1, 2, \cdots$,

1. H_i is a normal subgroup of H_{i+1}
2. H_{i+1}/H_i is simple and
3. H_{i+1}/H_i is abelian.

Theorem 1.18.1. *Every group of prime order is solvable.*

Proof. □

Definitions 1.18.2. Let G be a finite group. Consider G acting on itself by conjugation. Then,

$$|G| = |Z(G)| + \sum_{s+1}^{r} |Gx_i| \tag{1.16}$$

This is the class equation of G. And each orbit of G under conjugation by itself is a conjugate class in G.

Theorem 1.18.2. *The center of a finite, nontrivial p-group is nontrivial.*

Proof. □

Lemma 1.18.3. *Let G be a group containing normal subgroups H and K such that $H \cap K = \{e\}$ and $H \vee K = G$. Then G is isomorphic to $H \times K$.*

Proof. □

Theorem 1.18.4. *For a prime number p, every group of G of order p^2 is abelian.*

Proof. □

1.19 Sylow Theorem : Further Applications

Theorem 1.19.1. *If H, K are subgroups of a group G. Then,*

$$|HK| = \frac{(|H|)(|K|)}{|H \cap K|} \tag{1.17}$$

Proof. □

1.19.1 Analysis of Finite Groups

The following are a few sample tests for finite groups. For different numbers, you may have to use a combination of these tests.

Remark. Finite, non-abelian, simple groups are of order 60, 168, 360, ...

Type 1 : p^r

Groups of order p^r for $r > 1$, are not simple as they have normal subgroup of order p^{r-1} by first Sylow theorem.

For example, any group G of order 16 is not simple as by first Sylow theorem G has a normal subgroup of order 8.

Type 2 : pq

Suppose G is a group with order pq where p, q are primes and $q > p$. Clearly, $p \not\cong 1 \pmod{q}$. Thus, by third Sylow theorem, the Sylow p-subgroup of G is unique, say H.

Now, by second Sylow theorem, Sylow p-subgroups are conjugate subgroups. Thus, this Sylow p-subgroup H is its only conjugate. That is, $gHg^{-1} = H$ for every $g \in G$. Thus, it is a normal subgroup of G. Thus G is not simple.

Further, if $q \not\cong 1 \pmod{p}$ for $q > p$, then group G is cyclic.

For example, $5 \not\cong 1 \pmod{3}$. Thus any group G of order 15 is cyclic. Clearly, G is abelian and is not simple.

However, $7 \cong 1 \pmod{3}$. Thus, groups of order 21 are not simple, but are not necessarily cyclic.

Type 3 : Only one Sylow p-subgroup

In this case, we don't have a strict form. However, the nature of prime factor of the order suggests that G has only one Sylow p-subgroup for some prime factor p of its order.

For example, group G of order 20 has only one Sylow 5-subgroup say, H (by third Sylow theorem). Thus, this subgroup H is a normal subgroup of G (by second Sylow theorem). Therefore, groups of order 20 are not simple.

Type 4 : Enumerating elements of Sylow p-subgroups

Groups of order 30 are not simple. — The distinct Sylow p-subgroups suggests $p - 1$ elements are unique to each. By counting, we can show that not all p-subgroups have a conjugate.

Type 5 : Applying the relation for $|HK|$

Groups of order 48 are not simple. — $|HK| =$

Type 6 : Normaliser of $H \cap K$

Groups of order 36 are not simple. — Let H, K be distinct Sylow 3-subgroups of G. Then, the normaliser of $H \cap K$ suggests a normal subgroup which is either $H \cap K$ or its normliser.

Type 7 : Commutator of G

Groups of order 255 are not simple. — Its commutator subgroup has order 1. Thus, the group is cyclic. Clearly, cyclic groups are abelian and not simple.

1.20 Rings, Fields & Integral Domains

Subject 2

ME010102 Linear Algebra

2.1 Vector Spces

2.2 Linear Transformations

2.3 Determinants

2.4 Elementary Canonical Forms

Subject 3

ME010103 Basic Topology

Subject 4

ME010103 Real Analysis

Module 1 : Bounded Variation & Rectifiable Curves

4.1 Functions of Bounded Variation & Rectifiable Curves

4.1.2 Properties of Monotone Functions

Theorem 4.1.1. *Let f be an increasing function defined on closed interval $[a, b]$. And let $x_0 = a < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$. Then,*

$$\sum_{k=1}^{n-1} [f(x_k+) - f(x_k-)] \leq f(b) - f(a)$$

Proof. Let f be an increasing function on $[a, b]$. Let $\{x_0 = a,\ x_1, \ldots,\ x_n = b\}$ be a partition of $[a, b]$. Let $y_k \in (x_{k-1}, x_k),\ \forall k$. Then, $f(y_k) \leq f(x_k+)$ and $f(x_k-) \leq f(y_{k-1})$. Therefore,

$$\sum_{k=1}^{n-1} [f(x_k+) - f(x_k-)] \leq \sum_{k=1}^{n-1} [f(y_k) - f(y_{k-1})] \leq f(b) - f(a)$$

□

In other words, for monotonic functions the sum of jumps is bounded.

Theorem 4.1.2. *Let f be a monotonic function defined on closed interval $[a, b]$. Then the set of discontinuities of f is countable.*

Proof. Without loss of generality, let f be an increasing function on $[a, b]$. Let S_m be the set of all points on $[a, b]$ at which the jump exceeds $\frac{1}{m}$.

We know that, the sum of jumps of an increasing function is bounded above by $f(b) - f(a)$. Thus, cardinality of S_m given by,

$$|S_m| < m[f(b) - f(a)]$$

is finite for any positive integer m.

If f is discontinuous at a pont $x \in [a,b]$, then there exists some integer m' such that $0 < \frac{1}{m'} < x$ and $x \in S_{m'}$.

$$\text{Number of discontinuities } = \left| \bigcup_{m=1}^{\infty} S_m \right| \leq \sum_{m=1}^{\infty} |S_m| \text{ is countable.}$$

since countable sum of finite values is countable. Therefore, the number of discontinuities of f is countable. □

4.1.3 Function of Bounded Variation

Definitions 4.1.1 (partition)**.** Let $[a,b]$ be a compact interval. Let $x_0 = a$, $x_0 < x_1 < x_2 < \cdots < x_n$ and $x_n = b$. Then $P = \{x_0, x_1, \ldots, x_n\}$ is a **partition** of $[a,b]$. And (x_{k-1}, x_k) is the **kth subinterval** of the partition.

Definitions 4.1.2 (bounded variation)**.** Let f be a function defined on closed interval $[a,b]$. If there exists a positive real-number M such that

$$\sum_{k=1}^{n} |\Delta f_k| = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| \leq M$$

for any partition P on $[a,b]$. Then f is a function of **bounded variation**, where (x_{k-1}, x_k) is the kth subinterval of the partition.

In other words, a function is of bounded variation on $[a,b]$ if the sum of variations is bounded for any (finite) partition of $[a,b]$.

Theorem 4.1.3. *Let f be a monotonic function on $[a,b]$. Then f is of bounded variation on $[a,b]$.*

Proof. Without loss of generality, let f be an increasing function. Then for any partition $P = \{x_0, x_1, \ldots, x_n\}$ of $[a,b]$, we have

$$\sum_{k=1}^{n} [f(x_k) - f(x_{k-1})] \leq f(b) - f(a) = M$$

Therefore, f is of bounded variation on $[a,b]$. □

Theorem 4.1.4. *Let f be a continuous function on $[a,b]$ and its derivative f' exists and f' is bounded in (a,b). Then f is of bounded variation.*

Proof. Let f be a continuous function with derivative f' on (a,b). Since f is continuous and f' exists, from intermediate value theorem we have

$$\Delta f_k = f(x_k) - f(x_{k-1}) = f'(t_k)[x_k - x_{k-1}] \text{ where } t_k \in (x_{k-1}, x_k)$$

Since f' is bounded, $f'(t_k) \leq M'$ for any $t_k \in (a,b)$. Thus,

$$\sum_{k=1}^{n} |\Delta f_k| = \sum_{k=1}^{n} |f'(t_k)|(x_k - x_{k-1}) \leq M' \sum_{k=1}^{n} x_k - x_{k-1} = M'(b-a) = M$$

Therefore, f is of bounded variation. □

Theorem 4.1.5. *Let f be a function on $[a, b]$. If f is of bounded variation, then f is bounded.*

Proof. Let $x \in (a, b)$. Consider the partition $P = \{a, x, b\}$. Since f is of bounded variation, there exists a positive real-number M such that

$$\sum_{k=1}^{2} |\Delta f_k| = |f(x) - f(a)| + |f(b) - f(x)| \leq M$$

Clearly, $|f(x) - f(a)| \leq M$, since $|f(b) - f(x)| > 0$. We have,

$$|f(x)| = |f(x) - f(a) + f(a)| \leq |f(x) - f(a)| + |f(a)| \leq M + |f(a)|$$

Suppose $x = a$, then $|f(x)| = |f(a)|$.
Suppose $x = b$, then $|f(x)| = |f(b)| = M' + |f(a)|$ where $M' = |f(b)| - |f(a)|$.

Therefore, the function f is bounded on $[a, b]$. □

4.1.4 Total Variation

Definitions 4.1.3 (total variation)**.** Let f be a function of bounded variation on $[a, b]$. Let $\Sigma(P)$ be the sum of variations with respect to the partition P of $[a, b]$. Then the **total variation** of the function f on $[a, b]$ is given by,

$$V_f(a, b) = V_f = \sup\{\Sigma(P) : P \in \mathscr{P}[a, b]\}$$

Note 1 : V_f is finite, since f is of boundned variation on $[a, b]$.

Note 2 : $V_f \geq 0$, since $|\Delta f_k| \geq 0$ for any subinterval of $[a, b]$.

Note 3 : $V_f = 0$ if only if f is a constant function on $[a, b]$. (Why ?)

Theorem 4.1.6. *Let f, g be functions of bounded variation on $[a, b]$. Then their sum $f + g$, difference $f - g$, and product fg are of bounded variation. Also,*

$$V_{f \pm g} \leq V_f + V_g \quad \text{and} \quad V_{fg} \leq AV_f + BV_g$$

where $A = \sup\{|g(x)| : x \in [a, b]\}$ and $B = \sup\{|f(x)| : x \in [a, b]\}$.

Proof. Let f, g be functions of bounded variation on $[a, b]$. Then f, g are bounded and $\sup |f(x)|$ and $\sup |g(x)|$ exists.

Step 1 : $V_{f\pm g} \leq V_f + V_g$
We have,

$$|(f + g)(x_k) - (f + g)(x_{k-1})| \leq |f(x_k) - f(x_{k-1})| + |g(x_k) - g(x_{k-1})|$$

Then

$$\begin{aligned}
V_{f+g} &= \sup_{P \in \mathscr{P}} \sum_{k=1}^{n} |(f + g)(x_k) - (f + g)(x_{k-1})| \\
&\leq \sup_{P \in \mathscr{P}} \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| + \sup_{P \in \mathscr{P}} \sum_{k=1}^{n} |g(x_k) - g(x_{k-1})| \\
&= V_f + V_g
\end{aligned}$$

Similarly, we have $V_{f-g} \leq V_f + V_g$ since

$$|(f-g)(x_k) - (f-g)(x_{k-1})| \leq |f(x_k) - f(x_{k-1})| + |g(x_k) - g(x_{k-1})|$$

Step 2 : $V_{fg} \leq AV_f + BV_g$

We have,

$$\begin{aligned}|fg(x_k) - fg(x_{k-1})| &= |f(x_k)g(x_k) - f(x_{k-1}g(x_{k-1})| \\ &= |f(x_k)g(x_k) - f(x_{k-1})g(x_k) + f(x_{k-1}g(x_k) - f(x_{k-1}g(x_{k-1})| \\ &\leq |g(x_k)|\ |f(x_k) - f(x_{k-1})| + |f(x_{k-1}|\ |g(x_k) - g(x_{k-1})| \\ &\leq A|f(x_k) - f(x_{k-1})| + B|g(x_k) - g(x_{k-1})|\end{aligned}$$

where $A = \sup\{|g(x)| : x \in [a,b]\}$ and $B = \sup\{|f(x)| : x \in [a,b]\}$.

Therefore,

$$\begin{aligned}V_{fg} &\leq A \sup_{P\in\mathscr{P}} \left\{\sum_{k=1}^{n} |f(x_k) - f(x_{k-1}|\right\} + B \sup_{P\in\mathscr{P}} \left\{\sum_{k=1}^{n} |g(x_k) - g(x_{k-1})|\right\} \\ &= AV_f + BV_g\end{aligned}$$

□

Definitions 4.1.4 (bounded away from zero)**.** A function f is bounded away from zero on $[a,b]$ if there exists a real-number m such that $0 < m \leq f(x)$, $\forall x \in [a,b]$.

Let f be a function of bounded variation. Then $\frac{1}{f}$ is of bounded variation if and only if f is bounded away from zero.(Why ?)

Theorem 4.1.7. *Let f be a function of bounded variation on $[a,b]$ and f is bounded away from zero. Then, $g = \frac{1}{f}$ is a function of bounded variation on $[a,b]$ and $V_g \leq \frac{V_f}{m^2}$ whenever $0 < m \leq |f(x)|$.*

Proof. Suppose f is of bounded variation on $[a,b]$ and f is bounded away from zero. That is, there exists positive real-number m such that $0 < m \leq |f(x)|$, $\forall x \in [a,b]$. Then, $\frac{1}{|f(x)|} \leq \frac{1}{m}$, $\forall x \in [a,b]$.

Define $g = \frac{1}{f}$. Then, we have

$$|\Delta g_k| = \left|\frac{1}{f(x_k)} - \frac{1}{f(x_{k-1})}\right| = \frac{|f(x_k) - f(x_{k-1})|}{|f(x_k)|\ |f(x_{k-1}|} \leq \frac{|f(x_k) - f(x_{k-1})|}{m^2}$$

The total variation of g on $[a,b]$ is given by,

$$\begin{aligned}V_g &= \sup_{P\in\mathscr{P}} \left\{\sum_{k=1}^{n} |g(x_k) - g(x_{k-1})|\right\} \\ &\leq \frac{1}{m^2} \sup_{P\in\mathscr{P}} \left\{\sum_{k=1}^{n} |f(x_k) - f(x_{k-1})|\right\} \\ &\leq \frac{V_f}{m^2} \text{ where } 0 < m \leq |f(x)|,\ \forall x \in [a,b]\end{aligned}$$

□

In other words, if h, f are functions of bounded variation of $[a, b]$. And f is bounded away from zero, then $g = \frac{1}{f}$ and $\frac{h}{f} = hg$ is a function of bounded variation and $V_{\frac{h}{f}} = V_{hg} \leq AV_h + BV_f$.

Now, we have analyzed sum, difference, product and quotient of functions of bounded variation. And we are not surprised about preservation of bounded variation under function composition. (Why ?)

4.1.5 Additive Property of Total Variation

Theorem 4.1.8 (additive property)**.** *Let f be a function of bounded variation on $[a, b]$. Let $c \in (a, b)$. Then f is of bounded variation on both $[a, c]$ and $[c, b]$. And, $V_f(a, b) = V_f(a, c) + V_f(c, b)$.*

Proof. Let f be a function of bounded variation on $[a, b]$. Let P_1, P_2 be partitions of $[a, c]$ and $[c, b]$ respectively. Then $P_0 = P_1 \cup P_2$ is a partition of $[a, b]$. Thus,

$$\sum(P_1) + \sum(P_2) = \sum(P_0) \leq V_f(a, b)$$

Taking supremums on the left side, we get

$$V_f(a, c) + V_f(c, b) \leq V_f(a, b)$$

Clearly, $V_f(a, c) \leq V_f(a, b)$ and $V_f(c, b) \leq V_f(a, b)$. Therefore, function f is of bounded variation on both $[a, c]$ and $[c, b]$.

Let P be a paritition of $[a, b]$. Then $P_0 = P \cup \{c\}$ is refinement of P. Now, we have two partitions P_1, P_2 of $[a, c]$ and $[c, b]$ such that $P_0 = P_1 \cup P_2$. Thus,

$$\sum(P) \leq \sum(P_0) = \sum(P_1) + \sum(P_2) \leq V_f(a, c) + V_f(c, b)$$

Taking supremum on the left side, we get

$$V_f(a, b) \leq V_f(a, c) + V_f(c, b)$$

Therefore,

$$V_f(a, b) = V_f(a, c) + V_f(c, b), \quad \forall c \in (a, b)$$

□

4.1.6 Total Variation on $[a, x]$ as a function of x

By additive property, we have existence of $V_f(a, x)$ for every $x \in (a, b]$. Assigning $V(x) = V_f(a, x)$, we have a well-defined function on $(a, b]$.

Theorem 4.1.9. *Let f be a function of bounded variation on $[a, b]$. Define $V : [a, b] \to \mathbb{R}$ given by,*

$$V(x) = \begin{cases} V_f(a, x) & x \in (a, b] \\ 0 & x = a \end{cases}$$

Then,

1. *V is an increasing function on $[a, b]$.*

2. *$V - f$ is an increasing function of $[a, b]$.*

Proof. Suppose $x = a$. Then $V(x) = 0$ and $V(y) = V_f(a, y) \geq 0 = V(x)$.
Suppose $x \neq a$. Then, $a < x < y \leq b$. By additive property of total variation, we have $V(y) = V_f(a, y) = V_f(a, x) + V_f(x, y) = V(x) + V_f(x, y)$. Since $V_f(x, y) \geq 0$, we have $V(x) \leq V(y)$. Therefore, V is an increasing function on $[a, b]$.

Define $D : [a, b] \to \mathbb{R}$ given by $D(x) = (V - f)(x) = V(x) - f(x)$. Suppose $a \leq x < y \leq b$. Then, $D(y) - D(x) = V(y) - f(y) - V(x) + f(x) = [V(y) - V(x)] - [f(y) - f(x)]$. We have, $V(y) = V_f(a, y) = V_f(a, x) + V_f(x, y)$. Thus, $V(y) - V(x) = V_f(x, y)$. Also we have, f is of bounded variation on $[x, y]$. Consider the trivial partition $P = \{x, y\}$ of $[x, y]$. Then, we have $f(y) - f(x) = \sum(P) \leq V_f(x, y)$. Thus, $D(y) - D(x) = V_f(x, y) - [f(y) - f(x)] \geq 0$. Therefore, $D = V - f$ is an increasing function on $[a, b]$. □

4.1.7 Function of bounded variation expressed as the difference of increasing functions

Theorem 4.1.10. *Let f be a function on $[a, b]$. Function f is of bounded variation on $[a, b]$ if and only if f can be expressed as difference of two increasing functions.*

Proof. Let f be a function of bounded variation, then $f = V - D$ where total variation V and $D = V - f$ are both increasing.

Let f be function on $[a, b]$. Let $f = V - D$ where V, D are increasing functions. Then V, D are of bounded variation, since monotonic functions on $[a, b]$ are of bounded variation. Also, we have $V - D$ is of bounded variation, since for any two functions of bounded variation their difference is also of bounded variation. □

4.1.8 Continuous functions of bounded variation

Theorem 4.1.11. *Let f be a function of bounded variation on $[a, b]$. Let V be the total variation function of f defined on $[a, b]$. f is continuous at a point if and only if V is continuous at that point.*

Proof. Let f be a function of bounded variation on $[a, b]$. Let V be the total variation of f. Let $x, y \in [a, b]$, such that $x < y$. We have V is an increasing function on $[a, b]$. And f is difference of two increasing functions on $[a, b]$. Thus, $f(x+)$, $f(x-)$, $V(x+)$, $V(x-)$ exists for any $x \in (a, b)$. It remains to prove that V, f are continuous at $x \in (a, b)$.

Part 1 : V continuous $\implies$ f continuous
Suppose $x \neq a$, then $a < x < y \leq b$. Let P be any partition on $[a, x]$. Then there exists a partition P' on $[a, y]$ such that $P \subset P'$. Consider, $P' = P \cup \{y\}$. Then, $V(y) > V(x)$ and $V(y) - V(x) \geq |f(y) - f(x)|$. Thus,

$$0 \leq |f(y) - f(x)| \leq V(y) - V(x)$$

The inequality is true for any $y > x$. Therefore,

$$0 \le |f(x+) - f(x)| \le V(x+) - V(x) \text{ as } y \to x+$$

Similarly, let $a < z < x$. Then,

$$0 \le |f(x) - f(x-)| \le V(x) - V(x-) \text{ as } z \to x-$$

Clearly, if V is continuous at x then f is continuous x.

Part 2 : f continuous $\implies$ V continuous
Suppose f is continuous at $c \in (a, b)$. Let $\varepsilon > 0$. We have,

$$V_f(c, b) = \sup_{P \in \mathscr{P}} \sum(P)$$

Thus, there exists a partition $P_1 \in \mathscr{P}[c, b]$ such that $V_f(c, b) - \frac{\varepsilon}{2} < \sum(P_1)$. Since f is continuous at c, there exists $x_1 \in (c, b)$ such that $|f(x_1) - f(c)| < \frac{\varepsilon}{2}$. Then,

$$V_f(c, b) - \frac{\varepsilon}{2} < \frac{\varepsilon}{2} + V_f(x_1, b)$$

We have,

$$\begin{aligned} V(x_1) - V(c) &= V_f(a, x_1) - V_f(a, c) = V_f(c, x_1) \\ &= V_f(c, b) - V_f(x_1, b) < \varepsilon \end{aligned}$$

Clearly, $V(c + h) \to V(c)$ as $h \to 0$. That is, $V(c+) = V(c),\ \forall c \in [a, b)$. Similarly we have,

$$V_f(a, c) = \sup_{P \in \mathscr{P}} \sum(P)$$

And there exists $P_2 \in \mathscr{P}[a, c]$ such that

$$V_f(a, c) - \frac{\varepsilon}{2} < \sum(P_2)$$

And there exists $x_2 \in (a, c)$ such that $|f(c) - f(x_2)| < \frac{\varepsilon}{2}$, since f is continuous at c. Therefore,

$$V(c) - V(x_2) = V_f(a, c) - V_f(a, x_2) = V(x_2, c) < \varepsilon$$

Thus, $V(c-) = V(c),\ \forall c \in (a, b]$. Therefore, V is continuous at c. $\square$

Theorem 4.1.12. *Let f be a continuous function on $[a, b]$. Function f is of bounded variation on $[a, b]$ if and only if f can be expressed as difference of two increasing continuous functions.*

Proof. Let f be a continuous function on $[a, b]$. Then total variation V is a continuous increasing function on $[a, b]$. Clearly, $D = V - f$ is also a continuous, increasing function on $[a, b]$. Therefore, $f = V - D$ where V, D are continuous, increasing functions.

Let V, D be continuous, increasing functions on $[a, b]$. Then $f = V - D$ is also a continuous function on $[a, b]$. We have, V, D are increasing functions, therefore both V, D are of bounded variation and their difference f is also of bounded variation. $\square$

Suppose f is of bounded variation on $[a, b]$. Let $id : [a, b] \to [a, b]$ where $id(x) = x$. Then $V + id$ is a strictly increasing function and $D = V + id - f$ is also strictly increasing. Thus, any function of bounded variation on $[a, b]$ can be characterised as difference of two strictly increasing continuous functions on $[a, b]$.

4.1.9 Curves and Paths

Definitions 4.1.5 (path)**.** Let $f : [a, b] \to \mathbb{R}^n$ be a continuous, vector-valued function. Then f is a path in $\mathbb{R}^n$. And f is a motion if $[a, b]$ is a time interval.

4.1.10 Rectifiable paths and Arc length

Definitions 4.1.6 (rectifiable path)**.** Let $f : [a, b] \to \mathbb{R}^n$ be a path in $\mathbb{R}^n$. Let $P = \{t_0, t_1, \ldots, t_m\}$ be a partition of $[a, b]$.

$$\Lambda_f(P) = \sum_{k=1}^{m} \|f(t_k) - f(t_{k-1})\| = \sum_{k=1}^{m} \|\Delta f_k\|$$

If $\Lambda_f(P)$ is bounded for any partition $P \in \mathscr{P}[a, b]$, then path f is **rectifiable**. If $\Lambda_f(P)$ is unbounded, then f is **nonrectifiable**.

Definitions 4.1.7 (arc length)**.** Let $f : [a, b] \to \mathbb{R}^n$ be a rectifiable path. Then **arc length** of path f is given by,

$$\Lambda_f(a, b) = \sup_{P \in \mathscr{P}} \Lambda_f(P)$$

Theorem 4.1.13. *A path $f : [a, b] \to \mathbb{R}^n$ is rectifiable if and only if each component f_k of f is of bounded variation on $[a, b]$. Let $V_k(a, b)$ be the total variation of f_k on $[a, b]$. Then,*

$$V_k(a, b) \le \Lambda_f(a, b) \le V_1(a, b) + V_2(a, b) + \cdots + V_n(a, b)$$

Proof. Let $x_j \in \mathbb{R}^n$ for $j = 1, 2, \ldots, m$. Then,

$$|x_r| \le \sqrt{\sum_{j=1}^{n} |x_j|^2} \le \sum_{j=1}^{n} |x_j| \text{ since } \sum_{j=1}^{n} x_j^2 \le \left(\sum_{j=1}^{n} x_j\right)^2$$

Let $f : [a, b] \to \mathbb{R}^n$ be a path in $\mathbb{R}^n$. Then $f = (f_1, f_2, \ldots, f_n)$ where f_k's are components of the path f. Let $P = \{t_0, t_1, \ldots, t_m\}$ be a partition of $[a, b]$. Now $f_r(t_j) - f_r(t_{j-1}) \in \mathbb{R}^n$ for each subinterval of $[a, b]$ and each component of f. Thus for each subinterval (t_j, t_{j-1}) we have,

$$|f_r(t_j) - f_r(t_{j-1})| \le \|f(t_j) - f(t_{j-1})\| \le \sum_{j=1}^{n} |f_k(t_j) - f_k(t_{j-1})|$$

Adding inequalities for every subinterval of the partition, we get

$$\sum_{k=1}^{m} |f_k(t_j) - f_k(t_{j-1})| \le \sum_{k=1}^{m} \|f(t_j) - f(t_{j-1})\| \le \sum_{k=1}^{m} \sum_{j=1}^{n} |f_k(t_j) - f_k(t_{j-1})|$$

Rearranging summation, we get

$$\sum(P) \leq \Lambda_f(P) \leq \sum_{j=1}^{n} \left(\sum(P)_{f_j}\right)$$

Suppose f is a rectifiable path. Then f has finite arc length $\Lambda_f(a,b)$. Let f_k be a component function of f. Let P be a partition of $[a,b]$. Then,

$$\sum(P) \leq \Lambda_f(P) \leq \Lambda_f(a,b)$$

Thus, $\sum(P)$ is bounded for any partition P. Therefore, total variation $V_k(a,b)$ exists for each component function f_k. And $V_k(a,b) \leq \Lambda_f(a,b)$.

Suppose f_k's are of bounded variation. Then

$$\Lambda_f(P) \leq \sum_{j=1}^{n} \sum(P) \leq \sum_{j=1}^{n} V_j(a,b)$$

is bounded for any partition P. Thus, f is rectifiable. And, arc length $\Lambda_f(a,b) \leq \sum_{k=1}^{n} V_k(a,b)$. Combining the inequalities, we get

$$V_k(a,b) \leq \Lambda_f(a,b) \leq \sum_{j=1}^{n} V_j(a,b)$$

□

4.1.11 Additive and Continuity Properties of Arc length

Theorem 4.1.14 (additive). *Let $f : [a,b] \to \mathbb{R}^n$ be a rectifiable path. Let $c \in (a,b)$. Then,*

$$\Lambda_f(a,b) = \Lambda_f(a,c) + \Lambda_f(c,b)$$

Proof. Let P be a partition of $[a,b]$. Then $P' = P \cup \{c\}$ is a refinement of P such that $P' = P_1 \cup P_2$ where P_1, P_2 are partition of $[a,c]$ and $[c,b]$ respectively. We have,

$$\Lambda_f(P) \leq \Lambda_f(P') = \Lambda_f(P_1) + \Lambda_f(P_2)$$

This inequality if true for any partition of $[a,b]$. Thus,

$$\Lambda_f(a,b) \leq \Lambda_f(a,c) + \Lambda_f(c,b)$$

Let P_1, P_2 be partition of $[a,c]$ and $[c,b]$ respectively. Then,

$$\Lambda_f(P_1) + \Lambda_f(P_2) \leq \Lambda_f(P) \leq \Lambda_f(a,b)$$

This inequality if true for any paritions on $[a,c]$ and $[c,b]$. Thus,

$$\Lambda_f(a,c) + \Lambda_f(c,b) \leq \Lambda_f(a,b)$$

□

Theorem 4.1.15 (continuity)**.** *Let $f : [a, b] \to \mathbb{R}^n$ be rectifiable path. Let function $s : [a, b] \to \mathbb{R}$ defined by*

$$s(x) = \begin{cases} 0 & x = a \\ \Lambda_f(a, x) & x \in (a, b] \end{cases}$$

Then,

1. *function s is continuous and increasing on $[a, b]$.*
2. *if there is no subinterval of $[a, b]$ in which f is constant, then s is strictly increasing.*

Proof. Let $a \leq x < y \leq b$. Then, $s(y) - s(x) = \Lambda_f(x, y) = \Lambda(a, y) - \Lambda(a, x) \geq 0$. Therefore, s in increasing.

Suppose f is not constant in any subinterval $[x, y]$ of $[a, b]$. Suppose s is not strictly increasing. Then, there exists $x, y \in (a, b)$ such that $x < y$ and $s(y) - s(x) = 0$. Thus,

$$\Lambda_f(a, y) - \Lambda(a, x) = \Lambda_f(x, y) = 0$$

Thus, $V_k(x, y) = 0,\ \forall k$ which is a contradition since f is not constant in $[x, y]$. Therefore, s is strictly increasing. □

4.1.12 Equivalence of path, Change of parameter

Definitions 4.1.8 (change of parameter)**.** Let $f : [a, b] \to \mathbb{R}^n$ be a path. Let $g : [c, d] \to \mathbb{R}^n$ be another path. Then f, g are **equivalent** if there exists a continuous, real-valued function, $u : [c, d] \to [a, b]$ such that $g = f \circ u$. That is, $g(t) = f(u(t)),\ \forall t \in [c, d]$. In other words, f, g are different parametric representations of a common graph.

Function u defines a change of parameter. If u is strictly increasing, then f, g are in the same direction. And u is an orientation preserving change of parameter. If u is strictly decreasing, then f, g are in opposite directions. And u is an orientation reversing change of parameter.

Theorem 4.1.16 (change of parameter)**.** *Let $f : [a, b] \to \mathbb{R}^n$ and $g : [a, b] \to \mathbb{R}^n$ be two paths. Let f, g be both injective functions. Then f and g are equivalent if they have the same graph.*

Proof. Let $f : [a, b] \to \mathbb{R}^n$ and $g : [c, d] \to \mathbb{R}^n$ be continuous, injective, vector-valued functions. Suppose f, g are equivalent paths, then f, g have the same graph.

Suppose f, g have the same graph. Since f is injective and continuous on its domain $[a, b]$, function f^{-1} exists and is continuous on its graph.

$$\text{Define } u : [c, d] \to [a, b],\ u(t) = f^{-1}(g(t))$$

Then u is continuous and $g(t) = f(u(t))$. Suppose u is not a strictly monotonic function. Since u is continuous, there exists $t_1, t_2 \in [c, d]$ such that $u(t_1) = u(t_2)$. Then $f(u(t_1)) = f(u(t_2)) \implies g(t_1) = g(t_2)$ which is a contradiction since g is injective on $[c, d]$. □

Module 2 : Riemann-Stieltjes Integral

4.2 The Riemann-Stieltjes Integral

Definitions 4.2.1 (unti step)**.** The unit step function $I : \mathbb{R} \to \mathbb{R}$ is defined by

$$I(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$$

Definitions 4.2.2 (Riemann Integral)**.** Let f be a bounded real function defined on $[a, b]$. Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$. Let $M_k = \sup\{f(x) : x \in [x_{k-1}, x_k]\}$ and $m_k = \inf\{f(x) : x \in [x_{k-1}, x_k]$.

Then Riemann upper sum of function f with respect to parition P,

$$U(P, f) = \sum_P M_k \Delta x_k = \sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

And Riemann lower sum of f with respect to P,

$$L(P, f) = \sum_{k=1}^{n} m_k(x_k - x_{k-1})$$

Now, Riemann upper integral of f over $[a, b]$,

$$\overline{\int_a^b} f\ dx = inf\{U(P, f) : P \in \mathscr{P}[a, b]\}$$

And, Riemann lower integral of f over $[a, b]$,

$$\underline{\int_a^b} f\ dx = sup\{L(P, f) : P \in \mathscr{P}[a, b]\}$$

A function f is Riemann integrable over $[a, b]$ if Riemann lower and upper integrals of f over $[a, b]$ are the same. Then Riemann integral of f over $[a, b]$,

$$\int_a^b f\ dx = \overline{\int_a^b} f\ dx = \underline{\int_a^b} f\ dx$$

Definitions 4.2.3 (Riemann-Stieltjes Integral)**.** Let f be a bounded function on $[a, b]$. Let α be an increasing function on $[a, b]$. Let $P = \{x_0, x_1, \ldots, x_n\}$ be a partition of $[a, b]$. Then, the Riemann-Stieltjes upper sum of f with respect to partition P and increasing function α,

$$U(P, f, \alpha) = \sum_{k=1}^{n} M_k \Delta\alpha_k$$

where $M_k = \sup\{f(x) : x \in [x_{k-1}, x_k]\}$ and $\Delta\alpha_k = \alpha(x_k) - \alpha(x_{k-1})$. Similarly, Riemann-Stieltjes lower sum,

$$L(P, f, \alpha) = \sum_{k=1}^{n} m_k \Delta\alpha_k$$

where $m_k = \inf\{f(x) : x \in [x_{k-1}, x_k]\}$. And function f is Riemann-Stieltjes integrable if Riemann Stieltjes upper and lower integrals are the same.

$$\int_a^b f\ d\alpha = \overline{\int_a^b} f\ d\alpha = \underline{\int_a^b} f\ d\alpha$$

where $\overline{\int_a^b} f\ d\alpha = \overline{\int_a^b} f\ d\alpha(x) = \inf\{U(P, f, \alpha) : P \in \mathscr{P}[a, b]\}$ and
$\underline{\int_a^b} f\ d\alpha = \sup\{L(P, f, \alpha) : P \in \mathscr{P}[a, b]\}$.

We write $f \in \mathscr{R}(\alpha)$ on $[a, b]$, which means that a bounded, real function f is Riemann-Stieltjes integrable on $[a, b]$ with respect to the increasing function α.

Note : function $\alpha : [a, b] \to \mathbb{R}$ is monotonic (increasing), but not necessarily continuous.

Remark : With $\alpha = id$ identity function, Riemann-Stieltjes integral is Riemann integral itself. That is, Riemann integral is a special case of Riemann-Stieltjes integral.

Theorem 4.2.1. *Let P^* be a refinement of a parition P of $[a, b]$. Then, $L(P, f, \alpha) \le L(P^*, f, \alpha)$ and $U(P^*, f, \alpha) \le U(P, f, \alpha)$.*

Proof. Let $P^* = P \cup \{x\}$ where x belongs to the ith subinterval (x_{j-1}, x_j). Define $w_1 = \min\{f(x) : x \in (x_{j-1}, x)\}$ and $w_2 = \min\{f(x) : x \in (x, x_j)\}$. Clearly, $w_1, w_2 \ge \min\{f(x) : x \in (x_{j-1}, x_j)\} = m_i$.

$$\begin{aligned} L(P, f, \alpha) - L(P^*, f, \alpha) &= m_i \Delta\alpha_j - w_1(\alpha(x) - \alpha(x_{j-1})) - w_2(\alpha(x_j) - \alpha(x)) \\ &= (m_i - w_1)(\alpha(x) - \alpha(x_{j-1})) + (m_i - w_2)(\alpha(x_j) - \alpha(x)) \\ &\ge 0 \end{aligned}$$

By mathematical induction the inequality is true for any refinement P^* of P. Therefore, $L(P, f, \alpha) \ge L(P^*, f, \alpha)$.

Similarly, define $W_1 = \max\{f(x) : x \in (x_{j-1}, x)\}$ and $W_2 = \max\{f(x) : x \in (x, x_j)\}$ where $W_1, W_2 \le \max\{f(x) : x \in (x_{j-1}, x_j)\} = M_i$.

$$\begin{aligned} U(P, f, \alpha) - U(P^*, f, \alpha) &= M_i \Delta\alpha_j - W_1(\alpha(x) - \alpha(x_{j-1})) - W_2(\alpha(x_j) - \alpha(x)) \\ &= (M_i - W_1)(\alpha(x) - \alpha(x_{j-1})) + (M_i - W_2)(\alpha(x_j) - \alpha(x)) \\ &\le 0 \end{aligned}$$

Again, the result is true for any refinement P and we have

$$U(P, f, \alpha) \le U(P^*, f, \alpha)$$

□

Theorem 4.2.2. *Let f be a bounded, real function on $[a, b]$ and α increasing function on $[a, b]$. Then,*

$$\underline{\int_a^b} f\ d\alpha \le \overline{\int_a^b} f\ d\alpha$$

Proof. Let P_1, P_2 be any two partition of $[a, b]$. Let $P^* = P_1 \cup P_2$ be a refinement of both partitions. Then, we have $L(P^*, f, \alpha) \leq U(P^*, f, \alpha)$. And,

$$\underline{\int_a^b} f \; d\alpha \leq L(P_1, f, \alpha) \text{ and } U(P_2, f, \alpha) \leq \overline{\int_a^b} f \; d\alpha$$

Therefore,

$$L(P_1, f, \alpha) \leq L(P^*, f, \alpha) \leq U(P^*, f, \alpha) \leq U(P_2, f, \alpha)$$

Clearly, the inequality holds independent of the choice of the partition. Taking supremum on right and infimum on left, we get

$$\sup_{P \in \mathscr{P}} L(P, f, \alpha) \leq \inf_{P \in \mathscr{P}} U(P, f, \alpha)$$

Therefore,

$$\underline{\int_a^b} f \; d\alpha \leq \overline{\int_a^b} f \; d\alpha$$

□

Theorem 4.2.3 (criterion for integrability)**.** *Let f be a bounded, real function on $[a, b]$. Let α be an increasing function on $[a, b]$. Then, f is Riemann-Stieltjes integrable on $[a, b]$ with repesct to α if and only if for every $\varepsilon > 0$ there exists a partition P of $[a, b]$ such that $U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$.*

In other words, $f \in \mathscr{R}(\alpha)$ on $[a, b]$ if and only if

$$\forall \varepsilon > 0, \exists P \in \mathscr{P}[a, b] \text{ such that } U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$$

Proof. Let $\varepsilon > 0$. Suppose there exists a partition P of $[a, b]$ such that $U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$. We have,

$$L(P, f, \alpha) \leq \underline{\int_a^b} f \; d\alpha \leq \overline{\int_a^b} f \; d\alpha \leq U(P, f, \alpha)$$

$$\overline{\int_a^b} f \; d\alpha - \underline{\int_a^b} f \; d\alpha \leq U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$$

Clearly, $\underline{\int_a^b} f \; d\alpha = \overline{\int_a^b} f \; d\alpha$. Therefore, $f \in \mathscr{R}(\alpha)$.

Suppose $f \in \mathscr{R}(\alpha)$ on $[a, b]$. Then by the definition of infimum there exists a partition P_1 of $[a, b]$ such that

$$U(P_1, f, \alpha) - \overline{\int_a^b} f \; d\alpha < \frac{\varepsilon}{2}$$

Similarly, there exists a partition P_2 of $[a, b]$ such that

$$\underline{\int_a^b} f \; d\alpha - L(P_2, f, \alpha) < \frac{\varepsilon}{2}$$

Consider $P^* = P_1 \cup P_2$. Clearly, P^* is a refinement for both P_1 and P_2. Thus,

$$U(P^*, f, \alpha) - \overline{\int_a^b} f \, d\alpha < \frac{\varepsilon}{2}$$

And,

$$\underline{\int_a^b} f \, d\alpha - L(P^*, f, \alpha) < \frac{\varepsilon}{2}$$

Thus,

$$U(P^*, f, \alpha) - \overline{\int_a^b} f \, d\alpha + \underline{\int_a^b} f \, d\alpha - L(P^*, f, \alpha) < \varepsilon$$

Given, $f \in \mathscr{R}(\alpha)$ on $[a, b]$. Therefore, $U(P^*, f, \alpha) - L(P^*, f, \alpha) < \varepsilon$. □

Theorem 4.2.4. *Suppose $\varepsilon > 0$ and $U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$ for some partition P of $[a, b]$.*

1. *The inequaility is true for any refinement of P.*
2. *Let $s_i, t_i \in [x_{i-1}, x_i]$ for each subinterval of the partition P.*

$$\sum_{i=1}^{n} \left| f_(s_i) - f(t_i) \right| \; \Delta\alpha_i < \varepsilon$$

3. *If $f \in \mathscr{R}(\alpha)$ and $t_i \in [x_{i-1}, x_i]$ for each subinterval of the partition P, then*

$$\left| \sum_{i=1}^{n} f(t_i)\Delta\alpha_i - \int_a^b f \, d\alpha \right| < \varepsilon$$

Proof. 1. Let $\varepsilon > 0$. Suppose $U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$. Let P^* be a refinement of P. We have $U(P, f, \alpha) \geq U(P^*, f, \alpha)$ and $L(P, f, \alpha) \leq L(P^*, f, \alpha)$. Thus,

$$U(P^*, f, \alpha) - L(P^*, f, \alpha) \leq U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$$

2. Let $s_i, t_i \in (x_{i-1}, x_i)$. Clearly, for each subinterval (x_{i-1}, x_i), we have $m_i \leq f(s_i) \leq M_i$ and $m_i \leq f(t_i) \leq M_i$. Thus,

$$|f(t_i) - f(s_i)| \leq M_i - m_i$$

$$\sum_{i=1}^{n} |f(t_i) - f(s_i)|\Delta\alpha_i \leq \sum_{i=1}^{n}(M_i - m_i)\Delta\alpha_i \leq U(P, f, \alpha) - L(P, f, \alpha) \leq \varepsilon$$

3. Let $\varepsilon > 0$. Let P be a partition of $[a, b]$ such that $U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$. We have,

$$L(P, f, \alpha) \leq \underline{\int_a^b} f \, d\alpha \leq \overline{\int_a^b} f \, d\alpha \leq U(P, f, \alpha)$$

Suppose $f \in \mathscr{R}(\alpha)$ over $[a, b]$. Then,

$$L(P, f, \alpha) \leq \int_a^b f \, d\alpha \leq U(P, f, \alpha)$$

Let $t_i \in (x_{i-1}, x_i)$ for each subinterval of the partition P. Clearly, $m_i \leq f(t_i) \leq M_i$. Thus, we also have

$$L(P, f, \alpha) \leq \sum_{i=1}^{n} f(t_i)\Delta\alpha_i \leq U(P, f, \alpha)$$

Therefore,

$$\left| \sum_{i=1}^{n} f(t_i)\ \Delta\alpha_i - \int_a^b f\ d\alpha \right| \leq U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$$

□

Theorem 4.2.5. *If f is continuous on $[a, b]$, then $f \in \mathscr{R}(\alpha)$ on $[a, b]$.*

Proof. Let f be a continuous function on $[a, b]$. Since continuous functions defined on compact subsets metric spaces are uniformly continuous, we have f is uniformly continuous.

Let $\varepsilon > 0$. Choose $\eta > 0$ such that $[\alpha(b) - \alpha(a)]\eta < \varepsilon$. Since f is uniformly continuous, there exists $\delta > 0$ such that $|f(x) - f(t)| < \eta$ whenever $|x - t| < \delta$. Consider the partition P such that each subinterval is of length less than δ. For any $s_i, t_i \in (x_{i-1}, x_i)$, we have

$$\begin{aligned}
\sum_{i=1}^{n} [f(t_i) - f(s_i)]\ [\alpha(x_i) - \alpha(x_{i-1})] &\leq \sum_{i=1}^{n} \eta[\alpha(x_i) - \alpha(x_{i-1})] \\
&\leq \eta \sum_{i=1}^{n} [\alpha(x_i) - \alpha(x_{i-1})] \\
&\leq \eta[\alpha(b) - \alpha(a)] \\
&< \varepsilon
\end{aligned}$$

Clearly, the result is true for minimum and maximum values of f in each subinterval of P. Therefore, we have

$$U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$$

Thus, $f \in \mathscr{R}(\alpha)$ over $[a, b]$. □

Theorem 4.2.6. *If f is monotonic on $[a, b]$ and α is continuous on $[a, b]$, then $f \in \mathscr{R}(\alpha)$ on $[a, b]$.*

Proof. Let f be increasing on $[a, b]$. Let α is continuous on $[a, b]$. Let n be any integer. We can construct a partition P of $[a, b]$ such that each the variation of α in each subinterval is fixed. That is, $\Delta\alpha_i = \frac{\alpha(b) - \alpha(a)}{n}$. Since f is increasing, in each subinterval $[x_{i-1}, x_i]$, we have $M_i = f(x_i)$ and $m_i = f(x_{i-1})$. Now we have,

$$\begin{aligned}
U(P, f, \alpha) - L(P, f, \alpha) &= \sum_{i=1}^{n} M_i \Delta\alpha_i - \sum_{i=1}^{n} m_i \Delta\alpha_i \\
&= \sum_{i=1}^{n} (M_i - m_i)\Delta\alpha_i \\
&= \frac{\alpha(b) - \alpha(a)}{n} \sum_{i=1}^{n} (f(x_i) - f(x_{i-1})) \\
&= \frac{\alpha(b) - \alpha(a)}{n} (f(b) - f(a))
\end{aligned}$$

Thus, given $\varepsilon > 0$, there exists an integer n such that $(\alpha(b) - \alpha(a))(f(b) - f(a)) < n\varepsilon$. And, we get a partition P by fixing $\Delta\alpha_i = \frac{\alpha(b)-\alpha(a)}{n}$ such that $U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$.

In other words, $U(P, f, \alpha) - L(P, f, \alpha)$ depends on n and can be reduced to a value less than ε by increasing the value of n. Therefore, $f \in \mathscr{R}(\alpha)$ over $[a, b]$. □

Theorem 4.2.7. *If f bounded on $[a, b]$, f has only finitely many points of discontinuities on $[a, b]$ and α is continuous at every point at which f is discontinuous. Then $f \in \mathscr{R}(\alpha)$.*

Proof. Suppose f is bounded and has only finitely many discontinuities on $[a, b]$. Since f is bounded, we have $-M < f(x) < M$ for some real number M. Also, suppose that α is continuous at those points where f is discontinuous on $[a, b]$.

Let E be the set of all discontinuitites of f on $[a, b]$. Clearly, $|E|$ is finite.

Let $\varepsilon > 0$. Since, α is continuous at each point e_j of E, there exists open intervals (u_j, v_j) such that α is continuous on those intervals, e_j belongs to the interior of these intervals and the sum of variation of α in those intervals is less than ε.

Removing these open intervals from $[a, b]$, we get a compact subset K in which f is continuous. Since K is compact and f is continuous on K, we have f is uniformly continuous on K.

Given $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(t)| < \varepsilon$ whenever $|x - t| < \delta$. Define a partition P such that $u_j,\ v_j \in P$, P doesn't have any point in the interior of any (u_j, v_j). And if $x_{i-1} \neq u_j$, then x_i is so choosen that $x_i - x_{i-1} < \delta$, dividing K into subintervals of length less than δ. Now, we have

$$\begin{aligned}
U(P, f, \alpha) - L(P, f, \alpha) &= \sum_{i=1}^{n} (M_i - m_i)\Delta\alpha_i \\
&\leq \varepsilon \sum_{x_{i-1} \neq u_j} \Delta\alpha_i + 2M \sum_{x_{i-1} = u_j} \Delta\alpha_i
\end{aligned}$$

$$= \varepsilon \sum_{x_{i-1} \neq u_j} (\alpha(x_i) - \alpha(x_{i-1})) + 2M\varepsilon$$
$$\leq \varepsilon(\alpha(b) - \alpha(a)) + 2M\varepsilon$$

Clearly, $U(P, f, \alpha) - L(P, f, \alpha)$ is a function of ε which can reduced below any real number greater than zero by reducing the value of ε. Therefore, $f \in \mathscr{R}(\alpha)$ over $[a, b]$. □

Theorem 4.2.8. *Suppose $f \in \mathscr{R}(\alpha)$, $m \leq f \leq M$ on $[a, b]$, ϕ is continuous on $[m, M]$ and $h(x) = \phi(f(x))$. Then $h \in \mathscr{R}(\alpha)$ on $[a, b]$.*

Proof. Let f be a function on $[a, b]$ such that $m \leq f(x) \leq M$ and $f \in \mathscr{R}(\alpha)$ over $[a, b]$. Let ϕ be continuous function on $[m, M]$. Then ϕ is uniformly continuous on $[m, M]$, since $[m, M]$ is compact. Thus, given $\varepsilon > 0$ there exists $\delta > 0$ such that $|\phi(x) - \phi(t)| < \varepsilon$ whenever $|x - t| < \delta$. Without loss of generality, we may assume that $\delta < \varepsilon$. Otherwise choose a value less than ε as δ.

Since $f \in \mathscr{R}(\alpha)$ over $[a, b]$, there exists a partition P of $[a, b]$ such that

$$U(P, f, \alpha) - L(P, f, \alpha) < \delta^2$$

Consider the ith subinterval $[x_{i-1}, x_i]$ of the partition P. Let M_i, m_i be the maximum and minimum values of the function f in ith subinterval of P. Now, we divided the collection of subintervals into two sets depending on the value of $M_i - m_i$ in comparison with δ. Define A as the set of all subintervals of P such that $M_i - m_i < \delta$. And B as the set of all subintervals of P such that $M_i - m_i \geq \delta$. We know that,

$$\delta \sum_B \Delta\alpha_i \leq \sum_B (M_i - m_i)\Delta\alpha_i < U(P, f, \alpha) - L(P, f, \alpha) < \delta^2$$

Thus, $\sum_B \Delta\alpha_i < \delta$.

Define M_i^*, m_i^* as the maximum and minimum of the function $\phi \circ f$ each subinterval $[x_{i-1}, x_i]$ of the partition P. Now, we have

$$\begin{aligned} U(P, \phi \circ f, \alpha) - L(P, \phi \circ f, \alpha) &= \sum_{i=1}^{n} (M_i^* - m_i^*)\Delta\alpha_i \\ &= \sum_A (M_i^* - m_i^*)\Delta\alpha_i + \sum_B (M_i^* - m_i^*)\Delta\alpha_i \end{aligned}$$

Since ϕ is uniformly continuous,

$$\leq \varepsilon \sum_A \Delta\alpha_i + \sum_B (M_i^* - m_i^*)\Delta\alpha_i$$

Since, continuous functions defined on compact subset attains extrema. We have, $K = \sup\{|\phi(x)| : x \in [m, M]\}$. And $M_i^* - m_i^* < 2K$.

$$\begin{aligned} &\leq \varepsilon(\alpha(b) - \alpha(a)) + 2K \sum_B \Delta\alpha_i \\ &\leq \varepsilon(\alpha(b) - \alpha(a)) + 2K\delta \\ &\leq \varepsilon(\alpha(b) - \alpha(a)) + 2K\varepsilon \end{aligned}$$

Since $U(P, \phi \circ f, \alpha) - L(P, \phi \circ f, \alpha)$ is function of ε, it can be reduced to any sufficiently small real number of our choice. Therefore, $\phi \circ f \in \mathscr{R}(\alpha)$ over $[a, b]$. $\square$

4.2.1 Properties of the Riemann-Stieltjes Integral

Theorem 4.2.9. *Let f, f_1, f_2 be bounded real functions on $[a, b]$. Let $\alpha, \alpha_1, \alpha_2$ be increasing functions on $[a, b]$.*

1. *If $f_1, f_2, f \in \mathscr{R}(\alpha)$ on $[a, b]$, then $f_1 + f_2 \in \mathscr{R}(\alpha)$ on $[a, b]$. And,*

$$\int_a^b (f_1 + f_2)\ d\alpha = \int_a^b f_1\ d\alpha + \int_a^b f_2\ d\alpha$$

   If $c \in \mathbb{R}$, then $cf \in \mathscr{R}(\alpha)$ on $[a, b]$. And,

$$\int_a^b cf\ d\alpha = c \int_a^b f\ d\alpha$$

2. *If $f_1(x) \leq f_2(x)$ on $[a, b]$, then*

$$\int_a^b f_1\ d\alpha \leq \int_a^b f_2\ d\alpha$$

3. *If $c \in (a, b)$, then $f \in \mathscr{R}(\alpha)$ on $[a, c]$ and $[c, b]$, then*

$$\int_a^c f\ d\alpha + \int_c^b f\ d\alpha = \int_a^b f\ d\alpha$$

4. *If $|f(x)| \leq M$ on $[a, b]$, then*

$$\left| \int_a^b f\ d\alpha \right| \leq M[\alpha(b) - \alpha(a)]$$

5. *If $f \in \mathscr{R}(\alpha_1)$ and $f \in \mathscr{R}(\alpha_2)$ on $[a, b]$, then $f \in \mathscr{R}(\alpha_1 + \alpha_2)$. And,*

$$\int_a^b f\ d(\alpha_1 + \alpha_2) = \int_a^b f\ d\alpha_1 + \int_a^b f\ d\alpha_2$$

   If $f \in \mathscr{R}(\alpha)$ on $[a, b]$, and $c \in \mathbb{R}$, then $f \in \mathscr{R}(c\alpha)$ on $[a, b]$. And,

$$\int_a^b f\ d(c\alpha) = c \int_a^b f\ d\alpha$$

Proof. 1. Let $\varepsilon > 0$. Let $f_1, f_2 \in \mathscr{R}(\alpha)$ over $[a, b]$.
Let P_1 be a partition of $[a, b]$ such that $U(P_1, f_1, \alpha) - L(P_1, f_1, \alpha) < \frac{\varepsilon}{2}$.
Let P_2 be a partition of $[a, b]$ such that $U(P_2, f_2, \alpha) - L(P_2, f_2, \alpha) < \frac{\varepsilon}{2}$.
Let $P = P_1 \cup P_2$ be the refinedment of both the partitions. Then the above inequalities are true of the partition P as well.

We have,

$$\begin{aligned} L(P, f_1, \alpha) + L(P, f_2, \alpha) &\le L(P, f_1 + f_2, \alpha) \\ &\le U(P, f_1 + f_2, \alpha) \\ &\le U(P, f_1, \alpha) + U(P, f_2, \alpha) \end{aligned}$$

Thus,

$$\begin{aligned} &U(P, f_1 + f_2, \alpha) - L(P, f_1 + f_2, \alpha) \\ &\qquad \le U(P, f_1, \alpha) - L(P, f_1, \alpha) + U(P, f_2, \alpha) - L(P, f_2, \alpha) \\ &\qquad \le \varepsilon \end{aligned}$$

Therefore, $f_1 + f_2 \in \mathscr{R}(\alpha)$ over $[a, b]$.

---

2. Let c be any real number greater than zero. We have,

$$M_i' = \max\{cf(x) : x \in [x_{i-1}, x_i]\} = c \max\{f(x) : x \in [x_{i-1}, x_i]\} = cM_i$$

Similarly, $m_i' = cm_i$. Thus,

$$\sum_{i=1}^{n} M_i' \Delta\alpha_i = c \sum_{i=1}^{n} M_i \Delta\alpha_i \text{ and } \sum_{i=1}^{n} m_i' \Delta\alpha_i = c \sum_{i=1}^{n} m_i \Delta\alpha_i$$

Therefore, given $\varepsilon > 0$ we have,

$$\begin{aligned} L(P, cf, \alpha) &= cL(P, f, \alpha) \\ U(P, cf, \alpha) &= cU(P, f, \alpha) \end{aligned}$$

Since, $f \in \mathscr{R}(\alpha)$, there exists partition P such that $U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$. Thus,

$$U(P, cf, \alpha) - L(P, cf, \alpha) = cU(P, f, \alpha) - cL(P, f, \alpha) < c\varepsilon$$

Therefore, $cf \in \mathscr{R}(\alpha)$ over $[a, b]$.

$$\begin{aligned} \int_a^b cf \; d\alpha &\le U(P, cf, \alpha) \\ &\le cU(P, f, \alpha) \\ &\le c \int_a^b f \; d\alpha + c\varepsilon \\ &\le c \int_a^b f \; d\alpha \end{aligned}$$

Similarly, considering $-f$ and $-cf$ we get

$$\int_a^b cf \; d\alpha \ge c \int_a^b f \; d\alpha$$

Therefore, without loss of generality for any real number c,

$$\int_a^b cf \; d\alpha = c \int_a^b f \; d\alpha$$

---

3. Let $\varepsilon > 0$. Suppose $f_1, f_2 \in \mathscr{R}(\alpha)$ and $f_1 \leq f_2$. Let P_1, P_2 be the partitions such that $U(P_1, f_1, \alpha) - L(P_1, f_1, \alpha) < \varepsilon$ and $U(P_2, f_2, \alpha) - L(P_2, f_2, \alpha) < \varepsilon$. Consider the common refinement $P = P_1 \cup P_2$. Then, for each subinterval of the partition P, we have

$$\min_{x \in [x_{i-1}, x_i]} f_1(x) \leq \min_{x \in [x_{i-1}, x_i]} f_2(x) \text{ and } \max_{x \in [x_{i-1}, x_i]} f_1(x) \leq \max_{x \in [x_{i-1}, x_i]} f_2(x)$$

Thus, $L(P, f_1, \alpha) \leq L(P, f_2, \alpha)$ and $U(P, f_1, \alpha) \leq U(P, f_2, \alpha)$.

$$\begin{aligned} \int_a^b f_1 \, d\alpha &\leq U(P, f_1, \alpha) \\ &\leq U(P, f_2, \alpha) \\ &\leq \int_a^b f_2 \, d\alpha + \varepsilon \end{aligned}$$

Therefore,

$$\int_a^b f_1 \, d\alpha \leq \int_a^b f_2 \, d\alpha$$

---

4. Let $f \in \mathscr{R}(\alpha)$ on $[a, b]$. Let $\varepsilon > 0$. Then, there exists a parition P of $[a, b]$ such that $U(P, f, \alpha) - L(P, f, \alpha) < \varepsilon$. Let $c \in (a, b)$. Then $P^* = P \cup \{c\} = P_1 \cup P_2$ is refinement of P such that P_1, P_2 are partition of $[a, c]$ and $[c, b]$ respectively. Clearly, $U(P_1, f, \alpha) - L(P_1, f, \alpha) < \varepsilon$ and $U(P_2, f, \alpha) - L(P_2, f, \alpha) < \varepsilon$. Therefore, $f \in \mathscr{R}(\alpha)$ on both $[a, c]$ and $[c, b]$.

For any two partitions P_1, P_2 of $[a, c]$ and $[c, b]$, there exists partition $P^* = P_1 \cup P_2$ of $[a, b]$. Thus,

$$\begin{aligned} \int_a^b f \, d\alpha &\leq U(P^*, f, \alpha) \\ &\leq U(P_1, f, \alpha) + U(P_2, f, \alpha) \\ &\leq \int_a^c f \, d\alpha + \int_c^b f \, d\alpha + 2\varepsilon \end{aligned}$$

Since ε is arbitrary, we have

$$\int_a^b f \, d\alpha \leq \int_a^c f \, d\alpha + \int_c^b f \, d\alpha$$

And considering $-f$ instead of f, we get

$$\int_a^b -f \, d\alpha \leq \int_a^c -f \, d\alpha + \int_c^b -f \, d\alpha$$

Thus,

$$\int_a^b f \, d\alpha \geq \int_a^c f \, d\alpha + \int_c^b f \, d\alpha$$

Therefore,

$$\int_a^b f\ d\alpha = \int_a^c f\ d\alpha + \int_c^b f\ d\alpha$$

---

5. Let function $f \in \mathscr{R}(\alpha)$ on $[a, b]$ and $|f| \leq M$. Let P be any partition of $[a, b]$. We have,

$$L(P, f, \alpha) \leq \int_a^b f\ d\alpha \leq U(P, f, \alpha)$$

And,

$$L(P, f, \alpha) = \sum_{i=1}^{n} m_i \Delta\alpha_i \geq -M\Delta\alpha_i = -M\sum_{i=1}^{n} \Delta\alpha_i = -M[\alpha(b) - \alpha(a)]$$

Similarly,

$$U(P, f, \alpha) = \sum_{i=1}^{n} M_i \Delta\alpha_i \leq M\Delta\alpha_i = M\sum_{i=1}^{n} \Delta\alpha_i = M[\alpha(b) - \alpha(a)]$$

Therefore,

$$\left|\int_a^b f\ d\alpha\right| \leq M[\alpha(b) - \alpha(a)]$$

---

6. Let α_1, α_2 be monotonic functions on $[a, b]$. Then $|\alpha_1 + \alpha_2| \leq |\alpha_1| + |\alpha_2|$. Let $f \in \mathscr{R}(\alpha_1)$ on $[a, b]$ and $f \in \mathscr{R}(\alpha_2)$ on $[a, b]$. Let $\varepsilon > 0$. Then there exists partitions P_1, P_2 of $[a, b]$ such that

$$U(P_1, f, \alpha_1) - L(P_1, f, \alpha_1) < \varepsilon$$

$$U(P_2, f, \alpha_2) - L(P_2, f, \alpha_2) < \varepsilon$$

Consider the common refinement $P = P_1 \cup P_2$. Then, the inequalities are true for P as well. And for each subinterval $[x_{i-1}, x_i]$ of P, we have $\Delta(\alpha_1 + \alpha_2)_i \leq \Delta(\alpha_1)_i + \Delta(\alpha_2)_i$. Thus,

$$\begin{aligned} U(P, f, \alpha_1 + \alpha_2) - L(P, f, \alpha_1 + \alpha_2) &= \sum_{i=1}^{n} (M_i - m_i)\Delta(\alpha_1 + \alpha_2)_i \\ &\leq \sum_{i=1}^{n} (M_i - m_i)\Delta\alpha_{1,i} + \sum_{i=1}^{n} (M_i - m_i)\Delta\alpha_{2,i} \\ &\leq 2\varepsilon \end{aligned}$$

Therefore, $f \in \mathscr{R}(\alpha_1 + \alpha_2)$ on $[a, b]$.

$$\begin{aligned} \int_a^b f\ d(\alpha_1 + \alpha_2) &\leq U(P, f, \alpha_1 + \alpha_2) \\ &\leq U(P, f, \alpha_1) + U(P, f, \alpha_2) \\ &\leq \int_a^b f\ d\alpha_1 + \int_a^b f\ d\alpha_2 + 2\varepsilon \end{aligned}$$

Thus,

$$\int_a^b f\ d(\alpha_1+\alpha_2) \leq \int_a^b f\ d\alpha_1 + \int_a^b f\ d\alpha_2$$

Considering $-f$ instead of f,we get

$$\int_a^b f\ d(\alpha_1+\alpha_2) \geq \int_a^b f\ d\alpha_1 + \int_a^b f\ d\alpha_2$$

Therefore,

$$\int_a^b f\ d(\alpha_1+\alpha_2) = \int_a^b f\ d\alpha_1 + \int_a^b f\ d\alpha_2$$

Let $c>0$ and $f \in \mathscr{R}(\alpha)$ on $[a,b]$. Let $\varepsilon > 0$. Then there exists a partition P of $[a,b]$ such that $U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon$. Clearly, $U(P,f,c\alpha) - L(P,f,c\alpha) < c\epsilon$. Therefore, $f \in \mathscr{R}(c\alpha)$ on $[a,b]$.

$$\begin{aligned}\int fd(c\alpha) &\leq U(P,f,c\alpha)\\ &\leq cU(P,f,\alpha)\\ &\leq c\int fd\alpha + c\varepsilon\end{aligned}$$

Thus,

$$\int fd(c\alpha) \leq c\int fd\alpha$$

Taking $-f$ instead of f, we get

$$\int fd(c\alpha) \geq c\int fd\alpha$$

Therefore,

$$\int fd(c\alpha) = c\int fd\alpha$$

□

Theorem 4.2.10. *If $f,g \in \mathscr{R}(\alpha)$ on $[a,b]$, then*

1. $fg \in \mathscr{R}(\alpha)$ *on* $[a,b]$
2. $|f| \in \mathscr{R}(\alpha)$ *on* $[a,b]$*. And,*

$$\left|\int_a^b f\ d\alpha\right| \leq \int_a^b |f|\ d\alpha$$

Proof. 1. Let $f,g \in \mathscr{R}(\alpha)$ on $[a,b]$. By linearity of the integral we have $f+g \in \mathscr{R}(\alpha)$ on $[a,b]$. Let $\phi(t) = t^2$. Then ϕ is continuous. Thus, $\phi \circ f$ is continuous on $[a,b]$. Therefore, $\phi \circ f \in \mathscr{R}(\alpha)$ on $[a,b]$. Also we have, $\phi\circ(f+g) = (f+g)^2$ and $\phi\circ(f-g) = (f-g)^2$. Thus, $(f+g)^2, (f-g)^2 \in \mathscr{R}(\alpha)$ on $[a,b]$. Now, $4fg = (f+g)^2 - (f-g)^2$. Therefore, $4fg \in \mathscr{R}(\alpha)$ on $[a,b]$, from linearity of the integral. Take $c = \frac{1}{4}$, we get $fg \in \mathscr{R}(\alpha)$ on $[a,b]$ from linearlity of the integral.

---

2. Let $f \in \mathscr{R}(\alpha)$ on $[a,b]$. Let $\phi(t) = |t|$. Then ϕ is continuous. Therefore, $\phi \circ f = |f| \in \mathscr{R}(\alpha)$ on $[a,b]$.

   Let $c = \pm 1$ such that
$$c \int f d\alpha \geq 0$$
   Then,
$$\left| \int f d\alpha \right| = c \int f d\alpha = \int c f d\alpha \leq \int |f| d\alpha$$
   since $cf \leq |f|$.

□

Definitions 4.2.4 (step)**.** The unit step function, $I : \mathbb{R} \to [0,1]$ is defined by
$$I(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$$

Theorem 4.2.11. *Let f be bounded on $[a,b]$ and continuous at $s \in (a,b)$. Let $\alpha = I(t-s)$. Then,*
$$\int_a^b f d\alpha = f(s)$$

Proof. Let $P = \{a, s, x_2, b\}$ be a partition of $[a,b]$. Since f is bounded,
$$U(P,f,\alpha) = M_1 \Delta\alpha_1 + M_2 \Delta\alpha_2 + M_3 \Delta\alpha_3 = M_2$$
since $\Delta\alpha_1 = 0$, we have $\Delta\alpha_2 = \alpha(x_2) - \alpha(s) = I(x_2 - s) - I(0) = 1 - 0$ and $\Delta\alpha_3 = 0$. Similarly, $L(P,f,\alpha) = m_1\Delta\alpha_1 + m_2\Delta\alpha_2 + m_3\Delta\alpha_3 = m_2$. We have,
$$m_2 = L(P,f,\alpha) \leq \underline{\int_a^b} f d\alpha \leq \overline{\int_a^b} f d\alpha \leq U(P,f,\alpha) = M_2$$
We know that, M_2 is the maximum value of f in the subinteval $[s, x_2]$. The lower sum and upper sum remains the same as we reduce the length of the second subinterval. We also know that, f is continuous at s. As $x_2 \to s$, the maximum value of f tends to the value of f at s, say $f(s)$. Similarly, $m_2 \to f(s)$ as $x_2 \to s$. Thus,
$$f(s) \leq \underline{\int_a^b} f d\alpha \leq \overline{\int_a^b} f d\alpha \leq f(s)$$
Therefore, $f \in \mathscr{R}(\alpha)$ on $[a,b]$ and
$$\int_a^b f d\alpha = f(s)$$

□

Theorem 4.2.12. *Suppose $\sum_{i=1}^{\infty} c_n$ converges and $c_n \geq 0$. Let f be continuous on $[a,b]$. Let sequence $\{s_n\}$ be a strictly increasing sequence in (a,b). Let $\alpha = \sum_{i=1}^{\infty} c_n I(t - s_n)$. Then,*
$$\int_a^b f \; d\alpha = \sum_{i=1}^{\infty} c_n f(s_n)$$

Proof. Let $\sum_{n=1}^{\infty} c_n$ be a convergent series. Given $\varepsilon > 0$, there exists a natural number N such that

$$\sum_{n=N+1}^{\infty} c_n < \varepsilon$$

Let $s_1, s_2, s_3, \ldots$ be distinct points in (a, b). Without loss of generality, sequence $\{s_n\}$ is a striclty increasing sequence in $[a, b]$. Let $\alpha = \sum_{n=1}^{\infty} c_n I(t - s_n)$. Then, $\alpha(a) = 0$ and $\alpha(b) = \sum_{n=1}^{\infty} c_n$.

$$0 \leq \alpha(x) \leq \sum_{n=1}^{\infty} c_n, \quad \forall x \in [a, b]$$

By comparison test, α is convergent in $[a, b]$. We have,

$$\int_a^b f(t)\ d(c_1 I(t - s_1)) = c_1 \int_a^b f(t)\ d(I(t - s_1)) = c_1 f(s_1)$$

Let $\alpha = \alpha_1 + \alpha_2$ where $\alpha_1 = \sum_{n=1}^{N} c_n I(t - s_n)$ and $\alpha_2 = \sum_{n=N+1}^{\infty} c_n I(t - s_n)$. And from mathematical induction, we have

$$\begin{aligned}
\int_a^b f(t) d(\alpha_1) &= \int_a^b f(t)\ d\left(\sum_{n=1}^{N} c_n I(t - s_n)\right) \\
&= \sum_{n=1}^{N} \int_a^b f(t)\ d(c_n I(t - s_n)) \\
&= \sum_{n=1}^{N} c_n \int_a^b f dI(t - s_n) \\
&= \sum_{n=1}^{N} c_n f(s_n)
\end{aligned}$$

We have,

$$\left| \int_a^b f(t) d\alpha_2 \right| \leq M(\alpha_2(b) - \alpha_2(a)) = M \sum_{n=N+1}^{\infty} < M\varepsilon$$

where $M = \sup |f(x)|$. Thus,

$$\begin{aligned}
\left| \int_a^b f d\alpha_2 \right| &= \left| \int_a^b f d(\alpha_1 + \alpha_2) - \int_a^b f d\alpha_1 \right| \\
&= \left| \int_a^b f d\alpha - \sum_{n=1}^{N} c_n f(s_n) \right| \\
&< M\varepsilon
\end{aligned}$$

The inequality is true as $N \to \infty$. In other words, the sequence of partial sums converges to the value of integral of f. Therefore,

$$\sum_{n=1}^{\infty} c_n f(s_n) = \int_a^b f d\alpha$$

□

Theorem 4.2.13. *Let α be increasing function on $[a, b]$ and $\alpha' \in \mathscr{R}$. Let f be bounded real-valued function on $[a, b]$. Then $f \in \mathscr{R}(\alpha)$ if and only if $f\alpha' \in \mathscr{R}$. And*

$$\int_a^b f d\alpha = \int_a^b f(x)\alpha'(x)dx$$

Proof. Since $\alpha' \in \mathscr{R}$ on $[a, b]$, for any $\varepsilon > 0$, there exists a partition P of $[a, b]$ such that $U(P, \alpha') - L(P, \alpha') < \varepsilon$. We have, α is continuous on $[a, b]$, as it is differentiable on $[a, b]$. Then by intermediate value theorem, we have

$$\Delta\alpha_i = \alpha(x_i) - \alpha(x_{i-1}) = \alpha'(t_i)(x_i - x_{i-1})$$

where $t_i \in [x_{i-1}, x_i]$. Let $s_i \in [x_{i-1}, x_i]$. Then,

$$\sum_{i=1}^{n} |\alpha'(s_i) - \alpha'(t_i)|\Delta x_i \leq U(P, \alpha') - L(P, \alpha') < \varepsilon$$

Let $M = \sup |f|$. Let $u_i \in [x_{i-1}, x_i]$. [1] Consider,

$$\begin{aligned}\left|\sum_{i=1}^{n} f(u_i)\Delta\alpha_i - \sum_{i=1}^{n} f(s_i)\alpha'(s_i)\Delta x_i\right| &\leq M\left|\sum_{i=1}^{n} \Delta\alpha_i - \alpha'(s_i)\Delta x_i\right| \\ &\leq M\left|\sum_{i=1}^{n} \alpha'(t_i)\Delta x_i - \alpha'(s_i)\Delta x_i\right| \\ &\leq M\sum_{i=1}^{n} |\alpha'(t_i) - \alpha'(s_i)|\Delta x_i \\ &< M\varepsilon\end{aligned}$$

Clearly, the inequality is true independent of the choice of $u_i, s_i \in [x_{i-1}, x_i]$. Thus, selecting u_i such that $f(u_i) = M_i$, we get

$$\left|U(P, f, \alpha) - \sum_{i=1}^{n} \alpha'(s_i)\Delta x_i\right| < M\varepsilon$$

Therefore,

$$\sum_{i=1}^{n} f(s_i)\alpha'(s_i)\Delta x_i < U(P, f, \alpha) + M\varepsilon$$

Selecting s_i such that $f\alpha'$ attains maximum at s_i, we get

$$U(P, f\alpha') < U(P, f, \alpha) + M\varepsilon$$

Now, first we select s_i such that $f\alpha(s_i)$ is maximum in ith subinterval of P. Then,

$$\left|\sum_{i=1}^{n} f(u_i)\Delta\alpha_i - U(P, f\alpha')\right| < M\varepsilon$$

[1] In my opinion, using u_i instead of s_i in the first sum makes it simpler.

Therefore,

$$U(P, f\alpha') < \sum_{i=1}^{n} f(u_i)\Delta\alpha_i + M\varepsilon$$

And, now selecting u_i such that $f(u_i) = M_i$.

$$U(P, f\alpha') < U(P, f, \alpha) + M\varepsilon$$

Therefore, we have $U(P, f\alpha') = U(P, f, \alpha)$. Clearly, it is true for any refinement of P. Therefore,

$$\overline{\int_a^b} f\alpha' dx = \overline{\int_a^b} f d\alpha$$

Similary, by selecting u_i and s_i such that $f(u_i)$, $f\alpha'(s_i)$ is minimum in each subinterval of P, we get

$$\underline{\int_a^b} f\alpha' dx = \underline{\int_a^b} f d\alpha$$

Cleary, $f \in \mathscr{R}(\alpha) \iff f\alpha' \in \mathscr{R}$. And,

$$\int_a^b f\alpha' dx = \int_a^b f d\alpha$$

$\square$

Theorem 4.2.14 (change of variable)**.** *Let $\varphi : [A, B] \to [a, b]$ be a strictly increasing, continuous function onto $[a, b]$. Let $\alpha : [a, b] \to \mathbb{R}$ be an increasing function and $f \in \mathscr{R}(\alpha)$ on $[a, b]$. Define $\beta = \alpha \circ \varphi$ and $g = f \circ \varphi$. Then $g \in \mathscr{R}(\beta)$ on $[A, B]$ and*

$$\int_A^B g d\beta = \int_a^b f d\alpha$$

Proof. Let $P = \{x_0, x_1, \dots, x_n\}$ be any partition of $[a, b]$. Then there exists a partition $Q = \{y_0, y_1, \dots, y_n\}$ of $[A, B]$ such that $x_j = \varphi(y_j)$ since φ is a continuous, bijection. †[2] Similarly, for any partition $Q = \{y_0, y_1, \dots, y_n\}$ of $[A, B]$, there exists a partition $P = \{x_0, x_1, \dots, x_n\}$ of $[a, b]$ where $x_j = \varphi(y_j)$. Clearly, minimum/maximum of g in ith subinterval of Q is same as the minimum/maximum of f in ith subinterval of P. And

$$\Delta\beta_i = \beta(y_i) - \beta(y_{i-1}) = \alpha(\varphi(y_i)) - \alpha(\varphi(y_{i-1})) = \alpha(x_i) - \alpha(x_{i-1}) = \Delta\alpha_i$$

$$\begin{aligned}
\min\{g(y) : y \in [y_{i-1}, y_i]\} &= \min\{f(\varphi(y)) : y \in [y_{i-1}, y_i]\} \\
&= \min\{f(x) : x \in [x_{i-1}, x_i]\} \\
\implies L(Q, g, \beta) &= L(P, f, \alpha) \\
\max\{g(y) : y \in [y_{i-1}, y_i]\} &= \max\{f(\varphi(y)) : y \in [y_{i-1}, y_i]\} \\
&= \max\{f(x) : x \in [x_{i-1}, x_i]\} \\
\implies U(Q, g, \beta) &= U(P, f, \alpha)
\end{aligned}$$

[2] Strictly increasing functions are always injective.

Since $f \in \mathscr{R}(\alpha)$ on $[a,b]$, there exists a partition P of $[a,b]$ such that $U(P,f,\alpha) - L(P,f,\alpha) < \varepsilon$. Therefore, there exists a partition Q of $[A,B]$ such that $U(Q,g,\beta) - L(Q,g,\beta) < \varepsilon$. Thus, $g \in \mathscr{R}(\beta)$. And

$$\int_A^B g d\beta = \int_a^b f d\alpha$$

□

4.2.2 Integration and Differentiation

Theorem 4.2.15. *Let $f \in \mathscr{R}$ on $[a,b]$. Let $a \leq x \leq b$. Define*

$$F(x) = \int_a^x f(t)dt$$

Then F is continuous on $[a,b]$. Furthermore, if f is continuous at x_0, then F is differentiable at x_0. And $F'(x_0) = f(x_0)$.

Proof. Let $a \leq x < y \leq b$. Let $M = \sup|f|$. We have,

$$\begin{aligned} |F(y) - F(x)| &= \left| \int_a^y f(t)dt - \int_a^x f(t)dt \right| \\ &= \left| \int_a^x f(t)dt + \int_x^y f(t)dt - \int_a^x f(t)dt \right| \\ &= \left| \int_x^y f(t)dt \right| \\ &\leq M(y-x) \end{aligned}$$

Thus, given $\varepsilon > 0$, there exists $\delta > 0$ such that $|F(y) - F(x)| < \varepsilon$ whenever $|y - x| < \delta \leq \frac{\varepsilon}{M}$. Therefore, F is continuous on $[a,b]$.

Let f be continuous at $x_0 \in [a,b]$. Given $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(t) - f(x_0)| < \varepsilon$ whenever $|t - x_0| < \delta$ and $t \in [a,b]$.

Let $x - \delta < s \leq x_0 \leq t < x + \delta$. Clearly, $|f(u) - f(x_0)| < \varepsilon$ whenever $u \in [s,t]$. Consider,

$$\begin{aligned} \left| \frac{F(t) - F(s)}{t - s} - f(x_0) \right| &= \left| \frac{1}{t-s} \left(\int_a^t f(u)du - \int_a^s f(u)du \right) - \frac{1}{t-s} f(x_0)(t-s) \right| \\ &= \left| \frac{1}{t-s} \int_s^t f(u)du - \frac{1}{t-s} \int_s^t f(x_0)du \right| \\ &= \left| \frac{1}{t-s} \left(\int_s^t (f(u) - f(x_0))du \right) \right| \\ &< \frac{\varepsilon}{t-s} \int_s^t du \quad \text{since } |f(u) - f(x_0)| < \varepsilon \\ &< \varepsilon \end{aligned}$$

Clearly, as $\varepsilon \to 0$, $\delta \to 0$. Then, $s,t \to x_0$. Therefore, F differentiable at x_0 and

$$F'(x_0) = \lim_{s,t \to x_0} \frac{F(t) - F(s)}{t - s} = f(x_0)$$

□

Theorem 4.2.16 (fundamental theorem of calculus)**.** *Let $f \in \mathscr{R}$ on $[a, b]$. Let F be a differentiable function on $[a, b]$ such that $F' = f$. Then,*

$$\int_a^b f(x)dx = F(b) - F(a)$$

Proof. Let $\varepsilon > 0$. Let $f \in \mathscr{R}$ on $[a, b]$. Then, there exists a partition $P = \{x_0, x_1, \dots, x_n\}$ of $[a, b]$ such that $U(P, f) - L(P, f) < \varepsilon$.

Let F be differentiable function such that $F' = f$. Then, F is continuous. By intermediate value theorem, there exists $t_i \in [x_{i-1}, x_i]$ such that

$$F(x_i) - F(x_{i-1}) = F'(t_i)(x_i - x_{i-1}) = f(t_i)\Delta x_i$$

Clearly,

$$F(b) - F(a) = \sum_{i=1}^{n} (F(x_i) - F(x_{i-1})) = \sum_{i=1}^{n} f(t_i)\Delta x_i$$

Since $t_i \in [x_{i-1}, x_i]$, we have $m_i \le f(t_i) \le M_i$. Thus,

$$\left| \int_a^b f(x)dx - (F(b) - F(a)) \right| = \left| \int_a^b f(x)dx - \sum_{i=1}^{n} f(t_i)\Delta x_i \right| < \varepsilon$$

Therefore,

$$\int_a^b f(x)dx = F(b) - F(a)$$

□

Theorem 4.2.17 (integration by parts)**.** *Let F, G be differentiable $[a, b]$. Let $F' = f \in \mathscr{R}$ and $G' = g \in \mathscr{R}$. Then,*

$$\int_a^b F(x)g(x)dx = F(b)G(b) - F(a)G(a) - \int_a^b f(x)G(x)dx$$

Proof. Let F, G be differentiable functions on $[a, b]$ and $F' = f \in \mathscr{R}$ and $G' = g \in \mathscr{R}$. Let $H = FG$. Then, $H' = FG' + F'G = Fg + fG$.

$$\int_a^b H'(x)dx = \int_a^b F(x)g(x)dx + \int_a^b f(x)G(x)dx$$

By fundamental theorem of calculus, we also have

$$\int_a^b H'(x) = H(b) - H(a) = FG(b) - FG(a) = F(b)G(b) - F(a)G(a)$$

Rearranging the terms, we get

$$\int_a^b F(x)g(x)dx = F(b)G(b) - F(a)G(a) - \int_a^b f(x)G(x)dx$$

□

4.2.3 Integration of Vector-valued Functions

Definitions 4.2.5 (integrable)**.** Let $\bar{f} : [a, b] \to \mathbb{R}^k$ be a vector-valued function. Let α be a monotonic function on $[a, b]$. Let $f_1, f_2, \ldots, f_k$ be the component functions of $\bar{f}$. That is, $\bar{f}(x) = (f_1(x), f_2(x), \ldots, f_k(x))$. Then $\bar{f} \in \mathscr{R}(\alpha)$ on $[a, b]$ if and only if every component function $f_j \in \mathscr{R}(\alpha)$.

In other words, vector-valued function $\bar{f}$ is integrable if and only if every component function of $\bar{f}$ is integrable.

Theorem 4.2.18 (properties)**.** *Suppose $\bar{f}, \bar{g}$ be vector-valued functions from $[a, b]$ into $\mathbb{R}^k$.*

1. *Let $\bar{f}, \bar{g} \in \mathscr{R}(\alpha)$ on $[a, b]$. Then $\bar{f} + \bar{g} \in \mathscr{R}(\alpha)$. And,*

$$\int_a^b \bar{f} + \bar{g} d\alpha = \int_a^b \bar{f} d\alpha + \int_a^b \bar{g} d\alpha$$

   If $\bar{f} \in \mathscr{R}(\alpha)$ on $[a, b]$ and $c \in \mathbb{R}$, then $c\bar{f} \in \mathscr{R}(\alpha)$ on $[a, b]$. And,

$$\int_a^b c\bar{f} d\alpha = c \int_a^b \bar{f} d\alpha$$

2. *Let $c \in (a, b)$. Let $\bar{f} \in \mathscr{R}(\alpha)$ on $[a, b]$. Then $\bar{f} \in \mathscr{R}(\alpha)$ on both $[a, c]$ and $[c, b]$. And,*

$$\int_a^b \bar{f} d\alpha = \int_a^c \bar{f} d\alpha + \int_c^b \bar{f} d\alpha$$

3. *Let α_1, α_2 be monotonic functions on $[a, b]$. Let $\bar{f} \in \mathscr{R}(\alpha_1)$ on $[a, b]$ and $\bar{f} \in \mathscr{R}(\alpha_2)$ on $[a, b]$. Then, $\bar{f} \in \mathscr{R}(\alpha_1 + \alpha_2)$ on $[a, b]$. And,*

$$\int_a^b \bar{f} d(\alpha_1 + \alpha_2) = \int_a^b \bar{f} d\alpha_1 + \int_a^b \bar{f} d\alpha_2$$

Proof. 1. Let $\bar{f}, \bar{g} \in \mathscr{R}(\alpha)$ on $[a, b]$ where $\bar{f} = (f_1, f_2, \ldots, f_k)$ and $\bar{g} = (g_1, g_2, \ldots, g_k)$. By definition of integrability of vector-valued functions, the component functions $f_j, g_j \in \mathscr{R}(\alpha)$ on $[a, b]$ for $1 \le j \le k$.

We also know that, if $f_j, g_j \in \mathscr{R}(\alpha)$ on $[a, b]$, then $f_j + g_j \in \mathscr{R}(\alpha)$ on $[a, b]$. Thus, $f_1 + g_1, f_2 + g_2, \ldots, f_k + g_k \in \mathscr{R}(\alpha)$ on $[a, b]$ for $1 \le j \le k$. Therefore, $\bar{f} + \bar{g} = (f_1 + g_1, f_2 + g_2, \ldots, f_k + g_k) \in \mathscr{R}(\alpha)$ on $[a, b]$.

Let $c \in \mathbb{R}$. Let $\bar{f} \in \mathscr{R}(\alpha)$ on $[a, b]$. Then, $f_1, f_2, \ldots, f_k \in \mathscr{R}(\alpha)$ on $[a, b]$. We know that, $cf_1 \in \mathscr{R}(\alpha)$ on $[a, b]$, since $f_1 \in \mathscr{R}(\alpha)$ on $[a, b]$. Thus, $cf_1, cf_2, \ldots, cf_k \in \mathscr{R}(\alpha)$ on $[a, b]$. Therefore, $c\bar{f} = (cf_1, cf_2, \ldots, cf_k) \in \mathscr{R}(\alpha)$ on $[a, b]$.

---

2. Let $\bar{f} \in \mathscr{R}(\alpha)$ on $[a, b]$. Then, $f_1, f_2, \ldots, f_k \in \mathscr{R}(\alpha)$ on $[a, b]$. Let $c \in (a, b)$. We know that, if $f_j \in \mathscr{R}(\alpha)$ on $[a, b]$, then $f_j \in \mathscr{R}(\alpha)$ on both $[a, c]$ and $[c, b]$. Thus, $f_1, f_2, \ldots, f_k \in \mathscr{R}(\alpha)$ on both $[a, c]$ and $[c, b]$. Therefore, $\bar{f} \in \mathscr{R}(\alpha)$ on both $[a, c]$ and $[c, b]$.

---

3. Let $\bar{f} \in \mathscr{R}(\alpha_1)$ and $\bar{f} \in \mathscr{R}(\alpha_2)$ on $[a, b]$. Then $f_1, f_2, \ldots, f_k \in \mathscr{R}(\alpha_1)$ and $f_1, f_2, \ldots, f_k \in \mathscr{R}(\alpha_2)$ on $[a, b]$. We know that, if $f_j \in \mathscr{R}(\alpha_1)$ and $f_j \in \mathscr{R}(\alpha_2)$ on $[a, b]$, then $f_j \in \mathscr{R}(\alpha_1 + \alpha_2)$ on $[a, b]$.

□

Challenge 3. Let $\bar{f}, \bar{g} \in \mathscr{R}(\alpha)$ on $[a, b]$ where $\bar{f}, \bar{g} : [a, b] \to \mathbb{R}^k$. Define $\bar{f} \cdot \bar{g} : [a, b] \to \mathbb{R}$ by $(\bar{f} \cdot \bar{g})(x) = f_1(x)g_1(x) + f_2(x)g_2(x) + \cdots + f_k(x)g_k(x)$. Then, $\bar{f} \cdot \bar{g} \in \mathscr{R}(\alpha)$ on $[a, b]$. And,

$$\int_a^b \bar{f} \cdot \bar{g} d\alpha = \left(\int_a^b \bar{f} d\alpha\right) \cdot \left(\int_a^b \bar{g} d\alpha\right)$$

Theorem 4.2.19. *Let α be a monotonic function such that $\alpha' \in \mathscr{R}$ on $[a, b]$. Let $\bar{f}$ be a bounded function on $[a, b]$. Then $\bar{f} \in \mathscr{R}(\alpha)$ on $[a, b]$ if and only if $f\alpha' \in \mathscr{R}$ on $[a, b]$. And,*

$$\int_a^b \bar{f} d\alpha = \int_a^b \bar{f}\alpha' dx$$

Proof. Let $\bar{f} = (f_1, f_2, \ldots, f_k)$. Then, $\bar{f}\alpha' = (f_1\alpha', f_2\alpha', \ldots, f_k\alpha')$.
We already know that, $f_j \in \mathscr{R}(\alpha) \iff f_j\alpha' \in \mathscr{R}$.
Therefore,

$$\begin{aligned} \bar{f} \in \mathscr{R}(\alpha) &\iff f_1, f_2, \ldots, f_k \in \mathscr{R}(\alpha) \\ &\iff f_1\alpha', f_2\alpha', \ldots, f_k\alpha' \in \mathscr{R} \\ &\iff \bar{f}\alpha' \in \mathscr{R} \end{aligned}$$

Thus, $\bar{f} \in \mathscr{R}(\alpha) \iff \bar{f}\alpha' \in \mathscr{R}$.

We also know that,

$$\int_a^b f_j d\alpha = \int_a^b f_j\alpha' dx$$

Thus,

$$\begin{aligned} \int_a^b \bar{f} d\alpha &= \left(\int_a^b f_1 d\alpha, \int_a^b f_2 d\alpha, \ldots, \int_a^b f_k d\alpha\right) \\ &= \left(\int_a^b f_1\alpha' dx, \int_a^b f_2\alpha' dx, \ldots, \int_a^b f_k\alpha' dx\right) \\ &= \int_a^b \bar{f}\alpha' dx \end{aligned}$$

□

Theorem 4.2.20. *Let $\bar{f} \in \mathscr{R}$ on $[a,b]$ where $\bar{f} : [a,b] \to \mathbb{R}^k$. Define $\bar{F} : [a,b] \to \mathbb{R}^k$ defined by*

$$\bar{F}(x) = \int_a^x \bar{f}(t)dt$$

Then F is continuous on $[a,b]$. Furthermore, if $\bar{f}$ is continuous at $x_0 \in [a,b]$, then $\bar{F}$ is differentiable at x_0 and $\bar{F}'(x_0) = \bar{f}(x_0)$.

Proof. We know that, $\bar{f} = (f_1, f_2, \ldots, f_k)$. And from the definition of integral, we have

$$\int_a^b \bar{f}(t)dt = \left(\int_a^b f_1(t)dt, \int_a^b f_2(t)dt, \ldots, \int_a^b f_k(t)dt\right)$$

We know that, for $1 \leq j \leq k$, the function $F_j : [a,b] \to \mathbb{R}$ defined by

$$F_j(x) = \int_a^x f_j(t)dt$$

is continuous. And if f_j is continuous at x_0, then F_j is differentiable at x_0 and $F_j'(x_0) = f_j(x_0)$. Clearly, $\bar{F} = (F_1, F_2, \ldots, F_k)$ is continuous, since each component function is continuous. And, $\bar{F}$ is differentiable at x_0 and

$$\bar{F}'(x_0) = (F_1'(x_0), F_2'(x_0), \ldots, F_k'(x_0)) = (f_1(x_0), f_2(x_0), \ldots, f_k(x_0)) = \bar{f}(x_0)$$

□

Theorem 4.2.21 (fundamental theorem of calculus for vector-valued functions)**.** *Let $\bar{f} : [a,b] \to \mathbb{R}^k$. Let $\bar{F} : [a,b] \to \mathbb{R}^k$. If $\bar{f} \in \mathscr{R}$ on $[a,b]$ and $\bar{F}' = \bar{f}$, then*

$$\int_a^b \bar{f}(t)dt = \bar{F}(b) - \bar{F}(a)$$

Proof. By fundamental theorem of calculus, we have

$$\int_a^b f_j(t)dt = F_j(b) - F_j(a)$$

for $1 \leq j \leq k$. Therefore,

$$\begin{aligned}
\int_a^b \bar{f}(t)dt &= \left(\int_a^b f_1(t)dt, \int_a^b f_2(t)dt, \ldots, \int_a^b f_k(t)dt\right) \\
&= (F_1(b) - F_1(a), F_2(b) - F_2(a), \ldots, F_k(b) - F_k(a)) \\
&= (F_1(b), F_2(b), \ldots, F_k(b)) - (F_1(a), F_2(a), \ldots, F_k(a)) \\
&= \bar{F}(b) - \bar{F}(a)
\end{aligned}$$

□

Theorem 4.2.22. *Let $\bar{f} : [a,b] \to \mathbb{R}^k$. If $\bar{f} \in \mathscr{R}(\alpha)$ on $[a,b]$, then $|\bar{f}| \in \mathscr{R}(\alpha)$ on $[a,b]$. And*

$$\left|\int_a^b \bar{f}d\alpha\right| \leq \int_a^b |\bar{f}|d\alpha$$

Proof. Let $\bar{f} : [a,b] \to \mathbb{R}^k$. Then, we have $\bar{f} = (f_1, f_2, \ldots, f_k)$.
Suppose $\bar{f} \in \mathscr{R}(\alpha)$ on $[a,b]$. Then, $f_j \in \mathscr{R}(\alpha)$ on $[a,b]$ for $1 \leq j \leq k$. We have, $|\bar{f}| = \left(f_1^2 + f_2^2 + \cdots + f_k^2\right)^{\frac{1}{2}}$. We know that if $f_j \in \mathscr{R}(\alpha)$, then $f_j^2 \in \mathscr{R}(\alpha)$. Again, $\sum f_j^2 \in \mathscr{R}(\alpha)$.

Consider $g : [a,b] \to \mathbb{R}$ given by $g(x) = \sum_{j=1}^{k} f_j^2(x)$. We have $g \in \mathscr{R}(\alpha)$ on $[a,b]$. Thus g is bounded and there exists m, M such that $m \leq g \leq M$. Clearly, $g \geq 0$.

Consider, the function $\phi : [m, M] \to \mathbb{R}$ given by $\phi(x) = \sqrt{x}$. Clearly, ϕ is well-defined on $[m, M]$ since $0 \leq m$. And, ϕ is continuous on $[m, M]$. Thus, $|\bar{f}| = \phi \circ g = \sqrt{\sum f_j^2} \in \mathscr{R}(\alpha)$ on $[a,b]$.

Let $\bar{y} = \int_a^b \bar{f} d\alpha$. Then, $\bar{y} = (y_1, y_2, \ldots, y_k)$ and $y_j = \int_a^b f_j d\alpha$.

$$\begin{aligned}
|\bar{y}|^2 &= y_1^2 + y_2^2 + \cdots + y_k^2 \\
&= y_1 \int_a^b f_1 d\alpha + y_2 \int_a^b f_2 d\alpha + \cdots + y_k \int_a^b f_k d\alpha \\
&= \int_a^b (y_1 f_1 + y_2 f_2 + \cdots + y_k f_k) d\alpha
\end{aligned}$$

By Schwarz inequality,

$$\sum_{j=1}^{n} y_j f_j \leq \left(\sum_{j=1}^{n} y_j^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^{n} f_j^2\right)^{\frac{1}{2}} = |\bar{y}|\ |\bar{f}|$$

Thus,

$$|\bar{y}|^2 = \int_a^b \left(\sum_{j=1}^{k} y_j f_j\right) d\alpha \leq \int_a^b |\bar{y}|\ |\bar{f}| d\alpha = |\bar{y}| \int_a^b |\bar{f}| d\alpha$$

Therefore,

$$\left|\int_a^b \bar{f} d\alpha\right| = |\bar{y}| \leq \int_a^b |\bar{f}| d\alpha$$

$\square$

Module 3

4.3 Sequence and Series of functions

Definitions 4.3.1. Let sequence $\{f_n\}$ be a sequence of functions defined on E. Suppose sequence $\{f_n(x)\}$ converges forevery $x \in E$. Then, sequence $\{f_n\}$ converges. And **limit function** $f : E \to \mathbb{R}$ is defined by

$$f(x) = \lim_{n\to\infty} f_n(x)$$

Definitions 4.3.2. Let $f_n : E \to \mathbb{R},\ \forall n \in \mathbb{N}$. Suppose $\sum f_n(x)$ converges for every $x \in E$. Then, series $\sum f_n$ converges. And **sum** $f : E \to \mathbb{R}$ is defined by

$$f(x) = \sum_{n=1}^{\infty} f_n(x)$$

4.3.1 Counter Examples - Illustrating Main problem

Main Problem : Can we obtain the sufficient conditions for preserving desirable properties under convergence ?

Interchanging limits

$$\text{Generally,} \quad \lim_{n\to\infty}\lim_{m\to\infty} S_{n,m} \neq \lim_{m\to\infty}\lim_{n\to\infty} S_{n,m}$$

Proof. Consider, $S_{n,m} = \frac{m}{m+n}$.

$$\lim_{m\to\infty}\lim_{n\to\infty} S_{n,m} = \lim_{m\to\infty} 0 = 0$$
$$\lim_{n\to\infty}\lim_{m\to\infty} S_{n,m} = \lim_{n\to\infty} 1 = 1$$

□

Continuity

$$\{f_n\} \to f,\ f_n \text{ continuous} \not\Longrightarrow f \text{ continuous}$$

Proof. Consider $f_n(x) = \frac{x^2}{(1+x^2)^n}$. Clearly, $f_n : \mathbb{R} \to \mathbb{R}$ is continuous for every natural number n.

Case 1 : $x = 0$
Suppose $x = 0$. Then, $f_n(0) = 0$ and therefore $\lim_{n\to\infty} f_n(0) = 0$.

Case 2 : $x \neq 0$
Suppose $x \neq 0$. Then $\frac{1}{1+x^2} < 1,\ \forall x \in \mathbb{R},\ (x \neq 0)$.
Therefore,

$$f(x) = \lim_{n\to\infty} f_n(x) = x^2 \lim_{n\to\infty}\left(\frac{1}{1+x^2}\right)^n = x^2 \frac{1}{1-\frac{1}{1+x^2}} = x^2\frac{1+x^2}{x^2} = 1+x^2$$

Clearly from cases 1 and 2,

$$f : \mathbb{R} \to \mathbb{R},\ f(x) = \begin{cases} 0 & x = 0 \\ 1 + x^2 & x \neq 0 \end{cases} \quad \text{is not continuous at } 0$$

□

integrability

$$\{f_n\} \to f, f_n \in \mathscr{R} \not\Longrightarrow f \in \mathscr{R}$$

Proof. Consider, $f_m(x) = \lim\limits_{n\to\infty} (\cos m!\pi x)^{2n}$.

Suppose $m!x$ is not an integer. Then $f_m(x) = \lim\limits_{n\to\infty} (y^2)^n = 0$ since $-1 < y = \cos m!\pi x < 1$. Suppose $m!x$ is an integer. Then $\cos m!\pi x = \pm 1$ and $f_m(x) = \lim\limits_{n\to\infty} ((\pm 1)^2)^n = 1$.

We know that, if $x \in \mathbb{Q}$, then $x = \pm\frac{p}{q}$ where $p, q \in \mathbb{N}$. Clearly, for any $m > q$, $m!x$ is an integer and $f_m(x) = 1$ for any $m > q$. Therefore, $\lim\limits_{m\to\infty} f_m(x) = 1$ if $x \in \mathbb{Q}$. And $\lim\limits_{m\to\infty} f_m(x) = 0$ if $x \notin \mathbb{Q}$ since $m!x$ is not an integer for any $m \in \mathbb{N}$ and $f_m(x) = 0,\ \forall m \in \mathbb{N}$.

Now, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \lim\limits_{m\to\infty} f_m(x)$ is given by,

$$f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases} \quad \text{which is } \textbf{everywhere discontinuous.}$$

Clearly, f is not Riemann integrable since $\mathbb{Q}, \mathbb{R} - \mathbb{Q}$ are dense in any subinterval. Thus, $m_i = 0$ and $M_i = 1$ in any subinterval of any partition P.

$$\underline{\int_0^1} f(x)\ dx = 0 \neq 1 = \overline{\int_0^1} f(x)\ dx$$

Let $m \in \mathbb{N}$. Consider closed interval $E = [-1, 1]$. Then there exists finitely many rational numbers x on E such that $m!x$ is an integer. Thus, $f_m(x)$ has at most finitely many discontinuities in any bounded interval. $f_m(x) = \cos m!\pi x$ is discontinuous at $S = \left\{\frac{k}{m!} \in E : k \in \mathbb{Z}\right\}$. For example, suppose $m = 3$. Then $S = \left\{0, \pm\frac{1}{6}, \pm\frac{2}{6}, \pm\frac{3}{6}, \pm\frac{4}{6}, \pm\frac{5}{6}, \pm 1\right\}$. Let α be the identity function. Then α is continuous at those finite points where f_m is discontinuous. And f_m are bounded functions, since $|f_m| \leq 1$. Therefore, for any bounded interval E, f_m is Riemann integrable on E for each m. However, f is not Riemann integrable. □

Derivative

$$\{f_n\} \to f \not\Longrightarrow \{f'_n\} \to f'$$

Proof. Consider $f_n(x) = \frac{\sin nx}{\sqrt{n}}$.

$$f(x) = \lim_{n\to\infty} f_n(x) = 0,\ \forall x$$

However, $f_n'(x) = \sqrt{n}\cos nx$. And $f'(x) = 0$.

$$\lim_{n\to\infty} \frac{d}{dx} f_n(x) = \infty \neq f'(x) = \frac{d}{dx} \lim_{n\to\infty} f_n(x), \ \forall x$$

Clearly, convergence doesn't preserve derivatives. □

Integral

$$\{f_n\} \to f \not\Rightarrow \left\{\int f_n\right\} \to \int f$$

Proof. Consider $f_n : [0,1] \to \mathbb{R}$ defined by $f_n(x) = n^2 x(1-x^2)^n$.

We have,

$$\lim_{n\to\infty} \frac{n^\alpha}{(1+p)^n} = 0$$

Let $\frac{1}{1+p} = 1 - x^2$, then $1 + p = \frac{1}{1-x^2}$ and $p = \frac{x^2}{1-x^2} > 0$.
Now, we have

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} n^2 x(1-x^2)^n = x \lim_{n\to\infty} \frac{n^2}{\left(1 + \frac{x^2}{1-x^2}\right)^n} = 0$$

And, we have

$$\int_0^1 f(x)\ dx = 0$$

However,

$$\int_0^1 f_n(x)\ dx = n^2 \int_0^1 x(1-x^2)^n\ dx = n^2 \int_0^1 \frac{u^n}{2}\ du = n^2 \left[\frac{u^{n+1}}{2(n+1)}\right]_0^1 = \frac{n^2}{2n+2}$$

Clearly,

$$\lim_{n\to\infty} \int_0^1 f_n(x)\ dx = \lim_{n\to\infty} \frac{n^2}{2n+2} = \infty \neq 0 = \int_0^1 f(x)\ dx = \int_0^1 \lim_{n\to\infty} f_n(x)\ dx$$

□

4.3.2 Uniform Convergence

Uniform convergence is a partial solution to our main problem.

Definitions 4.3.3. Let sequence $\{f_n\}$ be a sequence of functions on E. Then $\{f_n\}$ converges uniformly to limit function f if for any $\varepsilon > 0$, there exists a natural number N such that for any $n > N$ and $x \in E$, $|f_n(x) - f(x)| \leq \varepsilon$.

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n > N,\ \forall x \in E,\ |f_n(x) - f(x)| \leq \varepsilon$$

The difference between pointwise convergence and uniform convergence is that the choice of natrual number N is independent of the choice of x in the case of uniform convergence.

Criterion for Uniform Convergence

Theorem 4.3.1 (Cauchy criterion)**.** *The sequence of functions $\{f_n\}$ defined on E converges uniformly on E if and only if for any $\varepsilon > 0$, there exists a natural number N such that for any $m, n \geq N$ and $x \in E$, $|f_n(x) - f_m(x)| \leq \varepsilon$.*

Proof. Suppose $\{f_n\} \to f$ uniformly on E. Let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that $\forall n, m > N$ and $\forall x \in E$, $|f_n(x) - f(x)| \leq \frac{\varepsilon}{2}$ and $|f_m(x) - f(x)| \leq \frac{\varepsilon}{2}$. Therefore, $\forall \varepsilon > 0$ we have, $N \in \mathbb{N}$ such that $\forall n, m > N$ and $\forall x \in E$,

$$|f_n(x) - f_m(x)| \leq |f_n(x) - f(x)| + |f_m(x) - f(x)| \leq \varepsilon$$

Let $\varepsilon > 0$. Suppose there exists $N \in \mathbb{N}$ such that $\forall n, m > N$ and $\forall x \in E$,

$$|f_n(x) - f_m(x)| \leq |f_n(x) - f(x)| + |f_m(x) - f(x)| \leq \varepsilon$$

By Cauchy criterion, $\{f_n\} \to f$ pointwise on E. It remains to prove that this convergence is uniform on E. Clearly, above inequality is true for any value of m greater than N. As $m \to \infty$ we have,

$$\lim_{m\to\infty} |f_n(x) - f_m(x)| \leq \varepsilon$$

$$\implies \left| f_n(x) - \lim_{m\to\infty} f_m(x) \right| = |f_n(x) - f(x)| \leq \varepsilon$$

Clearly, the convergence is uniform. That is, $\{f_n\} \to f$ uniformly on E. □

Theorem 4.3.2 (Supremum Test)**.** *Suppose sequence of function $\{f_n\} \to f$ pointwise on E. Suppose $M_n = \sup_{x\in E} |f_n(x) - f(x)|$. Then $\{f_n\} \to f$ uniformly on E, if and only if sequence $\{M_n\} \to 0$.*

Proof. Define $M_n = \sup_{x\in X} |f_n(x) - f(x)|$. Clearly, $M_n \geq 0$. Suppose $\{M_n\} \to 0$ on E. Let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that for every $n > N$, we have $|M_n - 0| = M_n \leq \varepsilon$. Thus for any $n \geq N$ and $x \in E$,

$$|f_n(x) - f(x)| \leq \sup |f_n(x) - f(x)| = M_n \leq \varepsilon$$

Therefore, $\{f_n\} \to f$ uniformly on E.

Suppose $\{f_n\} \to f$ uniformly on E. Let $\varepsilon > 0$. Then, there exists $N \in \mathbb{N}$ such that for every $n \geq N$ and $x \in E$, $|f_n(x) - f(x)| \leq \varepsilon$. This inequality is true for every $x \in E$. Thus,

$$\sup_{x\in E} |f_n(x) - f(x)| = M_n \leq \varepsilon, \quad \forall n \geq N$$

Therefore, $\{M_n\} \to 0$ on E. □

Theorem 4.3.3 (Weierstrass M-test)**.** *Suppose $\{f_n\}$ is a sequence of functions on E. Suppose $|f_n(x)| \leq M_n$. Then $\sum f_n$ converges on E if $\sum M_n$ converges.*

Proof. Suppose $|f_n| \leq M_n$. Then $M_n \geq 0$. Suppose $\sum M_n$ converges. Let $\{s_n\}$ be the sequence of partial sums. Let $\varepsilon > 0$. By Cauchy criterion, there exists $N \in \mathbb{N}$ such that $\forall n, m \geq N$, $|s_n - s_m| \leq \varepsilon$. Thus,

$$|s_n - s_m| = \left| \sum_{j=1}^{n} M_j - \sum_{j=1}^{m} M_j \right| = \sum_{j=n+1}^{m} M_j \leq \varepsilon, \quad (m > n)$$

Let $\{S_n\}$ be the sequence of partial sums for $\sum f_n$. Clearly, $\forall n, m \geq N$ and $\forall x \in E$, we have $|S_n - S_m| \leq \varepsilon$ since

$$|S_n - S_m| = \left| \sum_{j=1}^{n} f_j(x) - \sum_{j=1}^{m} f_j(x) \right| = \left| \sum_{j=n+1}^{m} f_j(x) \right| \leq \sum_{j=n+1}^{m} M_j \leq \varepsilon$$

By Cauchy criterion for uniform convergence, the sequence of partial sums, $\{S_n\}$ converges uniformly on E. Therefore, the series of functions, $\sum f_n$ converges uniformly on E. □

4.3.3 Uniform Convergence and Continuity

Theorem 4.3.4. *Suppose $\{f_n\} \to f$ uniformly on E. Let x be a limit point of E. Suppose $\lim\limits_{t \to x} f_n(t) = A_n$. Then $\{A_n\}$ converges, and $\lim\limits_{t \to x} f(t) = \lim\limits_{n \to \infty} A_n$. In other words,*

$$\lim_{t \to x} \lim_{n \to \infty} f_n(t) = \lim_{n \to \infty} \lim_{t \to x} f_n(t)$$

Proof. Suppose $\{f_n\} \to f$ uniformly on E. Suppose $\lim\limits_{t \to x} f_n(t) = A_n$ on E. Let $\varepsilon > 0$. By Cauchy criterion for uniform convergence, there exists $N \in \mathbb{N}$ such that for every $n \geq N$ and every $t \in E$, $|f_n(t) - f_m(t)| \leq \varepsilon/3$. As $t \to x$, we have $|A_n - A_m| \leq \varepsilon/3$. Clearly, $\{A_n\}$ is Cauchy and $\{A_n\} \to A$.

Now fix a natural number N such that for every $n \geq N$ and $t \in E$, $|f_n(t) - f(t)| < \varepsilon/3$ and $|A_n - A| \leq \varepsilon/3$. Also, choose a neighbourhood V of x such that $|f_n(t) - A_n| \leq \varepsilon/3$. Then, $\forall x \in V \cap E$ we have

$$|f(t) - A| \leq |f(t) - f_n(t)| + |f_n(t) - A_n| + |A_n - A| \leq \varepsilon$$

Therefore, $\lim\limits_{t \to x} f(t) = A$.

$$\lim_{t \to x} \lim_{n \to \infty} f_n(t) = \lim_{t \to x} f(t) = \lim_{n \to \infty} A_n = \lim_{n \to \infty} \lim_{t \to x} f_n(t)$$

□

Theorem 4.3.5. *If $\{f_n\}$ is a sequence of continuous functions on E, and if $f_n \to f$ uniformly on E, then f is continuous on E.*

Uniform convergence preserves continuity.

Proof. Suppose the sequence of continuous functions, $\{f_n\} \to f$ uniformly on E. Since f_n is continuous $\lim_{t \to x} f_n(t) = f_n(x)$, $\forall x \in E$. Since the convergence is uniform, the order of limits can be interchanged.

$$f(x) = \lim_{n \to \infty} f_n(x) = \lim_{n \to \infty} \lim_{t \to x} f_n(t) = \lim_{t \to x} \lim_{n \to \infty} f_n(t) = \lim_{t \to x} f(t)$$

Therefore, the limit function f is continuous. □

Remark. Suppose a sequence of continuous functions $\{f_n\}$ converges to a function f. Function f being continuous doesn't imply that the convergence is uniform.

Proof. Consider $f_n : (0,1) \to \mathbb{R}$ defined by $f_n(x) = \frac{1}{nx+1}$. Then, $\{f_n\} \to 0$. Clearly, $f_n, 0$ are continuous functions on $(0,1)$. However, the convergence is not uniform. □

Theorem 4.3.6. *Suppose K is compact, and*

1. *$\{f_n\}$ is sequence of continuous functions on K*
2. *$\{f_n\}$ converges pointwise to a continuous function f on K.*
3. *$f_n(x) \geq f_{n+1}(x)$, $\forall x \in K$*

Then, $\{f_n\} \to f$ uniformly on K.

Proof. Suppose $\{f_n\} \to f$ pointwise on K. Let function $g_n = f_n - f$. Then, g_n is continuous, $\{g_n\} \to 0$ pointwise and $g_n \geq g_{n+1}$. If sequence $\{g_n\} \to 0$ uniformly, then sequence $\{f_n\} \to f$ uniformly.

Let $\varepsilon > 0$. Let K_n be the set of all points $x \in K$ such that $g_n(x) \geq \varepsilon$. Then K_n is closed, since g_n is continuous. And K_n is compact since K_n is a closed subset of a compact set K.

Suppose $x \in K_{n+1}$. Then $g_{n+1}(x) \geq \varepsilon$. We have, $g(x) \geq g_{n+1}(x)$. Thus, $g(x) \geq \varepsilon$. Therefore, $K_n \supset K_{n+1} \supset \dots$.

Let $x \in K$. We have $\{g_n(x)\} \to 0$ pointwise. Then there exists $N \in \mathbb{N}$ such that $\forall n > N$, $g_n(x) < \varepsilon$. Thus, $x \notin K_n$, $\forall n \geq N$. Therefore $\cap K_n$ is empty.

Clearly, K_N is empty for some $N \in \mathbb{N}$. Thus, $\forall n \geq N$ and $\forall x \in K$ we have, $0 \leq g_n(x) < \varepsilon$. In other words, $\{g_n\}$ converges to 0 uniformly on K. Therefore, $\{f_n\}$ converges to f uniformly on K. □

Definitions 4.3.4. Let X be a metric space. Let $\mathscr{C}(X)$ be the set of all complex valued, continuous, bounded functions on X. Let $f \in \mathscr{C}(X)$. Then **supremum norm** on $\mathscr{C}(X)$ is defined by

$$\|f\| = \sup_{x \in X} |f(x)|$$

And, $\mathscr{C}(X)$ together with **distance function** $d : \mathscr{C}(X) \times \mathscr{C}(X) \to \mathbb{R}$ defined by $d(f,g) = \|f - g\|$ is a metric space.

Let $\mathscr{C}(X)$ be the set of all complex valued, continuous, bounded functions on metric space X. Let $f \in \mathscr{C}(X)$. Then f is bounded, $|f| < \infty$. Thus, $\sup |f(x)|$ exists. Therefore, $\|f\|$ is well-defined.

And, $\|f\| = \sup |f(x)| \geq 0$ since $|f(x)| \geq 0$. Let $f, g \in \mathscr{C}(X)$. Then $h = f + g \in \mathscr{C}(X)$ since sum of continuous functions is functions and sum of bounded functions is bounded. Also, we have

$$\|h\| = \sup_{x \in X} |(f+g)(x)| \leq \sup_{x \in X} |f(x)| + \sup_{x \in X} |g(x)| = \|f\| + \|g\|$$

Therefore, $\|.\| : \mathscr{C}(X) \to \mathbb{R}$ is a norm on $\mathscr{C}(X)$.

Consider the function $d : \mathscr{C}(X) \times \mathscr{C}(X) \to \mathbb{R}$ defined by $d(f,g) = \|f-g\|$. Clearly, $d(f,g) \geq 0$ since $|(f-g)(x)| \geq 0$. And,

$$d(f,g) = 0 \iff |(f-g)(x)| = 0, \forall x \in X \iff f = g$$

Also we have,

$$d(f,g) \leq \|(f-h)+(h-g)\| \leq \|f-h\| + \|h-g\| = d(f,h) + d(h,g)$$

Therefore, the function d is a distance function/metric on $\mathscr{C}(X)$.

Remark. A sequence $\{f_n\}$ converges to f in metric space $\mathscr{C}(X)$ if and only if $\{f_n\}$ converges to f uniformly on X.

Proof.

$$\begin{aligned}
\{f_n\} \to f \text{ in } \mathscr{C}(X) &\iff \forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \forall n > N, d(f_n, f) \leq \varepsilon \\
&\iff \forall n > N, \|f_n - f\| \leq \varepsilon \\
&\iff \forall n > N, \sup |f_n(x) - f(x)| \leq \varepsilon \\
&\iff \forall n > N, \forall x \in X,\ |f_n(x) - f(x)| \leq \varepsilon \\
&\iff \{f_n\} \to f \text{ uniformly on } X
\end{aligned}$$

□

Definitions 4.3.5 (uniformly closed)**.** Closed subset of $\mathscr{C}(X)$ is **uniformly closed** since every convergent sequence in it corresponds to a uniform convergent sequence in X.

Definitions 4.3.6 (uniform closure)**.** Let $\mathscr{A} \subset \mathscr{C}(X)$. Then its closure, $\bar{\mathscr{A}}$ is the **uniform closure** of $\mathscr{A}$.

Definitions 4.3.7 (complete metric space)**.** A metric space is **complete** if every Cauchy sequence in it converges.

For example, $\mathbb{R}$ is complete. But $\mathbb{Q}$ is not complete since a sequence of rational numbers converging to π in $\mathbb{R}$ is Cauchy sequence in $\mathbb{Q}$ which doesn't converge to any point in $\mathbb{Q}$.

Theorem 4.3.7. *Let X be a metric space. Let $\mathscr{C}(X)$ be the set of all complex valued, continuous, bounded functions on X. Let $f, g \in \mathscr{C}(X)$. Define norm $\|f\| = \sup_{x\in X} |f(x)|$ and metric $d(f,g) = \|f-g\|$. Then metric space $\langle \mathscr{C}(X), d\rangle$ is a complete metric space.*

Proof. Let sequence of functions $\{f_n\}$ be a Cauchy sequence in $\mathscr{C}(X)$. Let $\varepsilon > 0$. Then there exists a natural number N such that for any $n, m > N$, $d(f_n, f_m) < \varepsilon$. That is,

$$d(f_n, f_m) = \sup_{x\in X} |f_n(x) - f_m(x)| < \varepsilon$$

Define $M_n = \sup_{x\in X} |f_n(x)|$. Then,

$$|M_n - M_m| = \left| \sup_{x\in X} |f_n(x)| - \sup_{x\in X} |f_m(x)| \right| \leq \sup_{x\in X} |f_n(x) - f_m(x)| < \varepsilon$$

We have, $\{M_n\}$ is a Cauchy sequence in $\mathbb{R}$. We know that, every Cauchy sequence in $\mathbb{R}$ is convergent in $\mathbb{R}$ since $\mathbb{R}$ is complete. Thus, $\{M_n\}$ converges.

We have, corresponding sequence $\{f_n\}$ in X such that sequence $\{M_n\}$ defined by $M_n = \sup_{x\in X} |f_n(x)|$ converges. Therefore, $\{f_n\}$ converges to f uniformly on X by criterion of uniform convergence.

We have, f_n are continuous, bounded functions on X. And $\{f_n\} \to f$ uniformly on X.

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n > N, \forall x \in X, |f_n(x) - f_m(x)| < \varepsilon$$

Since, the convergence is uniform, the limit function f is continuous and bounded. Therefore, $\{f_n\}$ in $\mathscr{C}(X)$ converges to $f \in \mathscr{C}(X)$.

The Cauchy sequence $\{f_n\}$ is chosen arbitrarily. Thus, every Cauchy sequence in $\mathscr{C}(X)$ is convergent. Therefore, $\mathscr{C}(X)$ is complete. □

Remark. [Rudin, 1976, Exercise 7.1] Let $\{f_n\}$ be a sequence of bounded functions on a metric space X. Suppose $\{f_n\} \to f$ uniformly on X. Then f is bounded.

Proof. Suppose the sequence of bounded functions, $\{f_n\}$ converges to f uniformly on X. Consider $\{g_n\}$ where $g_n = |f_n - f|$ Then, $\{g_n\}$ converges to 0 uniformly. Suppose f is unbounded. Then g_n is an unbounded sequence which doesn't converge. This is a contradiction. Therefore, f is bounded. □

4.3.4 Uniform Convergence and Integration

Theorem 4.3.8. *Let α be monotonically increasing on $[a, b]$. Suppose $f_n \in \mathscr{R}(\alpha)$ on $[a, b]$. Suppose $\{f_n\} \to f$ uniformly on $[a, b]$. Then $f \in \mathscr{R}(\alpha)$ on $[a, b]$. And,*

$$\int_a^b f \; d\alpha = \lim_{n\to\infty} \int_a^b f_n \; d\alpha$$

Proof. Let $f_n \in \mathscr{R}(\alpha)$ on $[a, b]$. Suppose sequence $\{f_n\} \to f$ uniformly on $[a, b]$. Define

$$\varepsilon_n = \sup_{x\in[a,b]} |f_n(x) - f(x)|$$

Then, $-\varepsilon_n \leq f - f_n \leq \varepsilon_n, \quad \forall x \in [a, b]$. And,

$$f_n - \varepsilon_n \leq f \leq f_n + \varepsilon_n$$

Since $f_n \in \mathscr{R}(\alpha)$, we have $f_n + \varepsilon, f_n - \varepsilon \in \mathscr{R}(\alpha)$. From the definition of lower and upper integrals,

$$\int_a^b (f_n - \varepsilon_n) \; d\alpha \leq \underline{\int_a^b} f \; d\alpha \leq \overline{\int_a^b} f \; d\alpha \leq \int_a^b (f_n + \varepsilon_n) \; d\alpha$$

Thus,

$$0 \leq \overline{\int_a^b} f \; d\alpha - \underline{\int_a^b} f \; d\alpha \leq \int_a^b 2\varepsilon_n \; d\alpha = 2\varepsilon_n \left[\alpha(b) - \alpha(a)\right]$$

Clearly, $\varepsilon_n \to 0$ as $n \to \infty$. Then, $f \in \mathscr{R}(\alpha)$ on $[a, b]$ since the lower and upper integrals of f with respect to α are equal.

Also we have, $0 \leq |f - f_n| \leq \varepsilon_n$ for every $x \in [a, b]$.
Therefore,

$$0 \leq \left| \int_a^b f\ d\alpha - \int_a^b f_n\ d\alpha \right| \leq \int_a^b |f - f_n|\ d\alpha \leq \int_a^b \varepsilon_n\ d\alpha = \varepsilon_n \left[\alpha(b) - \alpha(a)\right]$$

As $n \to \infty$, we have $\varepsilon_n \to 0$. Thus,

$$\int_a^b f\ d\alpha = \lim_{n\to\infty} \int_a^b f_n\ d\ \alpha$$

□

Corollary 4.3.8.1. *If $f_n \in \mathscr{R}(\alpha)$ on $[a, b]$ and if*

$$f(x) = \sum_{n=1}^{\infty} f_n(x)$$

the series converging uniformly on $[a, b]$, then

$$\int_a^b f\ d\alpha = \sum_{n=1}^{\infty} \int_a^b f_n\ d\alpha$$

In other words, the series may be integrated term by term.

Proof. Suppose $f_n \in \mathscr{R}(\alpha)$ and $\sum f_n$ converges to f uniformly on $[a, b]$. Let $\{S_n\}$ be the sequence of partial sums converging to f uniformly on $[a, b]$. We have, $S_n \in \mathscr{R}(\alpha)$ on $[a, b]$ by linearity of the integral and mathematical induction.

$$\sum_{j=1}^{n} \int_a^b f_n\ d\alpha = \int_a^b \sum_{j=1}^{n} f_j\ d\alpha = \int_a^b S_n\ d\alpha$$

Since the uniform limit function of integrable functions is integrable,

$$\sum_{n=1}^{\infty} f_n = \lim_{n\to\infty} S_n = f \in \mathscr{R}(\alpha) \text{ and } \lim_{n\to\infty} \int_a^b S_n\ d\alpha = \int_a^b f\ d\alpha$$

By the definition of sum of series,

$$\sum_{n=1}^{\infty} \int_a^b f_n\ d\alpha = \lim_{n\to\infty} \sum_{j=1}^{n} \int_a^b f_j\ d\alpha = \lim_{n\to\infty} \int_a^b S_n\ d\alpha = \int_a^b f\ d\alpha = \int_a^b \sum_{n=1}^{\infty} f_n\ d\alpha$$

Thus, the series may be integrated term by term. □

4.3.5 Uniform Convergence and Differentiation

Theorem 4.3.9. *Suppose $\{f_n\}$ is a sequence of functions, differentiable on $[a, b]$ and $\{f_n(x_0)\}$ converges for some $x_0 \in [a, b]$. If $\{f_n'\}$ converges uniformly on $[a, b]$, then $\{f_n\}$ converges uniformly on $[a, b]$ to a function f. And,*

$$f'(x) = \lim_{n\to\infty} f_n'(x)$$

Proof. Suppose $\{f_n(x_0)\} \to f(x_0)$. Suppose $\{f'_n\}$ converges uniformly on $[a,b]$. Let $\varepsilon > 0$. Then there exists a natural number N such that

$$\forall n, m > N,\ |f_n(x_0) - f_m(x_0)| < \frac{\varepsilon}{2} \text{ and}$$
$$\forall n, m > N,\ |f'_n(t) - f'_m(t)| < \frac{\varepsilon}{2(b-a)}$$

Let $x, t \in [a,b]$.

$$|f_n(x) - f_m(x) - f_n(t) + f_m(t)| = |(f_n - f_m)(x) - (f_n - f_m)(t)|$$

By mean value theorem, there exists $y \in (x,t)$ such that

$$\begin{aligned} |f_n(x) - f_m(x) - f_n(t) + f_m(t)| &= |(x-t)\ (f_n - f_m)'(y)| \\ &\le \frac{|x-t|\varepsilon}{2(b-a)} \le \frac{\varepsilon}{2} \end{aligned}$$

since $|(f_n - f_m)'(x)| = |f'_n(x) - f'_m(x)| \le \frac{\varepsilon}{2(b-a)}$.

And the inequality is true of $t = x_0$. Thus,

$$\begin{aligned} |f_n(x) - f_m(x)| &= |f_n(x) - f_m(x) - f_n(x_0) + f_m(x_0) + f_n(x_0) - f_m(x_0| \\ &\le |f_n(x) - f_m(x) - f_n(x_0) + f_m(x_0)| + |f_n(x_0) - f_m(x_0| \le \varepsilon \end{aligned}$$

Thus, $\forall \varepsilon > 0$, we have a natural number N such that $\forall n, m > N$, $\forall x \in [a,b]$, $|f_n(x) - f_m(x)| \le \varepsilon$. In other words, $\{f_n\} \to f$ uniformly on $[a,b]$.

Fix $x \in [a,b]$. Define functions ϕ_n, ϕ on $[a,b]$ except at $t = x$,

$$\phi_n(t) = \frac{f_n(t) - f_n(x)}{t - x} \quad \text{and} \quad \phi(t) = \frac{f(t) - f(x)}{t - x}$$

Then,

$$\lim_{n\to\infty} \phi_n(t) = \lim_{n\to\infty} \frac{f_n(t) - f_n(x)}{t - x} = \frac{\lim\limits_{n\to\infty} f_n(t) - \lim\limits_{n\to\infty} f_n(x)}{t - x} = \frac{f(t) - f(x)}{t - x} = \phi(t)$$

since sequence $\{f_n\} \to f$ uniformly on $[a,b]$. Clearly, $\{\phi_n\}$ converges to ϕ pointwise on $[a,b] - \{x\}$.

We have,

$$\lim_{t\to x} \phi_n(t) = \lim_{t\to x} \frac{f_n(t) - f_n(x)}{t - x} = f'_n(x)$$
$$\lim_{t\to x} \phi(t) = \lim_{t\to x} \frac{f(t) - f(x)}{t - x} = f'(x)$$

Also we have,

$$|\phi_n(t) - \phi_m(t)| = \frac{|f_n(t) - f_n(x) - f_m(t) + f_m(x)|}{t - x} \le \frac{\varepsilon}{2(b-a)}$$

That is $\forall n, m > N$, $|\phi_n(t) - \phi_m(t)| < \varepsilon$. Now, by Cauchy criterion for uniform convergence, sequence $\{\phi_n\}$ converges to ϕ uniformly on $[a,b] - \{x\}$.

We know that continuity is preserved under uniform convergence. Therefore,

$$f'(t) = \lim_{t \to x} \phi(t) = \lim_{t \to x} \lim_{n \to \infty} \phi_n(t) = \lim_{n \to \infty} \lim_{t \to x} \phi_n(t) = \lim_{n \to \infty} f_n'(x)$$

$\square$

Theorem 4.3.10. *There exists a real continuous function on the real line which is nowhere differentiable.*

Proof. Define $\phi : [-1, 1] \to \mathbb{R}$ by $\phi(x) = |x|$. Extend ϕ from $[-1, 1]$ to $\mathbb{R}$ such that $\phi(x + 2) = \phi(x)$. Then $|\phi(s) - \phi(t)| \le |s - t|$, $\forall s, t \in \mathbb{R}$.

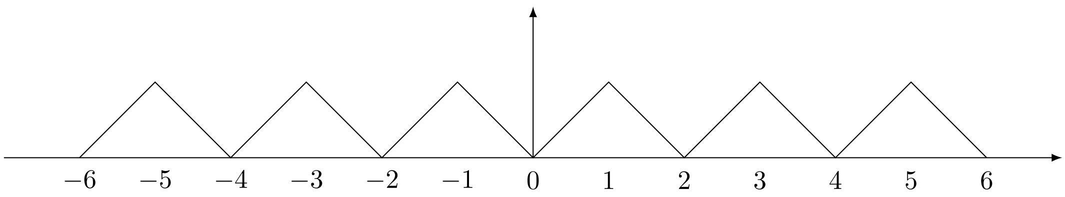


Figure 4.1: Graph of ϕ

Define $f_n(x) = \left(\frac{3}{4}\right)^n \phi(4^n x)$. Then $|f_n| \le \left(\frac{3}{4}\right)^n$. By Weierstrass M test, $\sum_{n=0}^{\infty} f_n$ converges uniformly on $\mathbb{R}$. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \sum_{n=0}^{\infty} \left(\frac{3}{4}\right)^n \phi(4^n x)$. This sum function f is continuous since the convergence is uniform and f_n are all continuous.

Fix a real number x and a positive integer m. Define $\delta_m = \pm\frac{1}{2}\ 4^{-m}$ such that no integer lies between $4^m x$ and $4^m(x + \delta_m)$. This is possible since $|4^m(x + \delta_m) - 4^m x| = |4^m \delta_m| = \frac{1}{2}$.

The choice of sign of δ_m is made depending on the value of $4^m x$. Suppose $x = 1.1$ and $m = 3$. Then $4^m x = 70.4$. Now $\delta = \frac{1}{2}\ 4^{-m} = \frac{1}{128}$ so that $4^m(x + \delta_m) = 70.9$. Suppose $x = 1.2$ and $m = 3$. Then $4^m x = 76.8$. Now $\delta = -\frac{1}{2}\ 4^{-m} = -\frac{1}{128}$ so that $4^m(x + \delta) = 76.3$.

Define

$$\gamma_n = \frac{\phi(4^n(x + \delta_m)) - \phi(4^n x)}{\delta_m}$$

If $n > m$, then $4^n \delta_m$ is an even integer and $\gamma_n = 0$.
If $n \le m$, then

$$|\gamma_n| = \left| \frac{\phi(4^n(x + \delta_m)) - \phi(4^n x)}{\delta_m} \right| \le \frac{4^n \delta_m}{\delta_m} \le 4^n$$

since $|\phi(s) - \phi(t)| \le |s - t|$.
In particular, if $n = m$, then

$$|\gamma_m| = \left| \frac{\phi(4^m(x + \delta_m)) - \phi(4^m x)}{\delta_m} \right| = 4^m$$

since $|\phi(4^m(x+\delta_m)) - \phi(4^m x)| = |4^m \delta| = \frac{1}{2}$ as there are no integers between $4^m(x+\delta_m)$ and $4^m x$.

From the definition of γ_n we have,

$$\begin{aligned}
\left|\frac{f(x+\delta_m)-f(x)}{\delta_m}\right| &= \left|\sum_{n=0}^{m}\left(\frac{3}{4}\right)^n \gamma_n\right| \\
&= \left|3^m + \sum_{n=0}^{m-1}\left(\frac{3}{4}\right)^n \gamma_n\right| \\
&\geq 3^m - \sum_{n=0}^{m-1} 3^n = \frac{3^m+1}{2}
\end{aligned}$$

Therefore, the function f is not differentiable at x since the following limit does not exist as $m \to \infty$.

$$\lim_{\delta_m \to 0}\left|\frac{f(x+\delta_m)-f(x)}{\delta_m}\right| \geq \lim_{m\to\infty}\frac{3^m+1}{2}$$

Since the choice of x is arbitrary, the function f is nowhere differentiable. $\square$

Module 4 : Weierstrass Approximation & Some Special Functions

4.4 Equicontinuous Family of Functions

Definitions 4.4.1. Let $\{f_n\}$ be a sequence of functions defined on E.
Sequence $\{f_n\}$ is a **bounded** sequence if every functions in the sequence is bounded.

$$\forall n \in N, \exists M \in \mathbb{R} \text{ such that } \forall x \in E, |f_n(x)| < M$$

Sequence $\{f_n\}$ is a **pointwise bounded** sequence if there exists a function $\phi(x)$ such that $|f_n(x)| < \phi(x)$. In other words, sequence $\{f_n\}$ is a pointwise bounded sequence if $\{f_n(x)\}$ is bounded for every $x \in E$

$$\forall x \in E, \exists M \in \mathbb{R} \text{ such that } \forall n \in \mathbb{N}, |f_n(x)| < M$$

Sequence $\{f_n\}$ is a **uniformly bounded** sequence if there exists a real number M such that $|f_n(x)| < M$ for every $x \in E$ and $n \in \mathbb{N}$.

$$\exists M \in \mathbb{R} \text{ such that } \forall x \in E, \forall n \in \mathbb{N}, |f_n(x)| < M$$

4.4.1 Two Problems

1. Whether uniform bounded sequence of uniformly bounded functions have a convergent subsequence ? **NO**.
   Sequence $\{f_n\}$ where $f_n(x) = \sin nx$ is uniformly bounded. But it doesn't have a convergent subseqeunce.

   Proof. Suppose $\{\sin nx\}$ has a convergent subsequence, say $\{\sin n_k x\}$. Then by Cauchy criterion, $(\sin n_k x - \sin n_{k+1} x) \to 0$ as $n_k \to \infty$.

   $$\lim_{n\to\infty} \sin n_k x - \sin n_{k+1} x = 0$$

   $$\lim_{n\to\infty} (\sin n_k x - \sin n_{k+1} x)^2 = 0$$

   By Lebesgue dominated convergence theorem,

   $$\lim_{n\to\infty} \int_0^{2\pi} (\sin n_k x - \sin n_{k+1} x)^2 dx = \int_0^{2\pi} \lim_{n\to\infty} (\sin n_k x - \sin n_{k+1} x)^2 dx = 0$$

   However, the integral on the left evaluates 2π which is a contradiction. Clearly, the uniformly bounded sequence $\{\sin nx\}$ of continuous functions on compact interval $[0, 2\pi]$ does not even imply existence of a subsequence which converges pointwise. □

2. Whether every uniformly bounded, convergent sequence has a uniformly convergent subsequence ? **NO**.
   Consider,

   $$f_n(x) = \frac{x^2}{x^2 + (1 - nx)^2}$$

   Sequence $\{f_n\}$ is a uniformly bounded sequence of functions on compact interval $[0, 1]$. doesn't have a convergent subsequence.

Proof. Suppose $\{f_n\}$ has a convergent subsequence $\{f_{n_k}\} \to f$. Then $\lim_{n\to\infty} f_{n_k}(x) = f(x)$. We have,

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \frac{x^2}{x^2 + (1-nx)^2} = 0$$

However,

$$\lim_{n\to\infty} f_n\left(\frac{1}{n}\right) = \lim_{n\to\infty} \frac{\left(\frac{1}{n}\right)^2}{\left(\frac{1}{n}\right)^2 + 0} = 1 \neq 0 = \left(\lim_{n\to\infty} f_n\right)(x)$$

Clearly, the sequence of functions $\{f_n\}$ is a uniformly bounded sequence of continuous functions on a compact interval $[0,1]$. Therefore, uniformly bounded, convergent sequence on compact set doesn't necessarily have a uniformly convergenct subsequence. □

Definitions 4.4.2. Let E be a subset of a metric space X. A family $\mathscr{F}$ of complex functions on E is **equicontinuous** if for any $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $d(x,y) < \delta$ for any $x, y \in E$ and $f \in \mathscr{F}$.

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ such that } \forall f \in \mathscr{F}, |f(x) - f(y)| < \varepsilon \text{ whenever } d(x,y) < \delta$$

In equicontinuity, the choice of δ is independent of the choice of f.

Theorem 4.4.1. *If $\{f_n\}$ is a sequence of pointwise bounded, complex functions on countable set E, then it has a subsequence $\{f_{n_k}\}$ such that $\{f_{n_k}(x)\}$ converges for every $x \in E$.*

Proof. Let E be a countable set. Let $\{f_n\}$ be a sequence of pointwise bounded, complex functions. Let $\{x_i\}$ be a sequence in E. Then $\{f_n(x_1)\}$ is a bounded sequence of complex numbers. By Bolzano-Weierstrass theorem, $\{f_n(x_1)\}$ has a subsequence S_1, say $\{f_{1,k}(x_1)\}$ which converges at $x_1 \in E$.

$$S_1 : f_{1,1}\ f_{1,2}\ f_{1,3} \dots$$

Consider, the subsequence $\{f_{2,k}(x_2)\}$ of the sequence $\{f_n(x_2)\}$ with the same values for k as in subsequence $\{f_{1,k}(x_2)\}$. Again, $\{f_{2,k}(x_2)\}$ is a bounded complex sequence. And by Bolzano-Weierstrass theorem, $\{f_{2,k}(x_2)\}$ has a convergent subsequence S_2, say $\{f_{2,k'}(x_2)\}$ which is a subsequence of $\{f_2(x_2)\}$. And more importantly, S_2 is a subsequence of S_1 such that S_2 converges for both $x_1, x_2 \in E$.

$$S_2 : f_{2,1}\ f_{2,2}\ f_{2,3} \dots$$

Continuing like this, we get a sequence of subsequences

$$\begin{array}{llllll} S_1 & : f_{n,1} & f_{n,2} & f_{n,3} & \dots \\ S_2 & : f_{n,1} & f_{n,2} & f_{n,3} & \dots \\ & \dots & & & \\ S_n & : f_{n,1} & f_{n,2} & f_{n,3} & \dots \end{array}$$

Consider the diagonal sequence, $S : f_{1,1}\ f_{2,2}\ f_{3,3} \dots$. We know that discarding finitely many first terms won't affect convergence of sequences. And for any natural number n, we have can obtain a subsequence of S_n by discarding a finite number of first terms from S. Thus, we have sequence S which converges for $x_1, x_2, \ldots, x_n \in E$ since S_n converges. Therefore, as $n \to \infty$ we have a convergent subsequence S of $\{f_n\}$ which converges for every $x \in E$. □

Theorem 4.4.2. *Let K be a compact metric space. If $\{f_n\}$ is a sequence of pointwise bounded, continuous, complex valued functions on K converges uniformly on K, then $\{f_n\}$ is equicontinuous on K.*

Proof. Let K be a compact metric space. Let $\varepsilon > 0$. Let $\{f_n\}$ converges uniformly on K. Then,

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N} \text{ such that } \forall n > N,\ \forall x \in K,\ \|f_n(x) - f_N(x)\| < \varepsilon$$

We have continuous functions on compact sets are uniformly continuous. Thus for any $\varepsilon > 0$ there exist $\delta > 0$ such that $|f_j(x) - f_j(y)| < \varepsilon$ whenever $d(x, y) < \delta$. Thus, $|f_N(x) - f_N(y)| \le \varepsilon$.

Let $1 \le j \le N$. Since the continuity is uniform, there exists $\delta > 0$ such that $|f_j(x) - f_j(y)| < \varepsilon$.†[3]

Let $n > N$ and $d(x, y) < \delta$. Then,

$$|f_n(x) - f_n(y)| \le |f_n(x) - f_N(x)| + |f_N(x) - f_N(y)| + |f_N(y) - f_n(y)| < 3\varepsilon$$

Therefore, for any $\varepsilon > 0$ there exists $\delta > 0$ such that $\forall n \in \mathbb{N}$, $|f_n(x) - f_m(x)| < \varepsilon$ whenever $d(x, y) < \delta$.

$$\forall \varepsilon > 0,\ \exists \delta > 0,\ \forall n \in \mathbb{N}, \underset{d(x,y)<\delta}{\forall\ y} \in K,\ |f_n(x) - f_n(y)| < \varepsilon$$

That is, the sequence $\{f_n\}$ is equicontinuous on K. □

Theorem 4.4.3. *Let K a compact. If $\{f_n\}$ be a sequence of pointwise bounded, equicontinuous, complex functions on K, then*

1. *$\{f_n\}$ is uniformly bounded on K.*
2. *$\{f_n\}$ has a uniformly convergent subsequence.*

Proof. Let $\{f_n\}$ be sequence of point-wise bounded, equicontinuous, complex functions on compact set K. Since f_n are equicontinuous, given $\varepsilon > 0$, there exists $\delta > 0$ such that $\forall n \in \mathbb{N}$, $|f_n(x) - f_n(y)| < \varepsilon$ whenever $d(x, y) < \delta$.

Let $\mathcal{C}$ be a cover of K of open balls of radius δ. Since, K is compact this cover has a finite subcover, say open balls with center p_j, $j = 1, 2, \ldots, r$. Thus there exists finitely many points, p_j such that every point in K is sufficiently close one of them. Then for any $x \in K$, there exists some p_j such that $d(x, p_j) < \delta$.

[3]For each functions f_j, given $\varepsilon > 0$, there exists $\delta_j > 0$ satisfying the ε-δ condition for uniform continuity. Define $\delta = \min\{\delta_j : j = 1, 2, \ldots, N\}$, then for this δ the condition is satified by functions $f_1, f_2, \ldots, f_N$.

Since $\{f_n\}$ is point-wise bounded, there exists M_j such that $|f_n(p_j)| < M_j$. Define $M = \max\{M_1, M_2, \ldots, M_r\}$. Then

$$\begin{aligned} |f_n(x)| &= |f_n(x) - f_n(p_j) + f_n(p_j)| \\ &\le |f_n(x) - f_n(p_j)| + |f_n(p_j)| \\ &\le \varepsilon + M \end{aligned}$$

Therefore, $\{f_n\}$ is uniformly bounded.

Let $\varepsilon > 0$. Choose $\delta > 0$ such that $|f_n(x) - f_n(y)| < \varepsilon$ whenever $d(x, y) < \delta$. Since K is compact, K has a countable dense subset, say $E = \{x_1, x_2, \ldots\}$. That is, given $\delta > 0$, for any $x \in K$, there exists $x_j \in E$ such that $d(x, x_j) < \delta$. Clearly, K has a cover of open balls with center x_js and radius δ. Since K is compact, there exists finitely many points $x_1, x_2, \ldots, x_m \in E$ such that $d(x, x_m) < \delta$.

Since E is countable and $\{f_n\}$ is point-wise bounded, $\{f_n\}$ on E has a subsequence, say diagonal sequence $f_{n_i} = g_i$ which converges for any $x_j \in E$. Thus, by Cauchy criterion there exists integer N such that

$$\forall i, j \ge N, \ |g_i(x_s) - g_j(x_s)| < \varepsilon, \quad s = 1, 2, \ldots, m$$

Let $x \in K$. Let $x_s \in E$ such that $d(x, x_s) < \delta$. Then,

$$|g_i(x) - g_j(x)| \le |g_i(x) - g_i(x_s)| + |g_i(x_s) - g_j(x_s)| + |g_j(x_s) - g_j(x)| \le 3\varepsilon$$

That is, $\{g_i\}$ is uniformly convergent on K. Therefore, $\{f_n\}$ has a uniformly convergent subseqeunce $\{f_{n_k}\}$. □

Theorem 4.4.4 (Weierstrass' Theorem)**.** *If f is a continuous, complex function on $[a, b]$, there exists a sequence of polynomials P_n such that*

$$\lim_{n\to\infty} P_n(x) = f(x)$$

uniformly on $[a, b]$. If f is real, the P_n may be taken real.

In other words, any continuous function has a polynomial approximation.

Proof. Without loss of generality, suppose $[a, b] = [0, 1]$ and $f(0) = f(1) = 0$.

Step 1 : WLoG $f(0) = f(1) = 0$ If theorem is true for continuous functions satisfying $f(0) = f(1) = 0$, then it is true for any continuous function.

Let f be any continuous function on $[0, 1]$. Then there exists a function $g : [0, 1] \to \mathbb{R}$ be defined by,

$$g(x) = f(x) - f(0) - x[f(1) - f(0)]$$

such that $g(0) = g(1) = 0$. Suppose g can be expressed as limit function of a uniformly convergent sequence of polynomials, say P_n. We have,

$$(f - g)(x) = x[f(1) - f(0)] + f(0)$$

is a polynomial, say $P(x)$. Then $P_n+P \to f$ uniformly as $n \to \infty$, since $P_n \to g$ uniformly and P is a constant polynomial independent of n.

$$\lim_{n\to\infty}(P_n+P)(x) = \lim_{n\to\infty}P_n(x)+P(x) = g(x)+(f-g)(x) = f(x)$$

Therefore, it is enough to prove the theorem is true for any continuous function f satisfying $f(0)=f(1)=0$.

Step 2 : Construction of $P_n(x)$

Since f is continous in $[0,1]$, f is uniformly continuous in $[0,1]$. Extend f such that $f(x)=0,\ \forall x \notin [0,1]$. Then, f is uniformly continuous on $\mathbb{R}$. Define $Q_n(x)=c_n(1-x^2)^n$ such that

$$\int_{-1}^{1} Q_n(x)\ dx = 1$$

Then we have,

$$\int_{-1}^{1}(1-x^2)^n\ dx = 2\int_{0}^{1}(1-x^2)^n\ dx \ \text{ since } (1-x^2)^n \text{ is even}$$

We have, Bernouli's inequality. $(1+x)^r \geq (1+rx),\ \forall x \geq -1,\ \forall r \geq 0$

$$\begin{aligned}\int_{-1}^{1}(1-x^2)^n\ dx &\geq 2\int_{0}^{1/\sqrt{n}}(1-nx^2)\ dx\\ &\geq 2\left(\frac{1}{\sqrt{n}}-\frac{n}{3n\sqrt{n}}\right) = \frac{4}{3\sqrt{n}} > \frac{1}{\sqrt{n}}\end{aligned}$$

Clearly, $c_n < \sqrt{n}$.

For $\delta > 0$, $Q_n(x) \leq \sqrt{n}(1-\delta^2)^n$ for $\delta \leq |x| \leq 1$. Then $Q_n \to Q$ uniformly for all such that $\delta \leq |x| \leq 1$. Define $P_n : [0,1] \to \mathbb{R}$ defined by

$$P_n(x) = \int_{-1}^{1} f(x+t)\ Q_n(t)\ dt$$

Then,

$$\begin{aligned}P_n(x) &= \int_{-1}^{1} f(x+t)\ Q_n(t)\ dt\\ &= \int_{-x}^{1-x} f(x+t)\ Q_n(t)\ dt \text{ since } f \text{ vanishes outside } [0,1]\\ &= \int_{0}^{1} f(t)\ Q_n(t-x)\ dt\end{aligned}$$

Continuous function f is uniformly continuous on compact interval $[0,1]$. Thus f is Riemann integrable on $[0,1]$. And $Q_n(t-x) = c_n[1-(t+x)^2]^n$. From integration by parts, we know that $\{P_n\}$ is a sequence of polynomials.

$$\begin{aligned}
P_n(x) &= \int_0^1 f(t)Q_n(t+x)dt \\
&= [f(t)Q_n'(t+x)]_0^1 - \int_0^1 Q_n'(t+x)\int_0^1 f(t)dt \\
&= f(1)Q_n'(1+x) - f(0)Q_n'(x) - [F(1)-F(0)]\int_0^1 Q_n'(t+x)dt \\
&= f(1)Q_n'(1+x) - f(0)Q_n'(x) - [F(1)-F(0)][Q_n(1+x)-Q_n(x)]
\end{aligned}$$

And for each natural number n, we have $P_n(x)$ is real, if f is real.

Step 3 : $P_n \to f$ uniformly

Let $\varepsilon > 0$. Since extended f is uniformly continuous on real line, there exists $\delta > 0$ such that $|f(y) - f(x)| < \frac{\varepsilon}{2}$ whenever $|y - x| < \delta$.

$$\begin{aligned}
|P_n(x) - f(x)| &= \left| \int_{-1}^1 f(x+t)Q_n(t)\ dt - f(x)\int_{-1}^1 Q_n(t)\ dt \right| \\
&\leq \int_{-1}^1 |f(x+t) - f(x)| Q_n(t)\ dt
\end{aligned}$$

Let $M = \sup f(x)$. Then $|f(x+t) - f(x)| \leq 2M$. And we have an upper bound for the value $Q_n(x)$ for $\delta \leq |x| \leq 1$. Therefore, we split the domain of integral into three parts so that we may apply uniform continuity on the middle part and bound of Q_n on other two parts.

$$|P_n(x) - f(x)| \leq 2M\int_{-1}^{-\delta} Q_n(t)dt + \frac{\varepsilon}{2}\int_{-\delta}^{\delta} Q_n(t)dt + 2M\int_{\delta}^1 Q_n(t)dt$$

We have an upper bound for Q_n, say $Q_n(x) \leq \sqrt{n}(1-\delta^2)^n$ for $\delta \leq |x| \leq 1$.

$$\begin{aligned}
|P_n(x) - f(x)| &\leq 2M\sqrt{n}(1-\delta^2)^n\int_{-1}^{-\delta} dt + \frac{\varepsilon}{2}\int_{-\delta}^{\delta} Q_n(t)dt + 2M\sqrt{n}(1-\delta^2)^n\int_{\delta}^1 dt \\
&\leq 4M\sqrt{n}(1-\delta^2)^n + \frac{\varepsilon}{2} \quad \text{ since } 2(1-\delta) < 2 \\
&\leq \varepsilon \text{ for sufficiently large n}
\end{aligned}$$

Therefore, there exists $N \in \mathbb{N}$ such that $\forall n > N,\ |P_n(x) - f(x)| < \varepsilon$. In other words, $P_n \to f$ uniformly on $[0,1]$. □

Corollary 4.4.4.1. *For every interval $[-a,a]$ there is a sequence of real polynomials P_n such that $P_n(0) = 0$ and*

$$\lim_{n\to\infty} P_n(x) = |x|$$

uniformly on $[-a,a]$.

Proof. By Weierstrass theorem, there exists a sequence $\{P_n^*(x)\}$ of real polynomials which converges to $|x|$ uniformly on $[-a,a]$. Thus, $P_n^*(0) \to 0$ as $n \to \infty$.

Define $P_n(x) = P_n^*(x) - P_n^*(0)$. Clearly, $P_n(0) = 0$ and the sequence $\{P_n(x)\}$ converges uniformly on $[-a, a]$. And,

$$\lim_{n\to\infty} P_n(x) = \lim_{n\to\infty} P_n^*(x) - \lim_{n\to\infty} P_n^*(0) = |x|$$

Therefore, $P_n(0) = 0$ and $P_n(x) \to |x|$ as $n \to \infty$. □

Remark (Out of Syllabus). On the other side of this corollary, we have Stone-Weierstrass theorem which study functions that doesn't vanish anywhere.

Let $\mathscr{A}$ be an algebra of functions defined on compact set K that separates points and vanishes nowhere. By Stone-Weierstrass theorem,†[4] any continuous function f on K has a sequence of functions in $\mathscr{A}$ which converges to f uniformly on K. The algebra of even polynomials doesn't separate points since $P_n(x) = P_n(-x)$.

4.4.2 Some special functions

Definitions 4.4.3 (analytic function)**.** A function f is (real) analytic if it can be represented by a power series.

$$f(x) = \sum_{j=1}^{\infty} c_n x^n$$

Remark. The open interval in which a power series $\sum c_n x^n$ converges is the **interval of convergence**.

Theorem 4.4.5. *Suppose series $\sum c_n x^n$ converges for $|x| < R$. Suppose function $f : (-R, R)$ is defined by*

$$f(x) = \sum_{n=0}^{\infty} c_n x^n, \ |x| < R$$

Then,for any $\varepsilon > 0$, the series $\sum c_n x^n$ converges uniformly on $[-R+\varepsilon, R-\varepsilon]$. Also the function f is continuous and differentiable in $(-R, R)$ and

$$f'(x) = \sum_{n=1}^{\infty} n c_n x^{n-1}, \quad |x| < R$$

Proof. Let $\varepsilon > 0$. Then for $|x| \leq R - \varepsilon$ we have $|c_n x^n| \leq |c_n(R-\varepsilon)^n|$.

We have,

$$\limsup_{n\to\infty} |R-\varepsilon| \sqrt[n]{|c_n|} = (R-\varepsilon) \limsup_{n\to\infty} \sqrt[n]{|c_n|} = 0 < 1$$

By raio test, the series $\sum c_n (R-\varepsilon)^n$ converges absolutely. By Weierstrass M test, series $\sum c_n x^n$ converges uniformly on $[-R+\varepsilon, R-\varepsilon]$. In the same fashion, the series $\sum n c_n x^{n-1}$ converges uniformly on $[-R+\varepsilon, R-\varepsilon]$ since

[4] [Rudin, 1976, §7.32] Let $\mathscr{A}$ be a family of functions defined on compact set K.
$\mathscr{A}$ separates points if $\forall x \in K,\ \exists f, g \in \mathscr{A}$ such that $f(x) \neq g(x)$.
$\mathscr{A}$ vanishes at a point $x \in K$ if $\forall x \in K,\ \exists f \in \mathscr{A}$ such that $f(x) \neq 0$

$\limsup_{n\to\infty} \sqrt[n]{n|c_n|} = 0$.

Let $x_0 \in (-R, R)$. Then there exists $\varepsilon > 0$ such that $x_0 \in [-R+\varepsilon, R-\varepsilon]$. The series $\sum nc_n x^{n-1}$ converges to $f'(x)$ uniformly on $[-R+\varepsilon, R-\varepsilon]$ and $f(x_0) = \sum c_n x_0^n$. Thus, f is differentiable on $[-R+\varepsilon, R-\varepsilon]$ and

$$f'(x) = \lim_{n\to\infty} \sum_{k=1}^{n} kc_k x^{k-1} = \sum_{n=1}^{\infty} nc_n x^{n-1}$$

We know that f is continuous at a point if it is differentiable at that point. Thus, f is continuous on $[-R+\varepsilon, R-\varepsilon]$. □

Corollary 4.4.5.1. *Suppose* $\sum_{n=0}^{\infty} c_n x^n$ *converges for* $|x| < R$ *and function* f *is defined by* $f(x) = \sum_{n=0}^{\infty} c_n x^n$. *Then* f *has derivatives of all ordres, say* $f^{(k)}(x)$ *given by*

$$f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)\dots(n-k+1)c_n x^{n-k}$$

In particular,

$$f^{(k)}(0) = k!\ c_k$$

Proof. Let $f(x) = \sum c_n x^n$. Then, we have

$$f'(x) = \sum_{n=1}^{\infty} nc_n x^n$$

By mathematical induction, we have

$$f^{(k)}(x) = \sum_{n=k}^{\infty} n(n-1)\dots(n-k+1)c_n x^{n-k}$$

When $x = 0$, we get

$$f^{(k)}(0) = \sum_{n=k}^{\infty} n(n-1)\dots(n-k+1)c_n 0^{n-k} = k!\ c_k$$

□

Theorem 4.4.6. *Suppose* $\sum_{n=0}^{\infty} c_n$ *converges. Define*

$$f(x) = \sum c_n x^n, \quad -1 < x < 1$$

Then,

$$\lim_{x\to 1} f(x) = \sum_{n=0}^{\infty} c_n$$

Proof. Let $s_n = c_0 + c_1 + \cdots + c_n$ and $s_{-1} = 0$. Suppose $\sum c_n$ converges, then s_n converges, say $s_n \to s$. Let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that $\forall n > N,\ |s - s_n| < \frac{\varepsilon}{2}$.

Also we have,

$$\begin{aligned}
\sum_{n=0}^{m} c_n x^n &= \sum_{n=0}^{m} (s_n - s_{n-1}) x^n \\
&= \sum_{n=0}^{m} s_n x^n - x s_{n-1} x^{n-1} \\
&= \sum_{n=0}^{m} s_n x^n - x \sum_{n=0}^{m} s_{n-1} x^{n-1} \\
&= \sum_{n=0}^{m} s_n x^n - x \sum_{n=-1}^{m-1} s_n x^n \\
&= \sum_{n=0}^{m-1} s_n x^n + s_m x^m - x \sum_{n=0}^{m-1} s_n x^n \text{ since } s_{-1} = 0 \\
&= (1-x) \sum_{n=0}^{m-1} s_n x^n + s_m x^m \\
\lim_{m\to\infty} \sum_{n=0}^{m} c_n x_n &= (1-x) \lim_{m\to\infty} \sum_{n=0}^{m-1} s_n x^n + \lim_{m\to\infty} s_m x^m \\
f(x) &= (1-x) \sum_{n=0}^{\infty} s_n x^n \text{ since } s_m \to s,\ |x| < 1 \text{ and } x^m \to 0
\end{aligned}$$

We know that,

$$(1-x) \sum_{n=0}^{m} x^n = (1-x)(1 + x + x^2 + \cdots + x^m) = 1 - x^{m+1}$$

Thus,

$$\lim_{m\to\infty} (1-x) \sum_{n=0}^{m} x^n = \lim_{m\to\infty} 1 - x^{m+1} = 1$$

Thus,

$$\begin{aligned}
|f(x) - s| &= \left| (1-x) \sum_{n=0}^{\infty} s_n x^n - s(1-x) \sum_{n=0}^{\infty} x^n \right| \\
&= (1-x) \left| \sum_{n=0}^{\infty} (s_n - s) x^n \right| \\
&\le (1-x) \sum_{n=0}^{\infty} |(s_n - s) x^n| \\
&\le (1-x) \sum_{n=0}^{\infty} |s_n - s|\ |x|^n
\end{aligned}$$

$$\leq (1-x)\sum_{n=0}^{N}|s_n - s|\ |x|^n + \frac{\varepsilon}{2}(1-x)\sum_{n=N+1}^{\infty} x^n$$

Let $1-x<\delta<1$.†[5]

$$\lim_{x\to 1}|f(x)-s| \leq \lim_{\delta\to 0}\delta\sum_{n=0}^{N}|s_n - s|(1-\delta)^n + \frac{\varepsilon}{2} = 0$$

Therefore,

$$f(1) = \lim_{x\to 1} f(x) = s = \sum_{n=0}^{\infty} c_n x^n$$

□

Theorem 4.4.7. *Given a double sequence $\{a_{ij}\}$. Suppose that,*

$$\sum_{j=1}^{\infty}|a_{ij}| = b_i$$

and $\sum b_i$ converges. Then,

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}$$

Proof. **Step 1 : Construction of f_i**
Given series $\sum_{j=1}^{\infty}|a_{ij}|$ converges to b_i. Let $E=\{x_0,x_1,x_2,\dots\}$ be a countable set such that $x_n\to x_0$ as $n\to\infty$. Define sequence of functions $\{f_i\}$ on E such that

$$f_i(x_n) = \sum_{j=1}^{n} a_{ij},\ \forall n\in\mathbb{N} \quad \text{and} \quad f_i(x_0) = \sum_{j=1}^{\infty} a_{ij}$$

Clearly, $f_i(x_0) = b_i$ and $f_i(x_n)\to f_i(x_0)$ as $n\to\infty$.

Step 2 : f_i is continuous at x_0
We have, $f_i(x_n)\to f_i(x_0)$. Then †[6],

$$\lim_{x_n\to x_0} f_i(x_n) = \lim_{n\to\infty}\sum_{i=1}^{n} a_{ij} = \sum_{i=1}^{\infty} a_{ij} = f_i(x_0)$$

Therefore, function f_i is continuous at x_0.

[5] In earlier version, the term $|s_n - s|$ was neglected. Corrected as per the seminar by Haripriya

[6] The method of contradiction was an unnecessary complication. Corrected as per the seminar by Mekha

Step 3 : Construction of g
Given $|f_i(x)| < b_i$ and $\sum b_i$ converges. Thus, $\sum f_i(x)$ converges. Define $g : E \to \mathbb{R}$ such that

$$g(x) = \sum_{i=1}^{\infty} f_i(x)$$

Since f_i are continuous functions defined on a countable set E, the convergence is uniform and the sum g is continuous. And we have,

$$\lim_{n\to\infty} \lim_{m\to\infty} \sum_{i=1}^{m} f_i(x_n) = \lim_{x_n\to x_0} g(x_n) = g(x_0) = \lim_{m\to\infty} \sum_{i=1}^{m} f_i(x_0) = \lim_{m\to\infty} \lim_{n\to\infty} \sum_{i=1}^{m} f_i(x_n)$$

Therefore,

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}$$

□

Theorem 4.4.8 (Taylor)**.** *Suppose $f(x) = \sum c_n x^n$, the series converging in $|x| < R$. If $-R < a < R$, then f can be expanded in a power series about the point $x = a$ which converges in $|x-a| < R - |a|$. And,*

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n, \quad |x-a| < R - |a|$$

Proof. Suppose

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} c_n[(x-a)+a]^n \\ &= \sum_{n=0}^{\infty} c_n \sum_{m=0}^{n} \binom{n}{m} a^{n-m}(x-a)^m \end{aligned}$$

Changing the order of summation, we may combine coefficients of $(x-a)^m$.

$$= \sum_{m=0}^{\infty}\sum_{n=m}^{\infty} \binom{n}{m} c_n a^{n-m}(x-a)^m$$

Therefore, it is enough to prove that the order of summation can be changed. We know that the order of summation can be changed if,

$$\sum_{n=0}^{\infty}\sum_{m=0}^{n} \left| c_n \binom{n}{m} a^{n-m}(x-a)^m \right| \text{ converges}$$

We know that,

$$\sum_{n=0}^{\infty}\sum_{m=0}^{n} \left| c_n \binom{n}{m} a^{n-m}(x-a)^m \right| = \sum_{n=0}^{\infty} |c_n| \ (|x-a| + |a|)^n$$

and it converges if $|x-a| + |a| < R - |a|$.

Also we know that, if $f(x) = \sum c_n x^n$ converges in $|x| < R$, then the convergence is uniform in $[-R+\varepsilon, R-\varepsilon]$ and it is differentiable in $(-R, R)$. And the derivatives are given by,

$$f^{(m)}(a) = \sum_{n=m}^{\infty} \frac{n!}{(n-m)!} c_n a^{n-m} = m! \sum_{n=m}^{\infty} \binom{n}{m} c_n a^{n-m}$$

Thus,

$$f(x) = \sum_{m=0}^{\infty} \sum_{n=m}^{\infty} c_n \binom{n}{m} a^{n-m} (x-a)^m = \sum_{m=0}^{\infty} \frac{f^{(m)}(a)}{m!} (x-a)^m$$

□

Theorem 4.4.9. *Suppose the series $\sum a_n x^n$ and $\sum b_n x^n$ converge in the segment $S = (-R, R)$. Let E be the set of all $x \in S$ at which*

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n$$

If E has a limit point on S, then $a_n = b_n$.

In other words, If two power series coverges to the same function in $(-R, R)$, then the series are identical.

Proof. —continue— □

Whether the power series representation is unique ?

Theorem 4.4.10. *Let e^x be defined on $\mathbb{R}$. Then*

1. *e^x is continuous and differentiable for all x.*
2. *$(e^x)' = e^x$*
3. *e^x is a strictly increasing function*
4. *$e^{x+y} = e^x e^y$*
5. *$e^x \to +\infty$ as $x \to \infty$ an d $e^x \to 0$ as $x \to \infty$ and $e^x \to 0$ as $x \to -\infty$*
6. *$\lim_{x\to\infty} x^n e^{-x} = 0$ for every n.*

Proof. Let $E(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!}$.

We know that,

$$E(1) = \sum_{k=0}^{\infty} \frac{1^n}{n!} = e \quad \text{and} \quad E(0) = 1$$

And,

$$\begin{aligned}
E(z)E(w) &= \sum_{n=0}^{\infty}\frac{z^n}{n!}\sum_{m=0}^{\infty}\frac{w^m}{m!}\\
&= \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}\frac{z^k w^{n-k}}{k!(n-k)!}\\
&= \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{k=0}^{\infty}\binom{n}{k} z^k w^{n-k}\\
&= \sum_{n=0}^{\infty}\frac{(z+w)^n}{n!}\\
&= E(z+w)
\end{aligned}$$

Then we have,

$$E(z)E(-z) = E(0) = 1$$

We $E(1) = e > 1$. Let M be any integer. Then $E(M) = e^M > M$. Thus there exists $x \in \mathbb{R}$ such that $E(x) > M$. Therefore, $E(x) \to +\infty$ as $x \to +\infty$.

From definition, $E(x) > 0$ for any $x > 0$. And for any $x < 0$, $-x > 0$ and $E(-x) > 0$. Therefore, $E(x) = 1/E(-x) > 0$. Thus, $E(z) > 0$ for any $x \in \mathbb{R}$. Let $\varepsilon > 0$. Then there exists an integer M such that $\varepsilon > 1/M$. We know that there exists $x \in \mathbb{R}$ such that $E(x) > M$. Then, $0 < E(-x) = 1/E(x) < 1/M < \varepsilon$. Thus, $E(x) \to 0$ as $x \to -\infty$.

$$\begin{aligned}
E'(z) &= \lim_{h\to 0}\frac{E(z+h)-E(z)}{h}\\
&= \lim_{h\to 0}\frac{E(z)E(h)-E(z)}{h}\\
&= E(z)\lim_{h\to 0}\frac{E(h)-1}{h}\\
&= E(z)\lim_{h\to 0}(h + \frac{h^2}{2!} + \frac{h^3}{3!} + \dots)/h\\
&= E(z)
\end{aligned}$$

Thus, $E(x)$ is differentiable at every $x \in \mathbb{R}$. Therefore, $E(x)$ is continuous at every $x \in \mathbb{R}$. Thus, the function $E : \mathbb{R} \to (0, \infty)$ is strictly increasing.

By mathematical induction

$$E(z_1 + z_2 + \dots + z_n) = E(z_1)E(z_2)\dots E(z_n)$$

Thus,

$$E(n) = E(1)^n = e^n, \ \forall n \in \mathbb{N}$$

Also we have,

$$E(-1) = \frac{1}{E(1)} = E(1)^{-1} = e^{-1}$$

$$E(-n) = E(-1)^n = \frac{1}{E(n)^n} = E(1)^{-n} = e^{-n}, \ \forall n \in \mathbb{N}$$

Thus,

$$E(n) = e^n, \ \forall n \in \mathbb{Z}$$

Let $p \in \mathbb{Q}$. Then $p = n/m$. And,

$$E(p)^m = E(mp) = E(n) = e^n \implies E(p) = e^p, \ \forall p \in \mathbb{Q}$$

Let $x \in \mathbb{R}$. Define

$$E(x) = \sup_{p<x, \ p \in \mathbb{Q}} E(p)$$

By continuity and monotonicity of E, we have $E(x) = e^x, \ \forall x \in \mathbb{R}$. Now rewriting the results, we get

$$e^{x+y} = e^x e^y \tag{4.1}$$

$$\frac{d}{dx} e^x = e^x \tag{4.2}$$

We have,

$$\begin{aligned} e^x &= \sum_{k=0}^{\infty} \frac{x^k}{k!} \\ &> \frac{x^{n+1}}{(n+1)!} \\ x^n e^{-x} &< \frac{(n+1)!}{x} \\ \lim_{x \to \infty} x^n e^{-x} &< (n+1)! \lim_{x \to \infty} \frac{1}{x} = 0, \ \forall n \in \mathbb{N} \end{aligned}$$

Thus, $x^n e^{-x} \to 0$ as $x \to \infty$. As $x \to +\infty$, $e^x \to +\infty$ faster than any positive power x. This is no surprise since $E(1) = e > 1 = E(0)$. □

Remark. The function $E : \mathbb{R} \to (0, \infty)$ is a bijection. Thus, there exists an inverse function $L : (0, \infty) \to \mathbb{R}$ such that $E \circ L = I_{(0,\infty)}$ and $L \circ E = I_{\mathbb{R}}$ such that

$$L(y) = \int_1^y \frac{1}{x} dx$$

Furthermore, $L(x) \to +\infty$ as $x \to +\infty$ and $L(x) \to 0$ as $x \to -\infty$.

Proof. Let $u, v \in (0, \infty)$. Then there exists $x, y \in \mathbb{R}$ such that $E(x) = u$ and $E(y) = v$. Thus,

$$L(uv) = L(E(x)E(y)) = L(E(x+y)) = x + y$$

Let $y = E(x)$. We have, $L(E(x)) = x$. And $L'(E(x))E'(x) = yL'(y) = 1$ since $E'(x) = E(x) = y$. Thus, $L'(y) = 1/y$.

From fundamental theorem for calculus,we have

$$\int_a^b f(t) \ dt = F(b) - F(a)$$

where $f = F'$ on (a, b). We have $E(x)$ is strictly monotonic and continuous, thus $L(x)$ is monotonic and continuous on $(0, \infty)$. And $L(x)$ is Riemann integrable on any interval $[a, b]$ subset of $(0, \infty)$. Put $F = L$. Then,

$$\int_a^b L'(x)dx = \int_a^b \frac{dx}{x} = L(b) - L(a)$$

We know that $L(1) = 0$. Put $a = 1$ and $b = y$. Then,

$$L(y) = \int_1^y \frac{dx}{x}$$

By mathematical induction,

$$L(x^m) = mL(x)\ \forall m \in \mathbb{N}$$

$$L(x) = L(x^{\frac{m}{m}}) = L(x^{\frac{1}{m}})^m = mL(x^{\frac{1}{m}}) \implies L(x^{\frac{1}{m}}) = \frac{1}{m}L(x)$$

Thus, $x^n = E(nL(x))$ and $x^{\frac{1}{m}} = E(\frac{1}{m}L(x))$. And,

$$x = E(L(x)) = E(L(\frac{1}{m}x)^m) = E(mL(\frac{1}{m}x))$$

Therefore,

$$x^\alpha = E(\alpha L(x)), \quad \forall x \in \mathbb{R},\ x > 0,\ \forall \alpha \in \mathbb{Q}$$

Rewriting the relation with $\log x$ instead of $L(x)$ and e^x instead of $E(x)$, we get $x^\alpha = e^{\alpha \log x}$ for any rational α.

We have,

$$\frac{d}{dx}x^\alpha = \frac{d}{dx}E(\alpha L(x)) = E'(\alpha L(x))\alpha L'(x) = E(\alpha L(x))\frac{\alpha}{x} = \alpha x^{\alpha-1}$$

Suppose $\alpha \neq -1$. For irrational values of x, $L(x)$ can be obtained from,

$$E(\alpha L(x)) = \int_1^x y^\alpha dy$$

We have,

$$\begin{aligned} x^{-\alpha} \log x &= x^{-\alpha} \int_1^x t^{-1}\ dt \\ &< x^{-\alpha} \int_t^x t^{\varepsilon-1}\ dt \\ &< x^{-\alpha} \frac{x^\varepsilon - 1}{\varepsilon} \end{aligned}$$

Clearly,

$$\lim_{x \to +\infty} x^{-\alpha} \log x = 0$$

As $x \to +\infty$, we have $\log x \to +\infty$ slower than any positive power of x. □

Theorem 4.4.11. *1. The function E is periodic with period $2\pi i$.*

2. *The functions C and S are periodic with period $2\pi i$.*

3. *If $0 < t < 2\pi$, then $E(it) \neq 1$.*

4. *If z is a complex number with $|z| = 1$, there is a unique $t \in [0, 2\pi]$ such that $E(it) = z$.*

Proof. Let $E(z) = \sum_{k=0}^{\infty} \frac{z^n}{n!}$. □

Theorem 4.4.12. *Suppose $a_0, a_1, \ldots, a_n$ are complex numbers.*

$$P(z) = \sum_{n=0}^{\infty} a_k z^k$$

Then $P(z) = 0$ for some complex number z.

Proof. Without loss of generality, suppose $a_n = 1$. Put $\mu = \inf |P(z)|$, infimum exists since $|P(z)| > 0$. It is enough to prove that $\mu = 0$.

Suppose $|z| = R$. Then

$$|a_n z^n| = |P(z) - a_0 + a_1 z + a_2 R^2 + \cdots + a_{n-1} R^{n-1}|$$
$$R^n \leq |P(z)| + |a_0| + |a_1| R + \cdots + |a_{n-1} R^{n-1}$$

Rearranging the terms, we get

$$\begin{aligned} |P(z)| &\geq R^n - |a_0| - |a_1| R - |a_2| R^2 - \cdots - |a_{n-1}| R^{n-1} \\ &\geq R^n \left(1 - |a_{n-1}| R^{-1} - |a_{n-2}| R^{-2} - \cdots - |a_0| R^{-n}\right) \end{aligned}$$

Thus $|P(z)| \to \infty$ as $R \to \infty$. Hence, there exists R_0 such that $|P(z)| > \mu$ if $|z| > R_0$. Since $P(z)$ is continuous, $|P(z)|$ is also continuous. And the continuous function $|P(z)|$ on closed disc centered at origin with radius R_0 attains minima, say μ, since closed disc is compact. Thus, $|P(z_0)| = \mu$ for some z_0 such that $|z_0| \leq R_0$.

Suppose $\mu > 0$. Then $P(z_0) \neq 0$. Define $Q(z) = P(z + z_0)/P(z_0)$. Then, Q is a nonconstant polynomial with $Q(0) = 1$ and $|Q(z)| \geq 1$, $\forall z$. Clearly, $Q(z)$ is a polynomial of at most n degrees.

$$Q(z) = b_0 + b_1 z + b_2 z^2 + \cdots + b_n z^n$$

where $b_0 = 1$ since $Q(0) = 1$. Let k be the smallest positive integer such that $b_k \neq 0$. That is, $1 \leq k \leq n$. We have $b_k = |b_k| e^{i\phi}$. Choose θ such that $\phi = \pi - k\theta$. Then $b_k = |b_k| e^{i(\pi - k\theta)} = -|b_k| e^{-ik\theta}$. Then $e^{ik\theta} b_k = -|b_k|$.

Let $z = re^{i\theta}$ such that $r > 0$ and $r^k |b_k| < 1$. Then,

$$|1 + b_k r^k e^{ik\theta}| = 1 - r^k |b_k|$$

Thus,

$$Q(z) = Q(re^{i\theta})| \leq 1 - r^k \left(|b_k| - r|b_{k+1}| - \cdots - r^{n-k} |b_n|\right)$$

For sufficiently small r, $|Q(re^{i\theta})| < 1$ which is a contradiction. Thus $\mu = 0$. Therefore, there exists $|z_0| \leq R_0$ such that $P(z_0) = 0$. □

In other words, complex field $\mathbb{C}$ is algebraically complete.

Proof. Since every polynomial with complex coefficients has a zero, there is no irreducible polynomial of degree greater than or equal to two. Thus, $\mathbb{C}$ doesn't have an algebraic field extension properly containing $\mathbb{C}$. Therefore, $\mathbb{C}$ is complete. $\square$

Subject 5

Graph Theory

Semester II

Subject 6

ME010201 Advanced Abstract Algebra

Module 1 Field Extensions

6.1 Extension Fields §29

Context

- Let R be a commutative ring with unity. If M is a maximal ideal in R, then R/M is a field. [Fraleigh, 2013, §27.9]
- Let F be a field. Every polynomial in $F[x]$ has a unique factorisation into irreducible polynomials except for order and unit. cite[§27.27]fraleigh
- If α is a zero of $f(x) \in F[x]$, then $f(\alpha) = 0$. cite[§22.10]fraleigh
- If $p(x)$ is irreducible over field F, then the principal ideal generated by $p(x)$, denoted by $\langle p(x) \rangle$ is a maximal ideal in $F[x]$. [Fraleigh, 2013, §27.25]
- Let R be a ring with unity. And N be an ideal of R containing a unit. Then $N = R$.[Fraleigh, 2013, §27.5]

Basic Goal Let F be a field and $f(x) \in F[x]$. Find a field E such that F is a subfield of E and there exists a zero of $f(x)$ in E ?

Extension Field Let F be a field. Field E is an extension field of F if F is a subfield of E.
Example : $\mathbb{Q} \leq \mathbb{R} \leq \mathbb{C}$

Tower of Fields A diagrammatic representation emphasising the hierarchy of field extensions in which extension fields appears above their subfields.

Theorem 6.1.1 (Kronecker)**.** *Let F be field. And $f(x)$ be a non-constant polynomial in $F[x]$. Then there exists an extension field E of F and an $\alpha \in E$ such that $f(\alpha) = 0$.*

Proof. Let $f(x) \in F[x]$. Then $f(x)$ has a unique factorisation into irreducible polynomials in $F[x]$ (except for order and unit). Let $p(x)$ be an irreducible factor of $f(x)$. If $f(x)$ is irreducible over F, then $f(x) = cp(x)$. It is enough to construct an extension field E containing both F and α such that $p(\alpha) = 0$.

If $p(x)$ is irreducible over F, then $\langle p(x) \rangle$ is maximal ideal in $F[x]$. Therefore, $F[x]/\langle p(x) \rangle$ is a field, say E.

Consider the function $\psi : F \to F[x]/\langle p(x) \rangle$ defined by $\psi(a) = a + \langle p(x) \rangle$. We claim that $\psi : F \to \psi[F]$ is a field isomorphism. ψ is a canonical homomorphism with trivial kernel. Thus, ψ is one-to-one.

Let $a, b \in F$. And suppose $\psi(a) = \psi(b)$. It is enough to prove that $a = b$. By the definition of ψ, we have $a + \langle p(x) \rangle = b + \langle p(x) \rangle \implies a - b \in \langle p(x) \rangle$. Suppose $a \neq b$. Then $a - b \neq 0$ and $degree(a - b) = 0$. Then, $\langle p(x) \rangle = F[x]$ which is a contradiction since $\langle p(x) \rangle$ is maximal ideal. Therefore, ψ is one-to-one.

We have $p(x)$ is a factor of $f(x)$. Thus $p(\alpha) = 0 \implies f(\alpha) = 0$. Thus, it remains to prove that there exists $\alpha \in F[x]/\langle p(x) \rangle$ such that $p(\alpha) = 0$.

Let $p(x) = a_0 + a_1 x + a_2 x^2 \cdots + a_n x^n$. Consider $\alpha = x + \langle p(x) \rangle$. Then $p(\alpha) = \phi_\alpha(p)$. Thus, $p(\alpha) = a_0 + a_1(x + \langle p(x) \rangle) + \cdots + a_n(x + \langle p(x) \rangle)^n$. Thus, $p(\alpha) = (a_0 + a_1 x + a_n x^n) + \langle p(x) \rangle = p(x) + \langle p(x) \rangle = \langle p(x) \rangle = 0$. Therefore, $p(\alpha) = 0$ and $f(\alpha) = 0$. □

algebraic over F Let $F \leq E$. An element $\alpha \in E$ is algebraic over a field F, if there exists $f(x) \in F[x]$ such that $f(\alpha) = 0$.

transcendental over F Let $F \leq E$. An element $\alpha \in E$ is transcendental over the field F, if it is not algebraic over F.

algebraic number We have, $\mathbb{Q} \leq \mathbb{C}$. A complex number $\alpha \in \mathbb{C}$ is algebraic if it is algebraic over $\mathbb{Q}$. Example : $2, \sqrt{2}, i$

transcendental number A complex number $\alpha \in \mathbb{C}$ is transcendental if it is not an algebraic number. Example : π, e (proof excluded)

Note 1 : A polynomial $f(x) \in F[x]$ is reducible/irreducible depending upon the choice of the field F. For example : $x^2 - 2$ is irreducible over $\mathbb{Q}$, but is reducible over $\mathbb{R}$.

Note 2 : An element $\alpha \in E$ is algebraic/transcendental depending on the choice of the field F. For example : $\sqrt{2} \in \mathbb{C}$ is algebraic over $\mathbb{Q}$ since $x^2 - 2 \in \mathbb{Q}[x]$.

Theorem 6.1.2. *Let E be an extension field of F, and $\alpha \in E$. Let function $\phi_\alpha : F[x] \to E$ be an evaluation homomoprhism. Then α is transcendental if and only if ϕ_α is one-to-one.*

Proof. An element $\alpha \in E$ is transcendental if and only if $f(\alpha) \neq 0$ for any nonzero $f(x) \in F[x]$ where $f(\alpha) = \phi_\alpha(f)$. Thus, kernel of ϕ_α is trivial. That is, $\ker(\phi_\alpha) = \{0\}$. Therefore, ϕ_α is one-to-one. □

Theorem 6.1.3. *Let E be an extension field of F and $\alpha \in E$ be algebraic over F. Then there exists a unique irreducible polynomial $p(x)$ with minimum degree in $F[x]$ and $p(\alpha) = 0$. If there exists a nonzero polynomial $f(x) \in F[x]$ with $f(\alpha) = 0$, then $p(x)$ divides $f(x)$.*

Proof. Consider evaluation homomorphism $\phi_\alpha : F[x] \to E$ defined by $\phi_\alpha(f) = f(\alpha)$. Then $\ker(\phi)$ is an ideal in $F[x]$. Since every ideal in $F[x]$ is principal, there exists $p(x) \in F[x]$ such that $\langle p(x) \rangle = \ker(\phi)$. And $f(\alpha) = 0 \implies f \in \ker(\phi) = \langle p(x) \rangle$. Thus, $p(x)$ divides $f(x)$.

Suppose $p(x) = r(x)s(x)$. Then $p(\alpha) = r(\alpha)s(\alpha) = 0$. However, E is a field and has no zero divisors. Thus, there exists a polynomial of lesser degree in $\langle p(x) \rangle$ which is a contradiction. □

monic polynomial A polynomial which has 1 as the coefficient of highest power of x. For example : $x^3 - 3x \in \mathbb{Q}[x]$

$irr(\alpha, F)$ The unique monic, irreducible polynomial $p(x) \in F[x]$ such that $p(\alpha) = 0$. For example, $irr(\sqrt{3}, \mathbb{Q}) = x^2 - 3$. And $\sqrt{3}$ is the **minimial** polynomial for $\sqrt{3}$ over $\mathbb{Q}$

$deg(\alpha, F)$ The degree of the unique monic, irreducible polynomial $p(x) \in F[x]$ such that $p(\alpha) = 0$. For example, $deg(\sqrt{3}, \mathbb{Q}) = 2$.

simple extension An extension field E of field F is a simple extension if $E = F(\alpha)$ for some $\alpha \in E$.

Theorem 6.1.4. *Let E be a simple extension $F(\alpha)$ of field F where $\alpha \in E$ is algebraic over F. Let $deg(\alpha, F) = n \geq 1$. Then any element $\beta \in E$ can be uniquely expressed in the form $\beta = b_0 + b_1\alpha + \cdots + b_{n-1}\alpha^{n-1}$ where $b_k \in F$.*

Proof. Consider the evaluation homomorphism $\phi_\alpha : F[x] \to E$ defined by $\phi_\alpha(f) = f(\alpha)$. Then $\phi_\alpha[F[x]] = F(\alpha)$.

Let $irr(\alpha, F) = p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$. Then, $p(\alpha) = 0$. Thus,

$$\alpha^n = -a_{n-1}\alpha^{n-1} - a_{n-2}\alpha^{n-2} - \cdots - a_1\alpha - a_0 \tag{6.1}$$

Clearly, any higher power of α can be eliminated from $f(\alpha)$ as shown below,

$$\begin{aligned} \alpha^{n+1} =& \alpha\alpha^n = -a_{n-1}\alpha^n - a_{n-2}\alpha^{n-1} - \cdots - a_1\alpha^2 - a_0\alpha \\ =& a_{n-1}\left(a_{n-1}\alpha^{n-1} + a_{n-2}\alpha^{n-2} + \cdots + a_1\alpha + a_0\right) \\ & - a_{n-2}\alpha^{n-1} - a_{n-3}\alpha^{n-2} - a_1\alpha^2 - a_0\alpha \end{aligned}$$

Thus, in the representation of the element in $F(\alpha)$, the maximum degree of α is $deg(\alpha, F) - 1$. Therefore, $\beta \in F(\alpha) \implies \beta = b_0 + b_1\alpha + b_2\alpha^2 + \cdots + b_{n-1}\alpha^{n-1}$.

We can also show that this representation is unique for any $\beta \in F(\alpha)$. Suppose $\beta = b_0' + b_1'\alpha + b_2'\alpha^2 + \cdots + b_{n-1}'\alpha^{n-1}$. Then $0 = (b_0 - b_0') + (b_1 - b_1')\alpha + \cdots + (b_{n-1} - b_{n-1}')\alpha^{n-1} = \phi_\alpha(g)$ where $g(x) = (b_0 - b_0') + (b_1 - b_1')x + \cdots + (b_{n-1} - b_{n-1}')x^{n-1}$. Clearly, degree of $g(x)$ is $deg(\alpha, F) - 1$ which is less than the minimum degree for a non-zero irreducible polynomial for α over F. Thus $g(x) = 0$. In other words, $b_j = b_j'$, $\forall j$ and representation for $\beta \in F(\alpha)$ is unique. □

6.1.1 Exercises §29

Irreducibility Conditions for Polynomials

1. Irreducibility over a finite field
   x^2+1 is irreducible in $\mathbb{Z}_3$ since for every $x \in \{0,1,2\}$, $x^2+1 \neq 0$.

2. Irreducibility in rational field : Eisentein's Criteria (§23.15)
   Consider $f(x) = x^3 + 60x^2 + 30x + 12$ since for $p = 3$, $f(x)$ satisfies Eisentein's Criteria. And thus is $f(x)$ irreducible over $\mathbb{Q}$.
   Note that for $p = 2, 5, 7, \ldots$ the Eisentein's Criteria is not satisfied.

Algebraic over a Field

1. $\sqrt{2}+\sqrt{3}$ is algebraic over $\mathbb{Q}$ [Fraleigh, 2013, Exercise 29.2]

$$\begin{aligned} \alpha =& \sqrt{2}+\sqrt{3} \\ \implies \alpha^2 =& 2 + 2\sqrt{6} + 3 = 5 + 2\sqrt{6} \\ \implies \alpha^2 - 5 =& 2\sqrt{6} \\ \implies (\alpha^2-5)^2 =& 24 \\ \implies \alpha^4 - 10\alpha^2 + 1 =& 0 \\ \implies \phi_\alpha(x^4 - 10x^2 + 1) =& 0 \end{aligned}$$

   We have, $irr(\sqrt{2}+\sqrt{3}, \mathbb{Q}) = x^4 - 10x^2 + 1$ and $deg(\sqrt{2}+\sqrt{3}, \mathbb{Q}) = 4$.

2. π, e are transcendental numbers. (proof excluded)
   However, π is algebraic over $\mathbb{Q}(\pi)$ since $x - \pi \in \mathbb{Q}(\pi)[x]$.

3. Consider $\alpha = \pi^2$ and $F = \mathbb{Q}(\pi^3)$. [Fraleigh, 2013, Exercise 29.16]

$$\begin{aligned} \alpha^3 =& \pi^6 = (\pi^3)^2 \\ \alpha^3 - (\pi^3)^2 =& 0 \\ \implies \phi_\alpha(x^3 - (\pi^3)^2) =& 0 \end{aligned}$$

   Thus, $irr(\pi^2, \mathbb{Q}(\pi^3)) = x^3 - (\pi^3)^2$ and $deg(\pi^2, \mathbb{Q}(\pi^3)) = 3$.
   Note that $x^3 - (\pi^3)^2 \in \mathbb{Q}(\pi^3)[x]$ since $\pi^3 \in \mathbb{Q}(\pi^3) \implies -(\pi^3)^2 \in \mathbb{Q}(\pi^3)$.
   However, $x^3 - (\pi^3)^2 \notin \mathbb{Q}[x]$.

Factorisation over Extended Field

1. Factorisation over Finite Extension of Finite Field [Fraleigh, 2013, Exercise 29.25]
   Let α be a zero of $f(x) = x^3 + x^2 + 1 \in \mathbb{Z}_2[x]$. Clearly, $f(x)$ is irreducible over $\mathbb{Z}_2$. And $x - \alpha$ is a factor of $f(x)$ in $\mathbb{Z}_2(\alpha)$. We have, $f(x) = (x-a)g(x)$

$$\text{By long division, } g(x) = \frac{x^3+x^2+1}{x-\alpha} = x^2 + (1+\alpha)x + (\alpha+\alpha^2)$$

   Therefore, $x^3 + x^2 + 1 = (x-\alpha)[x^2 + (1+\alpha)x + \alpha(1+\alpha)]$.

The elements of $\mathbb{Z}_2(\alpha)$ are of the form $a_0 + a_1\alpha + a_2\alpha^2$ where $a_0, a_1, a_2 \in \{0, 1\}$. Also we have, α is a zero of $x^3 + x^2 + 1$. Thus, $\alpha^3 = \alpha^2 + 1$.

In order to find a zero of $g(x)$ it is sufficient to evaluate $g(x)$ for all the eight elements $0, 1, \alpha, (1+\alpha), \alpha^2, (1+\alpha^2), (\alpha+\alpha^2), (1+\alpha+\alpha^2) \in \mathbb{Z}_2(\alpha)$. Clearly, $g(1) = \alpha^2 + 2\alpha + 2 = \alpha^2$. And $g(\alpha) = \alpha^2 + 2\alpha(1+\alpha) = \alpha^2$. However, $g(\alpha^2) = \alpha^4 + \alpha^3 + 2\alpha^2 + \alpha = \alpha(\alpha^2+1) + (\alpha^2+1) + 0\alpha^2 + \alpha = \alpha^3 + \alpha^2 + 2\alpha + 1 = \alpha^3 + \alpha^2 + 1 = 0$. Thus, α^2 is a zero of $g(x)$ and $g(x) = (x - \alpha^2)h(x)$.

By long division, $h(x) = \dfrac{x^2 + (1+\alpha)x + (\alpha+\alpha^2)}{x - \alpha^2} = x + (1 + \alpha + \alpha^2)$

Therefore, we have the following linear factoriation for $f(x)$,
$f(x) = (x-\alpha)(x-\alpha^2)(x-1-\alpha-\alpha^2) = (x+\alpha)(x+\alpha^2)(x+1+\alpha+\alpha^2)$ since $-\alpha = 0 - \alpha = 2\alpha - \alpha = \alpha$ in $\mathbb{Z}_2$.

Note : Students should be able to perform long division of polynomials over extended fields.

6.2 Algebraic Extensions §31

$(G : H)$ is the number of H-left cosets in G.

algebraic extension A extension field E of a field F is algebraic if every element in E is algebraic over F.
For example, $\mathbb{C}$ is algebraic over $\mathbb{R}$. But, $\mathbb{R}$ is not algebraic over $\mathbb{Q}$.

finite extension A extension field E of field F is a finite extension if E is a finite dimensional vector space over F. And $[E : F]$ is the **dimension of the vector space** E over F. Again, $[E : F]$ is the **degree of the finite extension** E over F.
For example, $\mathbb{C}$ is a finite extension of degree 2 over $\mathbb{R}$, $[\mathbb{C} : \mathbb{R}] = 2$. But, $\mathbb{R}$ is not a finite extension of $\mathbb{Q}$ and $[\mathbb{R} : \mathbb{Q}]$ is infinite.

Theorem 6.2.1. *Every finite field extensions is an algebraic extension.*

Proof. Let E be a fintie extension of degree n over F. Then $[E : F] = n$. Suppose $\alpha \in E$. Clearly, $\{1, \alpha, \alpha^2, \cdots, \alpha^n\}$ is set of $n+1$ vectors from the vector space E over F. We know that in a vector space of dimension n, any set having $n+1$ vector is linearly dependant. In other words, there exits scalars $c_0, c_1, \cdots, c_n \in F$ (not all zero) such that $c_0 + c_1\alpha + c_2\alpha^2 + \cdots + c_n\alpha^n = 0$.

Clearly, $f(x) = c_0 + c_1x + c_2x^2 + \cdots + c_nx^n$ is a polynomial in $F[x]$ such that $\phi_\alpha(f) = f(\alpha) = 0$. Since $\alpha \in E$ is arbitrary, every element in E is algebraic over F. □

Theorem 6.2.2. *If E is a finite extension of F and K is a finite extension of E. Then K is a finite extension of F. And $[K : F] = [K : E][E : F]$.*

Proof. Let $[E : F] = n$ and $[K : E] = m$. Let $\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ be basis for vector space $E(F)$ and let $\{\beta_1, \beta_2, \cdots, \beta_m\}$ be basis for vector space $K(E)$. We claim that $\{\alpha_i\beta_j : 1 \leq i \leq n,\ 1 \leq j \leq m\}$ is a basis for the vector space $K(F)$.

Let $\gamma \in K$. Then we have $b_1, b_2, \cdots, b_m \in E$ such that $\gamma = b_1\beta_1 + b_2\beta_2 + \cdots + b_m\beta_m$. Again, for each $b_j \in E$, we have $a_{ij} \in F$ such that $b_j = a_{1j}\alpha_1 + a_{2j}\alpha_2 + \cdots + a_{nj}\alpha_n$. Therefore,

$$\gamma = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}\alpha_i\beta_j$$

That is, $\{\alpha_i\beta_j : i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, m\}$ spans K. It remains to prove that $\{\alpha_i\beta_j\}$ is linearly independent. Suppose it is linearly dependent. Then there exists scalars $c_{i,j} \in F$ (not all zero) such that

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_{i,j}\alpha_i\beta_j = \sum_{j=1}^{m}\left(\sum_{i=1}^{n} c_{i,j}\alpha_i\right)\beta_j = 0$$

Let $\sum_{i=1}^{n} c_{i,j}\alpha_i = b_j$. We know that, $\{\beta_j : j = 1, 2, \cdots, m\}$ is linearly independent. Thus, $\sum_j b_j\beta_j = 0 \implies b_j = 0,\ \forall j$.

Again $b_j = 0 \implies \sum_{i=0}^{n} c_{i,j}\alpha_i = 0$. Once again, $\{\alpha_i : i = 1, 2, \cdots, n\}$ is linear independent. Thus, $c_{i,j} = 0,\ \forall i, j$. Thus, $\{\alpha_i\beta_j\}$ is linearly independent. Therefore, $\{\alpha_i\beta_j\}$ is a basis for the vector space $K(F)$. And $[K : F] = |\{\alpha_i\beta_j : i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m\}| = mn$. □

Corollary 6.2.2.1. *Let F_i be fields and F_{i+1} are finite extensions of F_is for $i = 1, 2, \cdots, r$. Then $[F_r : F_1] = [F_r : F_{r-1}][F_{r-1} : F_{r-2}] \cdots [F_2 : F_1]$.*

Proof. We have, F_3 is a finite extension of F_1 and

$$[F_3 : F_1] = [F_3 : F_2][F_2 : F_1] \tag{6.2}$$

Suppose F_k is a finite extension of F_1 and

$$[F_k : F_1] = [F_k : F_{k-1}][F_{k-1} : F_{k-2}] \cdots [F_2 : F_1] \tag{6.3}$$

$$\begin{aligned}[F_{k+1} : F_1] =& [F_{k+1} : F_k][F_k : F_1] \text{ since } F_k \text{ is a finite extension of } F_1\\ =& [F_{k+1} : F_k][F_k : F_{k-1}][F_{k-1} : F_{k-2}] \cdots [F_2 : F_1]\end{aligned}$$

□

Corollary 6.2.2.2. *If E is an extension field of F and $\alpha \in E$ is algebraic over F and $\beta \in F(\alpha)$, then deg(β, F) divides deg(α, F).*

Proof. We have, $\deg(\alpha, F) = [F(\alpha) : F]$ and $\deg(\beta, F) = [F(\beta) : F]$. Also given that $\beta \in F(\alpha) \implies F(\beta) \leq F(\alpha)$. Clearly, $F \leq F(\beta) \leq F(\alpha)$. Therefore, $[F(\alpha) : F] = [F(\alpha) : F(\beta)][F(\beta) : F]$. And $[F(\alpha) : F(\beta)] = [F(\alpha) : F]/[F(\beta) : F]$. Clearly, $[F(\beta) : F]$ divides $[F(\alpha) : F]$. □

Theorem 6.2.3 (algebraic closure)**.** *Let E be an extension field of F. Then $\bar{F}_E = \{\alpha \in E : \alpha$ is algebraic over $E\}$ is a subfield of E.*

Proof. Let $\alpha, \beta \in \bar{F}_E$. Then $\alpha, \beta \in E$ are algebraic over F. And $F(\alpha, \beta)$ is a finite extension field of F. Thus every element in $F(\alpha, \beta)$ are algebraic over F. Thus, $\alpha + \beta, \alpha\beta, \alpha - \beta, \alpha/\beta \in \bar{F}_E$. Therefore, $\bar{F}_E$ is a subfield of E. $\square$

Corollary 6.2.3.1. *The set of all algebraic numbers forms a field.*

Proof. Let α be an algebraic number. Then $\alpha \in \mathbb{C}$ and α is algebraic over $\mathbb{Q}$. Clearly, the set of all algebraic numbers, $\bar{\mathbb{Q}}_{\mathbb{C}}$ is a subfield of $\mathbb{C}$. $\square$

algebraic closure Let F be a field and E be an extension field of F. Then the (smallest) field containing all elements of E which are algebraic over F is the algebraic closure $\bar{F}_E$ of F in E.

algebraically closed Let F be a field. F is algebraically closed if every non-constant polynomial in $F[x]$ has a zero in $F[x]$.

Note : Let F be algebraically closed. Then every irreducible polynomial in $F[x]$ are linear since every non-constant polynomial has a linear factor.

Theorem 6.2.4. *A field F is algebraically closed if and only if every non-constant polynomial $f(x)$ can be factorised in $F[x]$ into linear factors.*

Proof. Let F be algebraically closed and $f(x)$ be a non-constant polynomial in $F[x]$. Then $f(x)$ has a zero $\alpha \in F$. Then $x - \alpha$ is a factor of $f(x)$. That is, $f(x) = (x - \alpha)g(x)$. If $g(x) \in F[x]$ is non-constant, then it has a zero in F. Continuing like this, $f(x)$ can be factorised in $F[x]$ into linear factors.

Suppose every non-constant polynomial in $F[x]$ can be factorised into linear factors. Let $f(x)$ be a non-constant polynomial in $F[x]$. Then $f(x)$ has a linear factor $(ax + b) \in F[x]$. Clearly, $-b/a$ is a zero of $f(x)$. $\square$

Theorem 6.2.5. *Algebraically closed field has no proper algebraic extensions.*

Proof. Let E be an algebraic extension field of F. Then if $\alpha \in E$, we have $irr(\alpha, F) = (x - \alpha)$ since F is algebraically closed, every irreducible polynomial in $F[x]$ are linear. Thus $\alpha \in F$. Since $\alpha \in E$ is arbitrary, $F = E$. $\square$

Theorem 6.2.6 (Fundamental Theorem of Algebra)**.** $\mathbb{C}$ *is algebraically closed.*

Proof. Let $f(z) \in \mathbb{C}[z]$. Suppose $f(z)$ has no zeroes in $\mathbb{C}$. Then $1/f(z)$ is an entire function and as $|z| \to \infty$, $|f(z)| \to \infty$. Thus, $\lim_{|z|\to\infty} \frac{1}{|f(z)|} = 0$. And $1/f(z)$ is bounded.

By Liouville's theorem, every bounded entire function is constant. Therefore, $1/f(z)$ is constant and $f(z)$ is also constant. Thus, every non-constant polynomial function in $\mathbb{C}[z]$ has a zero in $\mathbb{C}$. $\square$

POSET Partial Ordered Set - A set together with partial order (reflexive, antisymmetric, transitive relation).

For example $(\mathbb{R}, <)$, the set of all real numbers together with less than relation is a poset. However, in a poset it is not necessary that two arbitrary elements are comparable. $(\mathbb{C}, R)$ defined by aRb if $\Re(a) = \Re(b)$ and $Im(a) < Im(b)$ is a poset in which $2 + 3i$ and $3 + 3i$ are not comparable.

chain A subset of a poset in which any two elements are comparable. That is, $x, y \in T \implies x < y$ OR $y < x$.

For example, For above defined poset $(\mathbb{C}, R)$, $T = \{2 + ib \in \mathbb{C} : b \in \mathbb{R}\}$ is a chain in $\mathbb{C}$.

Lemma 6.2.7 (Zorn)**.** *If every chain in a poset S has an upper bound. Then S has at least one maximal element in it.*

Proof. Not required (I think, there is no proof. We just take it as an axiom - always true !. If it is not true for a collection then it is not a set !!) □

Theorem 6.2.8 (Existence of Algebraic Closure)**.** *Every field F has an algebraic closure $\bar{F}$.*

Proof. Not required (as per syllabus) □

6.2.1 Exercise §31

1.

6.3 Geometric Constructions §32

6.3.1 Basic Constructions

Finding Midpoint of a line

Let OA be a line. The line passing through the intersection of circles with center O and A with diameter greater than the length of the line gives a perpendicular line through its mid point (say, perpendicular bisection).

Drawing Perpendicular line through a point

Let OA be a line and B be a point on that line. Find points P, Q on OA which are equidistant from B. Then the perpendicular bisection of PQ is a line perpendicular to OA through B.

Drawing Parallel Line

Let OA be a line. Then the perpendicular line segment of any perpendicular line segment is a line segment parallel to OA.

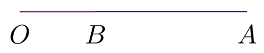
6.3.2 Constructible Numbers

Constructible Number A real number α is constructible if you can draw a line of length $|\alpha|$, given a line of unit length, in finite steps using straight-edge and compass.

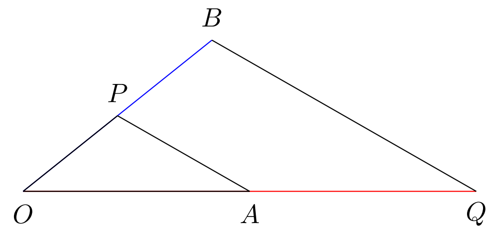
Theorem 6.3.1. *Let α, β be constructible real numbers. Then $\alpha + \beta$, $\alpha - \beta$, $\alpha\beta$, α/β $(\beta \neq 0)$ are also constructible.*

Proof. $\alpha + \beta$ Draw a line OA of length $|\alpha|$ and extend that line using straight edge OE. And draw the line AB of length $|\beta|$ on that extended line, at an point A of the former line extending it. Then OB is a line of length $|\alpha| + |\beta|$.

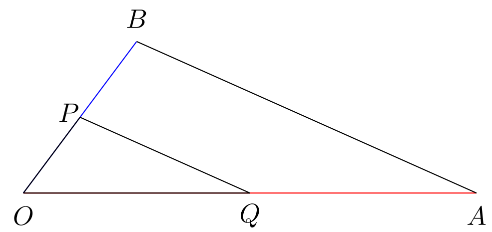
$\alpha - \beta$ Draw a line OA of length $|\alpha|$ and extend that line using straight edge OE. And draw the line AB of length $|\beta|$ on that extended line, at an end point A of former line, but in the opposite direction of extension(towards O). Then the line OB is of length $|\alpha| - |\beta|$.



$\alpha\beta$ Draw two lines OA and OB of length $|\alpha|$ and $|\beta|$ respectively (with a common end point O). Now draw the line of unit length OP along the line OB. Construct triangle OAP. Draw the line BQ parallel to the line PA through B such that OQ and OA are colinear. Now we have two similar triangles OPA and OBQ. Then, the line OQ of length $\|\alpha\beta\|$.



α/β Draw two lines OA and OB of length $\|\alpha\|$ and $\|\beta\|$, with a common end point O. Draw a line of unit length OP along the line OB. Now construct triangle OBA. Draw line PQ parallel to BA so that OQ and OA are colinear. Now the triangles OAB and OQP are similar. And the line OQ has length $\|\alpha/\beta\|$.



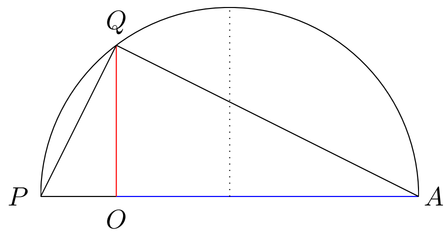
$\square$

Corollary 6.3.1.1. *The set of all constructible real numbers forms a subfield of the field of real numbers.*

Proof. The set of all constructible real numbers say H, contains both 0 and 1. since the line of length zero is trivial and line of unit length is provided. And we have, $\forall \alpha, \beta \in H,\ \alpha+\beta, \alpha-\beta, \alpha\beta, \alpha/\beta \in H$. Since $\alpha - \beta \in H$, $0-\beta = -\beta$ which is the additive inverse of β. And $1/\beta = \beta^{-1}$ is the multiplicative inverse of β. Thus, the set of all constructible numbers is a subfield of $\mathbb{R}$. □

Theorem 6.3.2. *The field of F of constructible numbers consists precisely of all real numbers that we can obtain from $\mathbb{Q}$ by taking square root of positive numbers a finite number of times and applying a finite number of field operations.*

Proof. The constructible numbers are closed under field operations and forms a subfield H of real numbers. We have, $\mathbb{Q}$ is the prime field of $\mathbb{R}$. That is, every subfield of $\mathbb{R}$ contains $\mathbb{Q}$. Thus, the subfield H of constructible numbers contains $\mathbb{Q}$. That is, all the rational numbers are constructible. Therefore, it remains to prove that if $\alpha > 0$ is constructible then $\sqrt{\alpha}$ is also constructible.



Let OA and OP be colinear lines such that OA be a line of length $\|\alpha\|$ and OP be a line of unit length. Find the mid point of PA and draw a circle of diameter PA. Then the length of the perpendicular OQ from PA to the circle is of length $\|\sqrt{\alpha}\|$ since ΔOAQ and ΔOQP are similar triangles.

Therefore, real numbers obtained from rational numbers through finite number of additions/subtractions, multplications/divisions, and square root are constructible. □

Note : For example $\sqrt[4]{5\sqrt[8]{3}-2}$ is constructible, but $\sqrt[6]{2}$ and π are not constructible. We skip the proof that a real number which can't be obtained from rationals by a finite number of these operations is not constructible. And assume that these three operations define the entire field of constructible numbers.

Corollary 6.3.2.1. *Let γ be constructible number which is not rational. Then there exists a sequence of real numbers $\alpha_1, \alpha_2, \ldots, \alpha_n = \gamma$ such that for every $i = 2, \ldots, n$, the extension field $\mathbb{Q}(\alpha_1, \alpha_2, \ldots, \alpha_i)$ is an extension of $\mathbb{Q}(\alpha_1, \alpha_2, \ldots, \alpha_{i-1})$ of degree two.*

In other words, for any constructible number γ, $[\mathbb{Q}(\gamma) : \mathbb{Q}] = 2^n$ for some positive integer n.

Proof. Let γ be a constructible number which can be obtained from rationals by n square root operations and a finite number of field operations. Then, we have a sequence of constructible numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that $[\mathbb{Q}(\alpha_1, \alpha_2, \ldots, \alpha_i) : \mathbb{Q}(\alpha_1, \alpha_2, \ldots, \alpha_{i-1})] = 2$. Therefore, $[\mathbb{Q}(\gamma) : \mathbb{Q}] = 2^n$. □

For example, $\gamma = \sqrt[4]{5\sqrt[8]{3} - 2}$. Then $\alpha_1 = \sqrt{3}$, $\alpha_2 = \sqrt[4]{3}$, $\alpha_3 = \sqrt[8]{3}$, $\alpha_4 = \sqrt{5\sqrt[8]{3} - 2}$ and $\alpha_5 = \gamma$. Clearly, the geometric construction of γ contains five instances of square root operation.

6.3.3 Impossible Problems from Ancient Times

Doubling the cube

Theorem 6.3.3. *There exists a cube such that it is impossible to construct the side of the cube with double the volume.*

Proof. Suppose the cube is of unit side. Then the side of the cube with double the volume is $\gamma = \sqrt[3]{2}$. And $irr(\gamma, \mathbb{Q}) = x^3 - 2$ and $[\mathbb{Q}(\sqrt[3]{2} : \mathbb{Q}] = 3 \neq 2^n$. Therefore, γ is not constructible. □

Squaring the circle

Theorem 6.3.4. *There exists a circle such that it is impossible to construct the side of the square with the same area.*

Proof. Suppose the circle of is of unit radius. Then area of the circle is 2π. And the side of the square with same area is $\sqrt{2\pi}$. Since π is transcendental, π,$\sqrt{\pi}$ and $\sqrt{2\pi}$ are not constructible. □

Trisecting an angle

Theorem 6.3.5. *There exists an angle which can be trisected.*

Proof. We have $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$. Consider $\theta = 20°$. Then $\gamma = \cos\theta$ is a root of the irreducible polynomial $4x^3 - 3x - 0.5$. That is, γ is a root of the monic irreducible polynomial $x^3 - \frac{3}{4}x - \frac{1}{8}$ of degree 3. Thus, γ is not constructible, since $[\mathbb{Q}(\gamma) : \mathbb{Q}] = 3 \neq 2^n$. □

6.3.4 Exercise §32

1.

6.4 Finite Fields §33

6.4.1 Structure of a Finite Field

Theorem 6.4.1. *Let E be a finite extension of degree n over a finite field F. If F has q elements, then E has q^n elements.*

Proof. We have, extension field E is an n-dimensional vector space over the field F. Let $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ be a basis of the vector space E over F. Then every elements of E can be uniquely written as linear combination of basis vectors.

$$\forall \beta \in E,\ \beta = b_1\alpha_1 + b_2\alpha_2 + \cdots + b_n\alpha_n$$

Suppose $\beta \in E$ has two distinct linear combinations. Then, the vectors $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ are not linearly independent, which is a contraditciton.

Since the representation is unique and F has q elements, there are q^n distinct linear combinations possible. Therefore, E has q^n elements. □

Corollary 6.4.1.1. *If E is a finite field of characteristic p. Then F contains exactly p^n elements for some integer n.*

Proof. We have, E is a finite field of characteristic p. Thus, $\mathbb{Z}_p$ is the prime subfield of E. And E is a finite extension of $\mathbb{Z}_p$. Thus, E is a finite dimensional vector space over $\mathbb{Z}_p$. Let n be the dimension of E over $\mathbb{Z}_p$. And $\mathbb{Z}_p$ has p elements. Therefore, E has p^n elements. □

Theorem 6.4.2. *Let E be a field of p^n elements contained in the algebraic closure $\overline{\mathbb{Z}_p}$ of $\mathbb{Z}_p$. Then the elements of E are precisely the zeroes of the polynomial $x^{p^n} - x \in \mathbb{Z}_p[x]$.*

Proof. We have E^* is a multiplicative group of non-zero elements in E. And E^* has $p^n - 1$ elements. Thus, order of any element $\alpha \in E^*$ should divide $p^n - 1$. In other words, if $\alpha \in E^*$, then $\alpha^{p^n-1} = 1$. Clearly, $\alpha^{p^n} - \alpha = 0$, $\forall \alpha \in E$. However, x^{p^n} can have atmost p^n zeroes in $\overline{\mathbb{Z}_p}$. Thus, E is precisely the set of all zeroes of $x^{p^n} - x \in \overline{\mathbb{Z}_p}$. □

nth Root of Unity α is an nth root of unity if $\alpha^n = 1$. ie, $\alpha = \sqrt[n]{1}$

Primitive nth Root of Unity α is a primitive nth root of unity if n is the smallest positive integer such that $\alpha^n = 1$.
That is, $\alpha^n = 1$ and $\forall m \in \mathbb{N},\ m < n \implies \alpha^m \neq 1$

Theorem 6.4.3. *The multiplicative group of non-zero elements of a finite field F is cyclic.*

Proof. Refer : [Fraleigh, 2013, Theorem 23.6] □

Corollary 6.4.3.1. *Finite extension of finite fields are simple extensions.*

Proof. Let E be a finite extension field of the finite field F. Then the multiplicative group of non-zero elements E^* is cyclic. Let α be a generator of the cyclic group E^*. Then, $E = F(\alpha)$. □

6.4.2 Galois Field $GF(p^n)$

Lemma 6.4.4. *If F is a field of prime characteristic p with algebraic closure $\overline{F}$, then $x^{p^n} - x$ has p^n distinct zeroes in $\overline{F}$.*

Proof. We have, $\overline{F}$ is algebraically closed. And $x^{p^n} - x \in \overline{F}[x]$. Thus, $x^{p^n} - x$ can be factorised into p^n linear components. It remains to prove that these factors are distinct.

Clearly, 0 is a zero of multiplicity 1, since $x^{p^n} - x = x(x^{p^n-1} - 1)$. Let $\alpha \neq 0$ be a zero of $x^{p^n} - x$. Then α is a zero of $x^{p^n-1} - 1$. ie, $\alpha^{p^n-1} = 1$.

$$\begin{aligned}(x-\alpha)g(x) =& x^{p^n-1} - 1\\ g(x) =& \frac{x^{p^n-1} - 1}{x-\alpha}\end{aligned}$$

By long division, we get

$$\begin{aligned}g(x) =& x^{p^n-2} + \alpha x^{p^n-3} + \alpha^2 x^{p^n-4} + \cdots + \alpha^{p^n-3}x + \alpha^{p^n-2}\\ g(\alpha) =& (p^n-1)\alpha^{p^n-2}\\ =& (p^n-1)\frac{\alpha^{p^n-1}}{\alpha}\\ =& p^n\frac{1}{\alpha} - \frac{1}{\alpha}\\ =& -\frac{1}{\alpha} \neq 0\end{aligned}$$

Thus, every zero of $x^{p^n-1} - x$ is of multiplicity 1. □

Lemma 6.4.5. *If F is a field of prime characteristic p, then $(\alpha+\beta)^{p^n} = \alpha^{p^n} + \beta^{p^n}$.*

Proof. Since F is a field of characteristic p, for every $\alpha \in F$, $p\alpha = 0$.

For $n = 1$, we have

$$\begin{aligned}(\alpha+\beta)^p =& \alpha^p + p\alpha^{p-1}\beta + \frac{p(p-1)}{2}\alpha^{p-2}\beta + \cdots + p\alpha\beta^{p-1} + \beta^p\\ =& \alpha^p + 0\alpha^{p-1}\beta + 0\alpha^{p-2}\beta + \cdots + 0\alpha\beta^{p-1} + \beta^p\\ =& \alpha^p + \beta^p\end{aligned}$$

Suppose $(\alpha+\beta)^{p^{n-1}} = \alpha^{p^{n-1}} + \beta^{p^{n-1}}$, then

$$\begin{aligned}(\alpha+\beta)^{p^n} =& \left[(\alpha+\beta)^{p^{n-1}}\right]^p\\ =& \left[\alpha^{p^{n-1}} + \beta^{p^{n-1}}\right]^p\\ =& \alpha^{p^n} + \beta^{p^n}\end{aligned}$$

Therefore, by mathematical induction the result is true. □

Theorem 6.4.6 (Existence of Galois Field)**.** *For every prime power p^n, a finite field of p^n elements exists.*

Hint : $\alpha^{p^n} = \alpha \iff \alpha^{p^n} - \alpha = 0 \iff \alpha$ is a zero of $x^{p^n} - x$

Proof. Consider the algebraic closure $\overline{\mathbb{Z}_p}$ of $\mathbb{Z}_p$. Let K be a subset of $\overline{\mathbb{Z}_p}$ containing all zeroes of $x^{p^n} - x \in \overline{\mathbb{Z}_p}$.

Let $\alpha, \beta \in K$. Then $\alpha^{p^n} = \alpha$ and $\beta^{p^n} = \beta$. Since $\overline{\mathbb{Z}_p}$ is a field of characteristic p, we have $(\alpha+\beta)^{p^n} = \alpha^{p^n} + \beta^{p^n} = \alpha + \beta$. Thus, $(\alpha+\beta)$ is a zero of $x^{p^n} - x$. ie, $(\alpha+\beta) \in K$. Clearly, $(\alpha\beta)^{p^n} = \alpha^{p^n}\beta^{p^n} = \alpha\beta$. Thus, $\alpha\beta \in K$.

Again $(-\alpha)^{p^n} = (-1 \cdot \alpha)^{p^n} = (-1)^{p^n} \cdot \alpha^{p^n} = -1 \cdot \alpha = -\alpha$. Thus, $-\alpha \in K$. Also $(\alpha^{-1})^{p^n} = \left(\frac{1}{\alpha}\right)^{p^n} = \frac{1}{\alpha^{p^n}} = \frac{1}{\alpha} = \alpha^{-1}$. Thus, $\alpha^{-1} \in K$.

Trivially, $0, 1 \in K$. Therefore, K is a subfield of $\overline{\mathbb{Z}_p}$ with p^n elements since the p^n zeroes of $x^{p^n} - x$ are distinct. □

Corollary 6.4.6.1. *If F is a finite field, then for every positive integer n, there exists an irreducible polynomial in $F[x]$ of degree n.*

Proof. Let F be a finite field. Then $\mathbb{Z}_p$ is prime field of F for some prime p. And F is of characteristic p and has p^r elements for some positive integer r.

Let K be the subfield of $\overline{F}$ containing precisely all the zeroes of the polynomial $x^{p^{rn}} - x \in \mathbb{Z}_p[x]$. Then, K has a subfield isomorphic to $\mathbb{Z}_p$ since F is of characteristic p and every subfield of F has a subfield isomorphic to $\mathbb{Z}_p$.

By existence theorem of Galois Fields, every element of F is a zero of the polynomial $x^{p^r} - x \in \mathbb{Z}_p[x]$. That is, $\alpha \in F \iff \alpha^{p^r} = \alpha$.

$$\begin{aligned}
\alpha^{p^{rn}} &= \left[\alpha^{p^r}\right]^{p^{r(n-1)}} && = \alpha^{p^{r(n-1)}} \\
&= \left[\alpha^{p^r}\right]^{p^{r(n-2)}} && = \alpha^{p^{r(n-2)}} \\
&\vdots && \vdots \\
&= \left[\alpha^{p^r}\right]^{p^r} && = \alpha^{p^r} = \quad \alpha
\end{aligned}$$

Thus, $\alpha \in F \implies \alpha^{p^{rn}} = \alpha \implies \alpha \in K$. Therefore F is a subfield of K. Clearly, K is a finite extension of the finite field F. And the vector space K over F is n-dimensional, since K has $p^{rn} = [p^r]^n$ elements and F has p^r elements.

Since every finite extension of finite fields are simple extensions, we have $K = F(\beta)$ and $irr(\beta, F) = n$. That is, there exists an unique monic, irreducible polynomial $p(x) \in F[x]$ of degree n such that $p(\beta) = 0$. Therefore, $\forall n \in \mathbb{Z}^+$, there exists an irreducible polynomial in $F[x]$ of degree n. □

Theorem 6.4.7 (Uniqueness of Galois Field)**.** *Let p be a prime and n a positive integer. If E and E' are fields of order p^n, then E and E' are isomorphic.*

*There exists a unique finite field of order p^n, say **Galois Field**, $GF(p^n)$.*

Proof. Let E, E' be fields of order p^n. Then both fields have $\mathbb{Z}_p$ as prime field. Thus, E is a simple extension of $\mathbb{Z}_p$ of degree n. ie, $[E : \mathbb{Z}_p] = n$. And there exits an irreducible polynomial $f(x)$ such that $E \simeq \mathbb{Z}_p[x]/\langle f(x)\rangle$.

Elements of E are zeroes of $x^{p^n} - x$, thus $f(x)$ is a factor of $x^{p^n} - x$. Clearly, elements of E' are zeroes of x^{p^n} and therefore E' has all zeroes of $f(x)$. And

$E' \simeq \mathbb{Z}_p / \langle f(x) \rangle \simeq E$. Therefore, there exists a unique field of order p^n (upto isomorphism), say $GF(p^n)$. $\square$

6.4.3 Exercise §33

1.

Module 2 : Unique Factorisation Domains

6.5 Unique Factorisation Domains §45

Definitions 6.5.1 (divides)**.** An element $a \in R$ divides b if there exists an element $c \in R$ such that $b = ac$.

Definitions 6.5.2 (associate)**.** Two elements a, b are associates if $a = bu$ where u is a unit.

Definitions 6.5.3 (UFD)**.** **Unique factorisation domain**, UFD is an integral domain such that

1. Every element can be factored into a finite number of irreducibles, except 0 and units †[1].

2. The above factorisation is unique except for order and associates.

For example, In $\mathbb{Z}$, $24 = 2 \times 2 \times 2 \times 3 = -2 \times -3 \times 2 \times 2$. Here 2 and -2 are associates. And 2 and 3 are not units, since $2^{-1}, 3^{-1} \notin \mathbb{Z}$.

Definitions 6.5.4 (PID)**.** An integral domain is D is a **Principal Ideal Domain** if every ideal in D is a principal ideal.

There are two important results on UFDs.

1. Every PID is a UFD.

2. If D is a UFD, then $D[x]$ is a UFD.

6.5.1 Every PID is UFD

Lemma 6.5.1. *Let R be a commutative ring and let $N_1 \subset N_2 \subset \dots$ be an ascending chain of ideals in R. Then $N = \cup_i N_i$ is an ideal of R.*

Proof. **Step 1 : N is a subring of R**
Suppose N_i, N_j are two ideals in the chain, $N_i \subset N_j$ and $a \in N_i$, $b \in N_j$. Clearly, $a \in N_j$.

And $a \pm b$, $ab \in N_j$. And $0 \in N_j \implies 0 \in N$. We have, 0 in every ideal†[2]. Take $a = 0$, we know that for every element $b \in N_j$, its additive inverse $-b \in N_j$. Thus $b \in N \implies b \in N_j \implies -b \in N_j \implies -b \in N$. Clearly, N is a subring of R.

Step 2 : N is a ideal of R
Let $a \in N$ and $r \in R$. We have, $a \in N \implies a \in N_J$ for some ideal N_j in the chain. Since N_j is an ideal $ar = ra \in N_j$. Therefore, $ar \in N$. And N is a ideal of R. □

[1]Unit is a element which has multiplicative inverse.
[2]Suppose $b \in N$. We have $-b \in R$ and $-b + b = 0 \in bN \subset N$.

Lemma 6.5.2 (**A**scending **C**hain **C**ondition)**.** *Let D be a PID. If $N_1 \subset N_2 \subset \dots$ is an ascending chain of ideal, then there exists a positive integer r such that $N_r = N_s$ for every $s \geq r$.*

In other words, every strictly[3] *ascending chain of ideals in a PID is of finite length.*

Proof. Let D be an integral domain. And $N_1 \subseteq N_2 \subseteq \dots$ be an ascending chain of ideals in D. Then $N = \cup_i N_i$ is an ideal in D. Since D is a PID, by definition of **p**rincipal **i**deal **d**omain every ideal in it is a principal ideal. Thus, N is a principal ideal in D. That is, there exists $c \in D$ such that $\langle c \rangle = N$.

Since $c \in N$, and $N = \cup_i N_i$ we have $c \in N_r$ for some $r \in \mathbb{N}$. Then $\langle c \rangle = N_r = N$. Again, $N_r \subset N_s \implies c \in N_s \implies \langle c \rangle = N_s = N$.

Suppose the chain of ideals are strictly ascending, that is every ideal in the chain is properly containing the former ideal. Then, the ideal N_r such that $c \in N_r$ is the last ideal in its chain. Thus, striclty ascending chain of ideals is finite. $\square$

Theorem 6.5.3. *Let D be a PID. Every element that is neither* 0 *nor a unit in D is a product of irreducibles.*

Proof. Suppose D is a PID. And suppose $a \in D$ is a neither a zero nor a unit. If a is an irreducible, then the result is trivial. Suppose a is not an irreducible.

Step 1 : a has an irreducible factor

By the definition of irreducibility, a has a factorisation $a = a_1 b_1$ where a_1, b_1 are non-units. And every element $ar \in \langle a \rangle$ can be expressed as $ar = a_1 b_1 r = a_1 r' \in \langle a_1 \rangle$. Thus, the ideal generated by a, $\langle a \rangle$ is contained in the ideal generated by a_1, $\langle a_1 \rangle$. That is, $\langle a \rangle \subset \langle a_1 \rangle$.

If a_1 is an irreducible, then the proof is complete. Suppose a_1 is not an irreducible. That is, $a_1 = a_2 b_2$. Then $\langle a_1 \rangle \subset \langle a_2 \rangle$. Also we have, $\langle a_1 \rangle \neq \langle a_2 \rangle$.

Suppose $\forall d \in D$, $a_2 d \in \langle a_1 \rangle \implies a_2 \cdot 1 = a_2 \in \langle a_1 \rangle \implies a_2 = a_1 c_2 \implies a_1 = (a_1 c_2) b_2 \implies c_2 b_2 = 1 \implies b_2$ has multiplicative inverse c_2. This contradicts the assumption that b_2 is a unit.

Continuing like this, we get a strictly ascending chain of ideals $\langle a \rangle \subsetneq \langle a_1 \rangle \subsetneq \langle a_2 \rangle, \dots$. And we know that, every strictly ascending chain of ideals is finite. Thus, a_r is an irreducible. And $a = a_1 b_1 = (a_2 b_2) b_1 = \dots = a_r (b_r b_{r-1} \dots b_1) = a_r b$. Clearly, a has an irreducible factor a_r.

Step 2 : a is a product of irreducibles

Suppose a is neither zero nor a unit. Then $a = p_1 c_1$ where p_1 is an irreducible and c_1 is not a unit. Continuing like this, we get another striclty ascending chain of ideals $\langle a \rangle \subsetneq \langle c_1 \rangle \subsetneq \langle c_2 \rangle \subsetneq \dots$ such that $a = p_1 c_1 = p_1 (p_2 c_2) = \dots$ where $p_1, p_2, \dots$ are irreducibles and c_j are non-units. Again, by ascending

[3] Strictly ascending chain : $N_1 \subsetneq N_2 \subsetneq N_3 \subsetneq \dots \subsetneq N_k$.

chain condition this strictly ascending chain is finite. That is, $a = p_1p_2 \dots p_kc_k$ for some $k \in \mathbb{N}$. Since $\langle c_k \rangle$ is a maximal ideal in D, c_k is an irreducible say, p_{k+1}. Thus, any element $a \in D$ (except zero and units) can be expressed as product of irreducibles. □

6.5.2 Irreducible element and Maximal Ideal

We have seen earlier that a non-trivial principal ideal $\langle p(x) \rangle$ in $F[x]$ is maximal if and only if the generator $p(x)$ is irreducible over F. Now we have a generalisation, which says ideals in PIDs are maximal if and only if the generator is irreducible. Remember that for any field F, $F[x]$ is a PID.

Lemma 6.5.4. *An ideal $\langle p \rangle$ in a PID is maximal if and only if p is an irreducible.*

Proof. **Part A : $\langle p \rangle$ maximal in $D \implies p$ irreducible in D**
Let D be a PID. Suppose $\langle p \rangle$ is a maximal ideal in D. Suppose $p = ab$ in D. That is, $p = ab$ where $a, b \in D$. Then, $\langle p \rangle \subset \langle a \rangle$.

Suppose $\langle p \rangle = \langle a \rangle$. Then $a = pc = (ab)c \implies bc = 1$ and b is a unit.

Suppose $\langle p \rangle \neq \langle a \rangle$. Then $\langle a \rangle = D$, since $\langle p \rangle$ is a maximal ideal. That is $\langle a \rangle = \langle 1 \rangle \implies a, 1$ are associates. And a is a unit.

Thus for any factorisation $p = ab$, either a or b is a unit. Therefore, p is an irreducible.

Part B : p is irreducible in $D \implies \langle p \rangle$ is maximal in D
Suppose p is an irreducible in D. Then $p = ab$ where $a, b \in D$ implies that either a or b is a unit. Since $p = ab$, $\langle p \rangle \subset \langle a \rangle$. We know that, if a is a unit then $\langle a \rangle = \langle 1 \rangle = D$.

Suppose a is not a unit. Then b is a unit. That is, there exists $u \in D$ such that $bu = 1$. Then $pu = abu = a$. Thus, $\langle a \rangle \subset \langle p \rangle \implies \langle p \rangle = \langle a \rangle$.

Thus, for any factorisation $p = ab$, the ideals generated by a, b are either D or $\langle p \rangle$ itself. In other words, there doesn't exist a proper ideal containing $\langle p \rangle$. Therefore, $\langle p \rangle$ is a maximal ideal in D. □

6.5.3 Every PID is UFD

We have seen earlier that $F[x]$ has unique factorisation of all its elements. Now, we generalise that result into "every PID has unique factorisation for all its elements".

Lemma 6.5.5. *In a PID, if an irreducible p divides ab, then either $p|a$ or $p|b$.*

Proof. Let D be a PID. Suppose $p \in D$ is irreducible. Suppose $p|ab$. Then $ab \in \langle p \rangle$. And we have, $\langle p \rangle$ is a maximal ideal. However, every maximal ideal in D is a prime ideal. Thus, $ab \in \langle p \rangle \implies a \in \langle p \rangle$ or $b \in \langle p \rangle$. In other words, $p|a$ or $p|b$. □

Corollary 6.5.5.1. *If p is an irreducible in a PID D and p divides $a_1a_2\dots a_n$ where $a_i \in D$, then $p|a_i$ for at least one i.*

Proof. We know that, $p|a_1a_2 \implies p|a_1$ or $p|a_2$.
Suppose $p|a_1a_2\dots a_k \implies p|a_1$ or $p|a_2\dots$ or $p|a_k$.
Suppose $p|a_1a_2\dots a_{k+1} \implies p|a_1a_2\dots a_k$ or $p|a_{k+1}$.
$\implies p|a_1$ or $p|a_2$ or $\dots p|a_k$ or $p|a_{k+1}$.
That is, if p divides a finite product, then it divides at least one of them. □

Definitions 6.5.5. A nonzero, nonunit element p in an integral domain D is a prime if $\forall a, b \in D,\ p|ab \implies p|a$ or $p|b$.

Theorem 6.5.6. *Every PID is a UFD.*

Proof. Let D be a PID. Let $a \in D$ is neither a zero nor a unit. Then a has a factorisation into irreducibles $a = p_1p_2\dots p_r$. Suppose a has another factorisation $a = q_1q_2\dots q_s$.

We have $p_1|q_1q_2\dots q_s$. Thus, p_1 divides at least one of them, say q_j. Without loss of generality, we assume that $p_1|q_1$ (rearranging the terms if necessary).

Since q_1 is an irreducible, $q_1 = p_1u_1 \implies u_1$ is a unit. Thus, $p_1p_2\dots p_r = p_1u_1q_2q_3\dots q_s$. That is, $p_2p_3\dots p_r = u_1q_2q_3\dots q_s$.

Continuing like this we get, $p_r = u_1u_2\dots u_{r-1}p_ru_rq_{r+1}q_{r+s}\dots q_s$. Thus $u_1u_2\dots u_rq_{r+1}q_{r+2}\dots q_s$ is a unit. We have $s = r$, otherwise $1 = u_1u_2\dots q_{r+1}\dots$ is a contradiction. Therefore, any element $a \in D$ has a unique factorisation except for order,associates and units. □

Corollary 6.5.6.1. *The integral domain $\mathbb{Z}$ is a UFD.*

Proof. The ideals in $\mathbb{Z}$ are of the form $n\mathbb{Z}$ where $n \in \mathbb{Z}$. Thus every ideal in $\mathbb{Z}$ are pricipal ideals. Therefore, $\mathbb{Z}$ is a PID. We know that every PID is a UFD. Thus, $\mathbb{Z}$ is a **u**nique **f**actorisation **d**omain. □

6.5.4 D is UFD $\implies D[x]$ is UFD

Definitions 6.5.6 (gcd)**.** Let D be a UFD and $a_1, a_2, \dots, a_n$ be nonzero elements in D. An element $d \in D$ is a **g**reatest **c**ommon **d**ivisor of a_i if d is a common divisor of all a_is and also divides any common divisor of a_is.

Definitions 6.5.7 (primitive)**.** Let D be a UFD. A nonconstant polynomial $a_0 + a_1x + \dots + a_nx^n \in D[x]$ is a primitive if 1 is the gcd of all a_is.

For exmaple, $4x^3 + 3x^2 + 2 \in \mathbb{Z}[x]$ is a primitive since $\gcd(4, 3, 2) = 1$.

Lemma 6.5.7. *If D is a UFD, then for every nonconstant $f(x) \in D[x]$ we have $f(x) = cg(x)$ where $c \in D$ and $g(x) \in D[x]$ and $g(x)$ is primitive. The element $c \in D$ is unique upto a unit factor in D and is the* ***content*** *of $f(x)$. Also $g(x)$ is unique upto a unit factor in D.*

Proof. **Step 1 : There exists $c \in D$ such that $f(x) = cg(x)$**
Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ be a non-constant polynomial in $D[x]$. Let $c = gcd(a_0, a_1, \ldots, a_n)$. In other words, c is the greatest common divisor of the coefficients of $f(x)$. Let $q_i = ca_i)$. Clearly, $gcd(q_1, q_2, \ldots, q_n) = 1$. Thus, we have $f(x) = cg(x)$ where $g(x) = q_0 + q_1x + \cdots + q_nx^n)$ is a primitive.

Step 2 : There exists unique c such that $f(x) = cg(x)$
Suppose $f(x) = cg(x) = dh(x)$ where $c, d \in D$ and $g(x), h(x)$ are primitives.

Clearly, c and d are associates. Otherwise, there exists an irreducible element p such that $p|cg(x)$ but $p \not| \, dh(x)$ which is a contradiction. Thus, we can cancel all irreducible factors of c to obtain, $ug(x) = vh(x)$ where u and v are units. Therefore, the content of $f(x)$, c unique upto associates.

Again, $f(x) = cg(x) = (cu)(u^{-1}g(x))$ and thus $g(x)$ is unique upto associates/units. □

Definitions 6.5.8 (content)**.** Let D be a UFD. Let $f(x) \in D[x]$. Then $f(x) = cg(x)$ for some primitive $g(x) \in D[x]$ and $c \in D$. The element c which is unique upto units, is the content of $f(x)$.

For example, Let $f(x) = 8x^3 + 6x^2 + 4 \in \mathbb{Z}[x]$. Then $f(x) = 2(4x^3 + 3x^2 + 2)$ where $4x^3 + 3x^2 + 2 \in \mathbb{Z}[x]$ is a primitive and $2 \in \mathbb{Z}$. Thus, $content(f) = 2$.

Lemma 6.5.8 (Gauss)**.** *If D is a UFD, then product of two primitive polynomials in $D[x]$ is again primitive.*

Proof. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$. Suppose $f(x)$ and $g(x)$ are primitives. Let $h(x) = f(x)g(x) = c_0 + c_1x + \cdots + c_{n+m}x^{n+m}$. Let p be an irreducible in D. Then $f(x)$ has a coefficient which is not divisible by p. Suppose p divides every coefficient $f(x)$, then $f(x)$ is not a primitive.

Let r be the smallest integer such that $p \not| a_r$. Similarly, let s be the smallest integer such that $p \not| \, b_s$. Now consider the coefficient c_{r+s} of $h(x)$.

$$c_{r+s} = a_0b_{r+s} + \cdots + a_{r-1}b_{s+1} + a_rb_s + a_{r+1}b_{s-1} + \cdots + a_{r+s}b_0$$

By the selection of r, $p|a_j$ for every $j < r$. Thus, $p|a_0b_s + \cdots + a_{r-1}b_{s+1}$. Similarly, by the selection of s, $p|b_k$ for every $k < s$. Thus, $p|a_{r+1}b_{s-1} + \cdots + a_0b_s$. Clearly, $p|c_{r+s} \iff p|a_rb_s \implies p|a_r$ or $p|b_s$. This is not possible, thus $p \not| c_{r+s}$. Thus, for any irreducible $p \in D$, $h(x)$ has a coefficient which is not divisible by p. Therefore, product of two primitives is always a primitive. □

Corollary 6.5.8.1. *If D is a UFD, then a finite product of primitive polynomials in $D[x]$ is again primitive.*

Proof. Let $\{f_k\}_{k=1}^n$ be a family of primitives. Then $f_1(x)f_2(x)$ is a primitive, since product of two primitives is always a primitive.

Suppose $g(x) = f_1(x)f_2(x) \ldots f_k(x)$ is a primitive.

Consider $h(x) = f_1(x)f_2(x)\dots f_{k+1}(x) = g(x)f_{k+1}(x)$. Since both $g(x)$ and $f_{k+1}(x)$ are primitives their product $h(x)$ is also a primitive.

Thus by finite mathematical induction, every finite product of primitives is a primitive. □

Lemma 6.5.9. *Let D be a UFD and let F be a field of quotients of D. Let $f(x) \in D[x]$ where $\deg f(x) > 0$. If $f(x)$ is an irreducible in $D[x]$, then $f(x)$ is also an irreducible in $F[x]$. Also, if $f(x)$ is primitive in $D[x]$ and irreducible in $F[x]$, then $f(x)$ is irreducible in $D[x]$.*

Proof. Let D be a UFD and F be the field of quotients of D. Then elements of F are of the form $(a, b) \in D \times D$. Also $(a, b) + (c, d) = (ad + bc, bd)$ and $(a, b) \cdot (c, d) = (ac, bd)$. And we say express (a, b) as a/b.

Part 1 : irreducible in $D[x] \implies$ irreducible in $F[x]$
Let $f(x) \in D[x]$ be an irreducible in $D[x]$. Suppose $f(x)$ is not irreducible in $F[x]$. Then $f(x) = r(x)s(x)$ where $r(x), s(x) \in F[x]$.

The coefficients of $r(x)$ are of the form a/b. By multiplying $r(x)$ with the product of all denominators, we can remove its denominators and obtain an associate $r_1(x) \in D[x]$. Similarly, we have $s(x) = vs_1(x)$ where $s_1(x) \in D[x]$. Thus, $f(x) = r(x)s(x) = ur_1(x)vs_1(x) = uvr_1(x)s_1(x)$. Therefore, $f(x) \in D[x]$ is not an irreducible, which is a contradiction. Thus, $f(x)$ is irreducible in the quotient field F as well.

Part 2 : irreducible in $F[x] \implies$ irreducible in $D[x]$
Suppose $f(x) \in D[x]$ is irreducible in $F[x]$. Clearly, F contains a UFD isomorphic to D. Thus $D[x] \leq F[x]$. Therefore, any polynomial irreducible in $F[x]$ is also irreducible in $D[x]$. □

We just saw that if F is the quotient field of a UFD D and $f(x) \in D[x]$. The irreducibility of $f(x)$ in $F[x]$ is necessary and sufficient for the irreducibility of $f(x)$ in $D[x]$. Now we prove that the factorisation is also unique upto a morphism.

Corollary 6.5.9.1. *If D is a UFD and F is a field of quotients of D, then a nonconstant polynomial $f(x) \in D[x]$ factors into a product of two polynomials of lower degrees r and s in $F[x]$ if and only if it has a factorization into polynomials of the same degress r and s.*

Proof. We know that if $f(x) \in D[x]$ has a factorisation $f(x) = r(x)s(x)$ in $F[x]$. Then it has a factorisation in $D[x]$ as well, $f(x) = cr_1(x)s_1(x)$ where $r_1(x), s_1(x)$ are associates of $r(x), s(x)$ in $F[x]$. Clearly, the associates are always of the same degree.

Trivially, any factorisation in $D[x]$ will also be a factorisation in $F[x]$. And the factors will have the same degree, unless $D[x]$ has an irreducible polymonial which is not irreducible in $F[x]$. This is not possible. □

Theorem 6.5.10. *If D is a UFD, then $D[x]$ is a UFD.*

Proof. Let $f(x) \in D[x]$. Suppose $deg(f) > 0$. Otherwise, $f(x)$ is a constant and we have trival factorisation.

Let $f(x) = g_1(x)g_2(x)\dots g_r(x)$ be a factorisation in $D[x]$ with maximum number of factors. Then $f(x) = c_1h_1(x)c_2h_2(x)\dots c_rh_r(x)$ where $g_k(x) = c_kh_k(x)$ and $h_k(x)$ are primitives. And $h_k(x)$ are irreducibles. Otherwise, $f(x)$ has a factorisation with greater number of factors than r which is a contradiction. Thus, $f(x) = uh_1(x)h_2(x)\dots h_r(x)$ is a factorisation of $f(x)$ into a product of irreducibles.

Suppose $f(x)$ has another factorisation $f(x) = G_1(x)G_2(x)\dots G_r(x)$. Then $f(x) = vH_1(x)H_2(x)\dots H_r(x)$ with $H_k(x)$ are all irreducibles. We know that, $h_k(x)|f(x) \implies h_k(x)|H_1(x)H_2(x)\dots H_r(x) \implies h_k(x)|H_k(x)$ (WLOG). Therefore, the factorisation is unique. □

For example, let x, y be two indeterminates, then an element $f(x) \in F[x, y]$ is of the form $\sum_{i,j=0}^{n,m} a_{ij}x^iy^j$.

Corollary 6.5.10.1. *If F is a field of quotients and $x_1, x_2, \dots, x_n$ are indeterminates, then $F[x_1, x_2, \dots, x_n]$ is a UFD.*

Proof. Suppose D is a UFD and F is its field of quotients, then $F[x]$ is a UFD.

Suppose $K = F[x_1, x_2, \dots, x_k]$ is a UFD.

Then, $F[x_1, x_2, \dots, x_k, x_{k+1}] = F[x_1, x_2, \dots, x_k][x_{k+1}] = K[x_{k+1}] = K[x]$ is a UFD. □

6.5.5 Exercise §45

1.

6.6 Euclidean Domain §46

Definitions 6.6.1 (Euclidean Norm)**.** Euclidean norm on an integral domain D is a function v which maps non-zero elements of D into non-negative integers such that

1. $\forall a, b \in D, (b \neq 0)$ there exists $q, r \in D$ such that $a = qb + r$ where either $r = 0$ or $v(r) < v(b)$ and
2. $\forall a, b \in D, (a, b \neq 0)$ $v(a) \leq v(ab)$

Definitions 6.6.2 (Euclidean Domain)**.** An integral domain D is a Euclidean Domain if there exists a Euclidean Norm in D.

For example, $v(n) = |n|$ is a Euclidean norm on Euclidean domain $\mathbb{Z}$. And $v(f(x)) = deg(f(x))$ is a Euclidean norm on Euclidean domain $F[x]$.

Theorem 6.6.1. *Every Euclidean domain is a PID.*

Proof. Let D be a Euclidean domain with Euclidean norm v. Let N be an ideal in D. If $N = \{0\}$, then $N = \langle 0 \rangle$ is a principal ideal.

Suppose $N \neq \{0\}$. Then there exists $b \in N$ $(b \neq 0)$. Clearly, $b \in N \implies \langle b \rangle \subset N$. Choose a $b \in N$ such that $v(b)$ is minimal in N.

Claim : $N = \langle b \rangle$
Let $a \in N$. Then there exists $q, r \in D$ such that $a = qb + r$ where $r = 0$ or $v(r) < v(b)$. Clearly, $r = a - qb \in N$. Thus, $r = 0$ since $v(r) < v(b)$ and $v(b)$ is minimal in N. That is, $a = qb$ or $b|a$ for every $a \in N$. Thus, $N \subset \langle b \rangle$. Therefore, $N = \langle b \rangle$. In other words, evey ideal in D is a principal ideal. □

Corollary 6.6.1.1. *Every Euclidean domain is a UFD.*

Proof. We know that, every Euclidean domain is a PID. And every PID is a UFD. Thus, every Euclidean domain is a UFD. □

6.6.1 Arithmetic in Euclidean Domain

Theorem 6.6.2. *For a Euclidean domain with Euclidean norm v, $v(1)$ is minimal and $u \in D$ is a unit if and only if $v(u) = v(1)$.*

Proof. We have, $v(a) \leq v(ab)$. Take $a = 1$, $v(1) \leq v(1b) \implies v(1) \leq v(b)$ for every $b \in D$. Thus, $v(1)$ is minimal.

Suppose u is a unit with multplicative inverse u^{-1}. Clearly, $v(1) \leq v(u)$ since $u \in D$. Take $a = u$ and $b = u^{-1}$. Then $v(u) \leq v(uu^{-1}) = v(1)$. Therefore, for every unit $u \in D$, we have $v(u) = v(1)$. □

Theorem 6.6.3 (division algorithm)**.** *Let D be a Euclidean domain with Euclidean norm v. And $a, b \in D$ $(a, b \neq 0)$.*

$$a = bq_1 + r_1, \text{ where } r_1 = 0 \text{ or } v(r_1) < v(b)$$

if $r_1 \neq 0$

$$b = r_1 q + r_2, \text{ where } r_2 = 0 \text{ or } v(r_2) < v(r_1)$$

In general, if $r_i \neq 0$

$$r_{i-1} = r_i q + r_{i+1} \text{ where } r_{i+1} = 0 \text{ or } v(r_{i+1}) < v(r_i)$$

Then the sequence $r_1, r_2, \ldots$ must terminate with some $r_s = 0$.

If $r_1 = 0$, then $gcd(a, b) = b$.
If $r_k = 0$, then $gcd(a, b) = r_{k-1}$.

Furthermore, if d is a gcd of a and b, then there exists $\lambda, \mu \in D$ such that $d = \lambda a + \mu b$.

Proof. **Step 1 : Finite Sequence** $r_1, r_2, \ldots, r_s$
Suppose $r_1, r_2, \ldots, r_{i-1} \neq 0$. Then $v(r_1) > v(r_2) > \cdots > v(r_{i-1})$ is a strictly decreasing sequence of non-negative integers. Thus, the sequence will terminate in finite numbers of steps. That is, $r_s = 0$ for some integer s.

Step 2 : $gcd(a, b) = r_{s-1}$
Suppose $r_1 = 0$. Then $a = bq$ and $gcd(a, b) = b$ since $b|a, b|b$ and there does not exist an integer greater than b that divides b.

Suppose $r_1 \neq 0$ and $gcd(a, b) = d$. Then $r_1 = a - bq$ and $d|r_1$.
And $d|b, d|r_1 \implies d|a$ since $a = qb + r_1$. Thus, $gcd(a, b) = gcd(b, r_1)$.
Continuing like this, we get $gcd(a, b) = gcd(r_{s-2}, r_{s-1})$ if $r_{s-1} \neq 0$.
And we have, $r_s = 0 \implies r_{s-2} = qr_{s-1}$.
Clearly, $gcd(r_{s-2}, r_{s-1}) = r_{s-1} = gcd(a, b)$.

Step 3 : $gcd(a, b) = \lambda a + \mu b$
Let b be a $gcd(a, b)$. Then, $b = 0a + 1b$ and result is complete.

Suppose $gcd(a, b) = r_{s-1}$. And $r_{s-1} = r_{s-3} - qr_{s-2}$. Clearly, $r_i = \lambda_i r_{i-2} + \mu_i r_{i-1}$. Keep on substituting using these equations until we reach $r_{s-1} = \lambda a + \mu b$.

Suppose d' is another $gcd(a, b)$. Then $d' = ud$ is an associate of d. And $d' = ud = (\lambda u)a + (\mu u)b$. And the result is true for any $gcd(a, b)$. □

6.7 Gaussian Integers and Multiplicative Norms §47

6.7.1 Gaussian Integers and Norm

Definitions 6.7.1 (Gaussian integers)**.** Gaussian integers are complex numbers of the form $a + ib$ where $a, b \in \mathbb{Z}$.

Note : Let $\alpha = a + ib$. Clearly, Gaussian integers are elements of $\mathbb{Z}[i]$. Here, $\mathbb{Z}[i]$ is a simple extension of the integral domain $\mathbb{Z}$ with a zero of $x^2 + 1$.

Definitions 6.7.2 (Gaussian Norm)**.** Let $\alpha = a + ib$ be a Gaussian integer. Then $N(\alpha) = |\alpha|^2 = a^2 + b^2$ is a Euclidean norm on $\mathbb{Z}[i]$.

Lemma 6.7.1. *The Gaussian norm* $N : \mathbb{Z}[i] \to \mathbb{Z}$ *defined by* $N(a+ib) = a^2+b^2$ *has the following properties*

1. $N(\alpha) \geq 0, \ \forall \alpha \in \mathbb{Z}[i]$
2. $N(\alpha) = 0 \iff \alpha = 0$
3. $N(\alpha\beta) = N(\alpha)N(\beta)$

Proof. We have $\alpha \in \mathbb{Z}[i]$. Then $\alpha = a + ib$ where $a, b \in \mathbb{Z}$. Clearly, $N(a + ib) = a^2 + b^2 \geq 0$.

$$N(\alpha) = 0 \iff a^2 + b^2 = 0 \iff a = 0, b = 0 \iff \alpha = 0 + i0 = 0.$$

Also, $N(\alpha\beta) = |\alpha\beta|^2 = |\alpha|^2|\beta|^2 = N(\alpha)N(\beta)$. □

Lemma 6.7.2. *$\mathbb{Z}[i]$ is an integral domain.*

Proof. Let $\alpha, \beta \in \mathbb{Z}[i]$.
Then, $\alpha + \beta = (a+ib) + (c+id) = (a+c) + i(b+d) = (c+a) + i(d+b) = \beta + \alpha$. And for any $\alpha \in \mathbb{Z}[i]$, we have $\alpha \cdot 1 = (a+ib) \cdot (1+i0) = (a+ib)$. Thus, we have $\mathbb{Z}[i]$ is a commutative ring with unity.

It remains to prove that $\mathbb{Z}[i]$ has no zero divisors. Suppose $\alpha\beta = 0$. $\alpha\beta = 0 \implies N(\alpha\beta) = 0 \implies N(\alpha)N(\beta) = 0$. But, $N(\alpha), N(\beta) \in \mathbb{Z}$ and $\mathbb{Z}$ has no zero divisors. Thus, $N(\alpha) = 0$ or $N(\beta) = 0 \implies \alpha = 0$ or $\beta = 0$. Therefore $\mathbb{Z}[i]$ has no zero divisors. □

Theorem 6.7.3 (Gaussian integers is a Euclidean Domain)**.** *The function v defined by $v(\alpha) = N(\alpha)$ for every non-zero $\alpha \in \mathbb{Z}[i]$ is a Euclidean norm.*

Proof. **Step 1 :** $v(a) \leq v(ab)$
Let $\beta = b_1 + ib_2$ and $\beta \neq 0$. Then $N(\beta) = b_1^2 + b_2^2 \geq 1$. Thus, $N(\alpha) \leq N(\alpha)N(\beta) = N(\alpha\beta)$.

Step 2 : Division algorithm
That is, $\forall \alpha, \beta \in \mathbb{Z}[i], (\beta \neq 0)\ \exists \sigma, \rho \in \mathbb{Z}[i]$ such that $\alpha = \beta\sigma + \rho$, and $\rho = 0$ or $v(\rho) < v(\beta)$.

Since $\beta \neq 0$, α/β exists. And we have, $\frac{\alpha}{\beta} = \frac{a_1+ia_2}{b_1+ib_2} = \frac{(a_1+ia_2)(b_1-ib_2)}{(b_1+ib_2)(b_1-ib_2)} = \frac{a_1b_1+a_2b_2}{b_1^2+b_2^2} + i\frac{a_2b_1-a_1b_2}{b_1^2+b_2^2} = r + is$ where $r, s \in \mathbb{Q}$. Consider integers q_1, q_2 that are nearest to the rational numbers r, s. Define $\sigma = q_1 + iq_2 \in \mathbb{Z}[i]$ and $\rho = \alpha - \beta\sigma$.

Suppose $\rho \neq 0$. If $\rho = 0$, then the proof is complete. By definition of σ, $|q_1 - r| \leq 0.5$ and $|q_2 - s| \leq 0.5$. And we have,

$$N\left(\frac{\alpha}{\beta} - \sigma\right) = N((r - q_1) + i(s - q_2)) = |q_1 - r|^2 + |q_2 - s|^2 \leq 0.5$$

Since $\alpha = \beta\sigma + \rho$, we have $\rho = \beta(\frac{\alpha}{\beta} - \sigma)$

$$N(\rho) = N\left(\beta\left(\frac{\alpha}{\beta} - \sigma\right)\right) = N(\beta)N\left(\frac{\alpha}{\beta} - \sigma\right) \leq 0.5N(\beta)$$

Thus, $N(\rho) < N(\beta)$. Therefore, Gaussian integers is a Euclidean domain. □

6.7.2 Multiplicative Norm

Definitions 6.7.3 (multiplicative norm)**.** Let D be an integral domain. A function $N : D \to \mathbb{Z}$ is a multiplicative norm if

1. $N(\alpha) = 0 \iff \alpha = 0$
2. $N(\alpha\beta) = N(\alpha)N(\beta),\ \forall \alpha, \beta \in D$

Note : Guassian norm is a multiplicative norm.

Theorem 6.7.4. *Let D be an integral domain with multiplicative norm N. Then $N(1) = 1$ and $|N(u)| = 1$ for any unit u. Suppose $|N(\alpha)| = 1$ only for units in D. Then $N(\alpha)$ is a prime in $\mathbb{Z}$ implies α is an irreducible in D.*

Proof. We know, N is a multiplicative norm. Thus, $N(1) = N(1 * 1) = N(1)N(1) \implies N(1) = 1$.

Suppose u is a unit in D. Then $1 = N(1) = N(uu^{-1}) = N(u)N(u^{-1})$ Since 1 has only two factors $\pm 1 \in \mathbb{Z}$, we have $|N(u)| = 1$.

Suppose $|N(\alpha)| = 1$ only if α is a unit in D. Suppose $\alpha \in D$ and $|N(\alpha)| = p$, where p is a prime in $\mathbb{Z}$. Suppose α is not an irreductible. Then α has a factorisation $\alpha = \beta\gamma$ where $\beta, \gamma \in D$ are non-units. Then $p = N(\alpha) = N(\beta\gamma) = N(\beta)N(\gamma)$. We have, β, γ are non-units and $|N(\beta)| \neq 1$ and $|N(\gamma)| \neq 1$. This is not possible. Therefore, α is an irreducible in D. □

6.7.3 Applications of multiplicative norm

The integral domain $\mathbb{Z}[\sqrt{-5}]$ is not a UFD.(§47.9)

Proof. Elements of $\mathbb{Z}[\sqrt{-5}]$ are of the form $a + b\sqrt{-5}$ and $N(a + b\sqrt{-5}) = a^2 + 5b^2$. If $N(a + b\sqrt{-5}) = a^2 + 5b^2 = 1$, then, $a = \pm 1$ and $b = 0$. Clearly, $|N(\alpha)| = 1$ only if α is a unit in $\mathbb{Z}[\sqrt{-5}]$.

The element $3 \in \mathbb{Z}[\sqrt{-5}]$ is an irreducible. Suppose 3 is not an irreducible. That is, there is a factorisation $3 = \beta\gamma$ where β, γ are non-units. Then $9 = N(3) = N(\beta)N(\gamma)$ where β, γ are non-units. The integer factors of 9 are $1, 3, 9$. Thus, $N(\beta) = N(\gamma) = 3$. Otherwise $N(\beta) = 1$ or $N(\gamma) = 1$. But, there doesn't exist an element $a + b\sqrt{-5}$ such that $N(a + b\sqrt{-5}) = 3$. Similarly, $7 \in \mathbb{Z}[\sqrt{-5}]$ is an irreducible. We have, $49 = N(7) = N(\beta)N(\gamma) \implies N(\beta) = N(\gamma) = 7$ is not possible since there are no elements of norm 7 in $\mathbb{Z}[\sqrt{-5}]$.

Suppose $(1 + 2\sqrt{-5})$ is not an irreducible. Then, $21 = N(1 + 2\sqrt{-5}) = N(\beta)N(\gamma)$. However, we know that the integer factors of 21 are $1, 3, 7, 21$. Without loss of generality, we have $N(\beta) = 3$ which is not possible. Thus, $1 + 2\sqrt{-5}$ is an irreducible.

However neither 3 nor 7 is an associate of $1+2\sqrt{-5}$. Since $|N(1+2\sqrt{-5})| \neq N(3)$ and $N(1 + 2\sqrt{-5}) \neq N(7)$. Clearly, the factorisation of $21 \in \mathbb{Z}[\sqrt{-5}]$ is not unique. Therefore, $\mathbb{Z}[\sqrt{-5}]$ is not a UFD. □

Fermat's $p = a^2 + b^2$ theorem

Theorem 6.7.5 (Fermat). *Let p be an odd prime integer. Then $p = a^2 + b^2$ where $a, b \in \mathbb{Z}$ if and only if $p \cong 1 \pmod 4$*

Proof. Suppose there exists an odd prime p such that $p = a^2 + b^2$. Then a, b are of different parity. Without loss of generality, a is even and b is odd. Let $a = 2r$ and $b = 2s + 1$. Then $a^2 + b^2 = 4r^2 + 4s^2 + 1 \cong 1 \pmod 4$.

Suppose $p \cong 1 \pmod 4$. Then $\mathbb{Z}_p$ is a cyclic group of order $p - 1$ which has a element n of order 4 since $4|(p - 1)$. Then $n^2 = -1 \in \mathbb{Z}_p$. And $n^2 + 1 \cong 0$

(mod p). In other words, $p|(n^2+1)$.

We claim that, p is not an irreducible in $\mathbb{Z}[i]$. Suppose p is an irreducible then $p|(n^2+1) \implies p|(n+i)$ or $p|(n-i)$. And $p|(n+i) \implies n+i = p(a+bi) \implies n = pa$ and $1 = pb$. But, $1 = pb$ is not possible. Similarly, $p|(n-i) \implies n-i = p(a+bi) \implies -1 = pb$ is also not possible. Thus if $p \cong 1$ (mod 4), then p is not an irreducible in $\mathbb{Z}[i]$.

Since p is not an irreducible in $\mathbb{Z}[i]$. $p = (a+bi)(c+di)$ where $a+bi, c+di$ are non-units in $\mathbb{Z}[i]$. Considering the Gaussian norm N on $\mathbb{Z}[i]$ defined by $N(a+bi) = a^2+b^2$. We have, $p^2 = N(p) = N(a+bi)N(c+di) = (a^2+b^2)(c^2+d^2)$. And $N((a+bi) \neq 1$ since for Gaussian norm $N(\alpha) = 1 \iff \alpha = 1$. Similarly, $N(c+di) \neq 1$. Therefore, without loss of generality $N(a+bi) = p$ since $a^2+b^2 > 0$. In other words, $p = a^2+b^2$. $\square$

6.7.4 Exercise §47

1. We have 5 is an odd prime and $5 \cong 1 \mod 4$. Therefore 5 is not an irreducible Gaussian integer. Clearly, $2^2+1^2 = 5 \implies (2+i)(2-i) = 5$.

   And $N(2+i) = 5$ is a prime, thus $2+i$ is an irreducible. Similarly, $2-i$ is also an irreducible. Therefore $5 = (2+i)(2-i)$ is factorisation using irreducibles Gaussian integers. (hint : $2^2+1^=5$. But, $1*1+2*2 = 5$)

   Note : We have $5 = (2+i)(2-i) = (1+2i)(1-2i)$. However 5 has a unique factorisation since $(2+i)$ and $(1-2i)$ are associates $(2+i)(0-i) = 1-2i$ where $-i \in \mathbb{Z}[i]$ is a unit with multiplicative inverse i.

2. We have 7 is an odd prime. However $7 \cong 3$ (mod 4). Therefore 7 is an irreducible Gaussian integer.

3. We have $N(4+3i) = 25$. Clearly, $(4+3i) = \alpha\beta \implies N(\alpha) = N(\beta) = 5$. Therefore, $4+3i = (2-i)(1+2i)$ is a factorisation of $4+3i$. (Hint : $2^2+1^2 = 5$. But, $1*2+1*2 = 4$.)

4. We have $N(6-7i) = 85 = 5*17$. Clearly $6-7i = (2+i)(1-4i)$. (Hint : $2^2+1^2 = 5$ and $4^2+1^2 = 17$. But $1*2+1*4 = 6$.)

5. We have $6 = 2*3$ in $\mathbb{Z}[\sqrt{-5}]$. Then, $N(\alpha\beta) = 4 \implies a^2+5b^2 = 2$ is not possible. And, $N(\alpha\beta) = 9 \implies a^2+5b^2 = 3$ is also not possible. Therefore, $6 = 2*3$ is a factoriation in $\mathbb{Z}[\sqrt{-5}]$.

   However $N(\alpha\beta) = 6 \implies a^2+5b^2 = 6 \implies |a| = |b| = 1$. Clearly, $6 = (1-\sqrt{-5})(1+\sqrt{-5})$ is another factorisation.

Module 3 Field Automorphisms

6.8 Automorphism and Fields §48

Context

1. Let F be a field. Then there exists a unique field $\bar{F}$ which is algebraically closed.

2. Let F be a field. Let α be algebraic over F. Then there exists a unique, monic, irreducible, minimal polynomial $irr(\alpha, F) = p(x) \in F[x]$ such that $p(\alpha) = 0$.

3. $F(\alpha)$ is a vector space of dimension $[F(\alpha) : F] = \deg(\alpha, F)$ over F.

6.8.1 Conjugation Isomorphisms

Definitions 6.8.1 (conjugates). Let E be an algebraic extension of F. Let $\alpha, \beta \in E$. Then α and β are **conjugates over** F if $irr(\alpha, F) = irr(\beta, F)$.

For example, $i, -i$ are conjugates over $\mathbb{R}$. $irr(i, \mathbb{R}) = x^2 + 1 = irr(-i, \mathbb{R})$.

Theorem 6.8.1 (conjugation isomorphism). *Let F be a field. Let α, β be algebraic over F with $\deg(\alpha, F) = n$. The map $\psi_{\alpha,\beta} : F(\alpha) \to F(\beta)$ defined by*

$$\psi_{\alpha,\beta}\left(c_0 + c_1\alpha + \cdots + c_{n-1}\alpha^{n-1}\right) = c_0 + c_1\beta + \cdots + c_{n-1}\beta^{n-1}$$

where $c_j \in F$ is an isomorphism from $F(\alpha)$ onto $F(\beta)$ if and only if α, β are conjugates over F.

Proof. Let $\psi_{\alpha,\beta} : F(\alpha) \to F(\beta)$ be an isomorphism. Let $\deg(\alpha, F) = n$ and $irr(\alpha, F) = q(x)$ given by

$$irr(\alpha, F) = q(x) = a_0 + a_1x + \cdots + a_nx^n$$

Then

$$a_0 + a_1\alpha + \cdots + a_n\alpha^n = 0$$

Thus,

$$\psi_{\alpha,\beta}(0) = \psi_{\alpha,\beta}(a_0 + a_1\alpha + \cdots + a_n\alpha^n) = a_0 + a_1\beta + \cdots + a_n\beta^n = 0$$

Let $irr(\beta, F) = p(x) \in F[x]$. Then, $p(\beta) = 0$. We already have, $q(\beta) = 0$. Let $\phi_\beta : F[x] \to F(\beta)$ be evaluation homomorphism. Then, $\phi_\beta(q(x)) = q(\beta)$. Therefore, $q(x) \in \ker(\phi_\beta)$ However, $\ker(\phi_\beta)$ is the ideal generated by $p(x)$. Therefore, $p(x)$ must divide $q(x)$ in $F[x]$. Thus, $irr(\beta, F)$ divides $irr(\alpha, F)$.

Similarly, $\psi_{\alpha,\beta}^{-1} : F(\beta) \to F(\alpha)$ is an isomorphism and $irr(\alpha, F)$ divides $irr(\beta, F)$. We know that, both the polynomials are monic polynomials. Therefore, $irr(\alpha, F) = irr(\beta, F)$. That is, α, β are conjugates over F.

Let $irr(\alpha, F) = irr(\beta, F) = p(x)$. Let $\phi_\alpha : F[x] \to F(\alpha)$ and $\phi_\beta : F[x] \to F(\beta)$ be evaluation homomorphisms. We know that for any polynomial $f(x) \in F[x]$ with $f(\alpha) = 0$, $p(x)$ divides $f(x)$. Thus, $\ker(\phi_\alpha) = \langle p(x) \rangle$. Similarly, $\ker(\phi_\beta) = \langle p(x) \rangle$.

$$\begin{tikzcd}
& F[x] \arrow[dl, "\phi_\alpha"'] \arrow[d, "\gamma"] \arrow[dr, "\phi_\beta"] & \\
F(\alpha) & F[x]/\langle p(x)\rangle \arrow[l, "\psi_\alpha"'] \arrow[r, "\psi_\beta"] & F(\beta)
\end{tikzcd}$$

Figure 6.1: $\psi_{\alpha,\beta} = \psi_\beta \circ \psi_\alpha^{-1}$

By fundamental homomorphism theorem, there exists a canonical isomorphism

$$\psi_\alpha : F[x]/\langle p(x)\rangle \to \phi_\alpha[F[x]], \quad \psi_\alpha(r(x) + \langle p(x)\rangle) = \phi_\alpha(r(x)) = r(\alpha)$$

Similary, there exists another canonical isomorphism

$$\psi_\beta : F[x]/\langle p(x)\rangle \to \phi_\beta[F[x]], \quad \psi_\beta(r(x) + \langle p(x)\rangle = \phi_\beta(r(x)) = r(\beta)$$

Therefore, $\psi_{\alpha,\beta} = \psi_\beta \circ \psi_\alpha^{-1} : F(\alpha) \to F(\beta)$ is an isomorphism such that

$$\psi_\beta \circ \psi_\alpha^{-1}(c_0 + c_1\alpha + \cdots + c_{n-1}\alpha^{n-1}) = c_0 + c_1\beta + \cdots + c_{n-1}\beta^{n-1}$$

□

Challenge 4. No algebraic number over $\mathbb{R}$ has multiple conjugates.(proved later) However, there exists fields over which there exists algebraic numbers with multiple conjugates.

Proof. Let $F = \mathbb{Z}_2$. Polynomial $x^3 + x + 1$ is an irreducible over $\mathbb{Z}_2$ since $\phi_0(x^3 + x + 1) = 0^3 + 0 + 1 \neq 0$ and $\phi_1(x^3 + x + 1) = 1^3 + 1 + 1 \neq 0$. Let α be a zero of x^3+x+1. Then α is algebraic over $\mathbb{Z}_2$. And $\phi_\alpha(x^3+x+1) = \alpha^3+\alpha+1 = 0$. Then α has two conjugates β, γ.

We have, $[\mathbb{Z}(\alpha) : \mathbb{Z}_2] = 3$. Thus, $|\mathbb{Z}(\alpha)| = 8$. And $\mathbb{Z}_2(\alpha) = \{0, 1, \alpha, \alpha^2, \alpha + \alpha^2, 1 + \alpha, 1 + \alpha^2, 1 + \alpha + \alpha^2\}$. Clearly, $\alpha^3 = \alpha + 1$. We can test each element of $\mathbb{Z}_2(\alpha)$ to find other zeroes of the polynomial $x^3 + x + 1$. Thus, $\phi_{\alpha^2} = 0$ and $\phi_{\alpha+\alpha^2} = 0$. Let $\beta = \alpha^2$ and $\gamma = \alpha + \beta$.

Now the we have conjugate isomorphisms $\psi_{\alpha,\beta} : \mathbb{Z}_2(\alpha) \to \mathbb{Z}_2(\beta)$ and $\psi_{\alpha,\gamma} : \mathbb{Z}_2(\alpha) \to \mathbb{Z}_2(\gamma)$ defined by

Clearly, there are 6 conjugate isomorphisms on $\mathbb{Z}_2(\alpha)$. However, only two of them are distinct. □

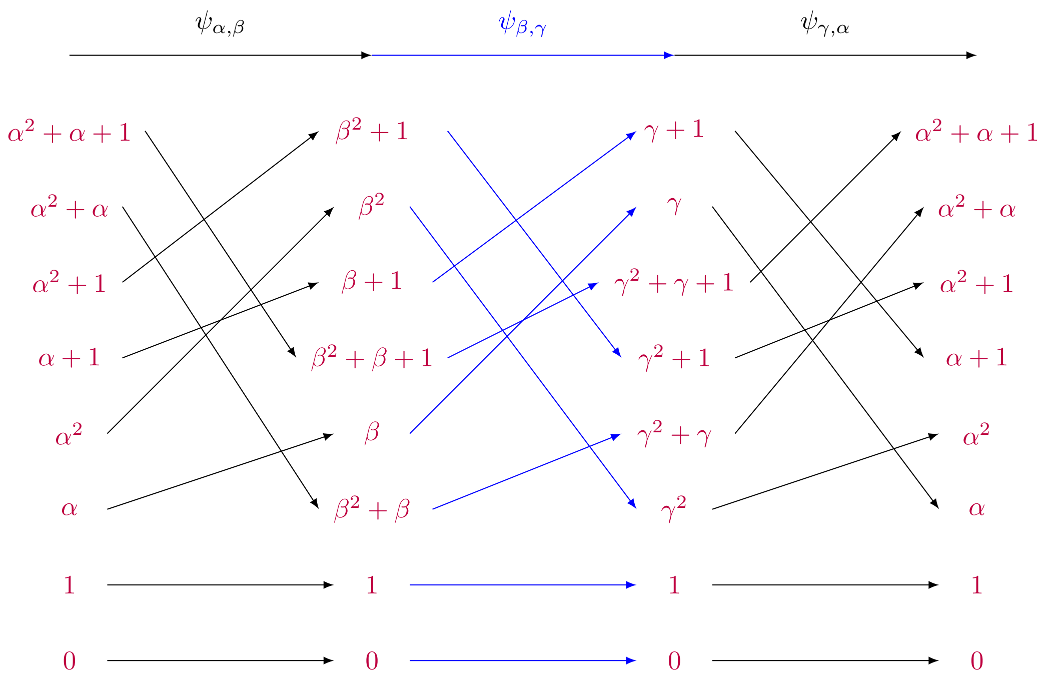


Figure 6.2: $\psi_{\alpha,\beta} : \mathbb{Z}_2(\alpha) \to \mathbb{Z}_2(\beta)$, $\psi_{\beta,\gamma} : \mathbb{Z}_2(\beta) \to \mathbb{Z}_2(\gamma)$, $\psi_{\gamma,\alpha} : \mathbb{Z}_2(\gamma) : \mathbb{Z}_2(\alpha)$.

Corollary 6.8.1.1. *Let α be algebraic over F. Let $E \leq \bar{F}$. Then every isomorphism $\psi : F(\alpha) \to E$ with $\psi(a) = a,\ \forall a \in F$ maps α into its conjugate over F. And for any conjugate β of α over F, there exists a unique isomorphism $\psi_{\alpha,\beta} : F(\alpha) \to \bar{F}$ such that $\psi_{\alpha,\beta}(\alpha) = \beta$ and $\forall a \in F,\ \psi_{\alpha,\beta}(a) = a$.*

Proof. Let $\psi : F(\alpha) \to E$ be an isomorphism such that $\psi(a) = a,\ \forall a \in F$. Let $irr(\alpha, F) = p(x) = a_0 + a_1 x + \cdots + a_n x^n$. Then $\phi_\alpha(p(x)) = p(\alpha) = a_0 + a_1\alpha + \cdots + a_n\alpha^n = 0$. And, $\psi(0) = \psi(p(\alpha)) = a_0 + a_1\psi(\alpha) + \cdots + a_n\psi(\alpha)^n = 0$. That is, $\psi(\alpha)$ is a zero of the polynomial $p(x)$. Thus, $\alpha, \psi(\alpha)$ are conjugates over F.

Let α, β be conjugates over F. Then, the conjugation isomorphism $\psi_{\alpha,\beta} : F(\alpha) \to F(\beta)$ defined by

$$\psi_{\alpha,\beta}(a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1}) = a_0 + a_1\beta + \cdots + a_{n-1}\beta^{n-1}$$

is an isomorphism which maps $a \to a,\ \forall a \in F$ and $\alpha \to \beta$. Since every element in $F(\alpha)$ is of the form $a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1}$. And the definition $a \to a,\ \alpha \to \beta$ compltelely determines the isomorphism. Therefore, there exists a unique conjugate isomorphism which maps α to its conjugate β. ☐

Corollary 6.8.1.2. *Let $f(x) \in \mathbb{R}[x]$. Let $a + ib \in \mathbb{C}$. If $f(a + ib) = 0$, then $f(a - ib) = 0$. Also, complex zeroes of polynomials with real-coefficients occur in conjugate pairs.*

In other words, complex conjugates are conjugates over $\mathbb{R}$.

Proof. We have $\mathbb{C} = \mathbb{R}(i)$ where i is algebraic over $\mathbb{R}$, $irr(i, \mathbb{R}) = x^2 + 1$. Clearly, $i, -i$ are conjugates over $\mathbb{R}$. And $\psi_{i,-i} : \mathbb{C} \to \mathbb{C}$ is a conjugate isomorphism which maps $a + ib \to a - ib$.

Let $a + ib \in \mathbb{C}$. Let $f(x) \in \mathbb{R}[x]$ such that $irr(a + ib, \mathbb{R}) = f(x)$. Since $f(x) \in \mathbb{R}[x]$,

$$f(x) = c_0 + c_1 x + \dots c_n x^n$$

Since $irr(a + ib, \mathbb{R}) = f(x)$,

$$\phi_{a+ib}(f(x)) = f(a + ib) = c_0 + c_1(a + ib) + \cdots + c_n(a + ib)^n = 0$$

Field $\mathbb{R}$ is fixed under the conjugate isomorphism, $\psi_{i,-i}$. Thus, $\psi_{i,-i}(0) = 0$,

$$\psi_{i,-i}(0) = \psi_{i,-i}(f(a + ib)) = f(a - ib) = c_0 + c_1(a - ib) + \cdots + c_n(a - ib)^n = 0$$

Therefore, $(a - ib)$ is a zero of $f(x)$. Let $irr(a - ib, \mathbb{R}) = p(x)$. Then $f(x) \in \langle p(x) \rangle$. That is, $p(x)$ divides $f(x)$. Similarly, $\psi_{-i,i}(p(a - ib)) = p(a + ib) = 0$. And $f(x)$ divides $p(x)$. Therefore, $irr(a - ib, \mathbb{R}) = p(x) = f(x) = irr(a + ib, \mathbb{R})$. Clearly, $a + ib, a - ib$ are conjugates over $\mathbb{R}$.

Let $w = c + id \in \mathbb{C}$ be another conjugate of $z = a + ib$ where $a, b, c, d \in \mathbb{R}$. Suppose $\psi_{z,w} : \mathbb{C} \to \mathbb{C}$ is a conjugation isomorphism. If $a + ib$ is a zero of some $g(x) \in \mathbb{R}[x]$. Then $c + id$ is also a zero of $g(x)$ by conjugation isomorphism. We have, $\psi_{z,w}(a) = a, \psi_{z,w}(b) = b$ since $a, b \in \mathbb{R}$. Thus, $\psi_{z,w}(a + ib) = \psi_{z,w}(a) + \psi_{z,w}(i)\psi_{z,w}(b) = a + \psi_{z,w}(i)b = c + id$. Clearly, $\psi_{z,w}(i) = \frac{c-a}{b} + i\frac{d}{b}$. Thus, $a = c$ and $d = \pm b$ since $i, -i$ are the only zero of $x^2 + 1$. Therefore, conjuates over $\mathbb{R}$ occur in pairs. □

Remark. The notion of complex conjugates

1. Complex conjugates $z, \bar{z}$ are conjugates over the real field $\mathbb{R}$ with irreducible, monic polynomial $irr(z, \mathbb{R}) = (x - \Re(z))^2 + \Im(z)^2$.

2. Any non-real number, $z \notin \mathbb{R}$ which is algebraic over $\mathbb{R}$ has a unique conjugate $\bar{z}$ over $\mathbb{R}$.

Remark. $\sqrt{2}, -\sqrt{2}$ are conjugates over $\mathbb{Q}$ with $irr(\sqrt{2}, \mathbb{Q}) = x^2 - 2$. Thus, $\psi_{\sqrt{2},-\sqrt{2}} : \mathbb{Q}(\sqrt{2}) \to \mathbb{Q}(\sqrt{2})$ is a conjugation isomorphism. Elements of $\mathbb{Q}(\sqrt{2})$ are of the form $a + \sqrt{2}b$ where $a, b \in \mathbb{Q}$. Clearly, $\psi_{\sqrt{2},-\sqrt{2}}(a + \sqrt{2}b) = a - \sqrt{2}b$.

Result : If $\sqrt{2}$ is a zero of a polynomial with rational coefficients, then $-\sqrt{2}$ is also a zero of it. Therefore, $\sqrt{2}$ is an eigen value of a rational matrix, if and only if $-\sqrt{2}$ is also an eigen value of it. Similarly, $a + ib$ is an eigen value of a real matrix, if and only if $a - ib$ is also an eigen value of it.

6.8.2 Automorphism and Fixed Fields

Definitions 6.8.2 (automorphism)**.** Automorphism of a field is an isomorphism of a field onto itself.

Definitions 6.8.3 (fixed)**.** Let $\sigma : F \to F$ be a field automorphism. The set $S = \{a \in F : \sigma(a) = a\}$ of elements invariant under σ is fixed under σ. Let E be a subfield of F such that $\forall a \in E, \sigma(a) = a$. Then E is left fixed by σ.

For example, the conjugation map $\psi_{\sqrt{3},-\sqrt{3}} : \mathbb{Q}(\sqrt{2},\sqrt{3}) \to \mathbb{Q}(\sqrt{2},\sqrt{3})$ leaves $\mathbb{Q}(\sqrt{2})$ fixed.

We have, $irr(\sqrt{2},\mathbb{Q}) = x^2 - 2$. Elements of $\mathbb{Q}(\sqrt{2})$ are of the form $a + b\sqrt{2}$ where $a, b \in \mathbb{Q}$.

$$\mathbb{Q}(\sqrt{2},\sqrt{3}) = \mathbb{Q}(\sqrt{2})(\sqrt{3})$$

And $irr(\sqrt{3},\mathbb{Q}(\sqrt{2})) = x^2 - 3$. Thus, elements of $\mathbb{Q}(\sqrt{2})(\sqrt{3})$ are of the form $x + y\sqrt{3}$ where $x, y \in \mathbb{Q}(\sqrt{2})$.

$$\left(a + b\sqrt{2}\right) + \left(c + d\sqrt{2}\right)\sqrt{3} = a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$$

Elements of $\mathbb{Q}(\sqrt{2},\sqrt{3})$ are of the form $a+b\sqrt{2}+c\sqrt{3}+d\sqrt{6}$ where $a, b, c, d \in \mathbb{Q}$.

$$\psi_{\sqrt{3},-\sqrt{3}}\left(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}\right) = a + b\sqrt{2} - c\sqrt{3} - d\sqrt{2}\sqrt{3}$$

Thus, an element of $\mathbb{Q}(\sqrt{2},\sqrt{3})$ is left fixed by $\psi_{\sqrt{3},-\sqrt{3}}$ if and only if $c, d = 0$.

Theorem 6.8.2. *Let $\{\sigma_i : i \in I\}$ be a collection of automorphisms of field E. Then, the $E_{\{\sigma_i\}} = \{a \in E : \sigma_i(a) = a, \forall i \in I\}$ is a subfield of E.*

The set fixed by an automorphism of a field is a **subfield**.

Quick Revision : How to test whehter F is a subfield of field E ?

1. Preservation of identitites, $0, 1 \in F$ (usually trivial)
2. Closure of binary operations, $\forall a, b \in F,\ a \pm b, ab \in F$ (hard to prove)
3. Existence of inverses (easy to prove)
   (a) $\forall a \in F,\ -a \in F$
   (b) $\forall a \in F,\ (a \neq 0), a^{-1} \in F$.

Proof. Let $a, b \in E_{\{\sigma_i\}}$. Then $\sigma_i(a) = a,\ \forall i \in I$ and $\sigma_i(b) = b,\ \forall i \in I$. Since σ_i's are homomorpisms on E,

$$\sigma_i(a \pm b) = \sigma_i(a) \pm \sigma_i(b) = a \pm b$$

$$\sigma_i(ab) = \sigma_i(a)\sigma_i(b) = ab$$

Thus, $a, b \in E_{\{\sigma_i\}} \implies a \pm b, ab \in E_{\{\sigma_i\}}$.

We have, σ_i's are automorphisms. Thus, $\sigma_i(0) = 0$ and $\sigma_i(1) = 1$ for every $i \in I$. We have, $\sigma_i(0) = 0$.

$$0 = \sigma_i(0) = \sigma_i(a + (-a)) = \sigma_i(a) + \sigma_i(-a) = a - \sigma_i(a)$$

Thus, $\sigma_i(-a) = -a,\ \forall i \in I$. And, $\forall a \in E_{\{\sigma_i\}},\ -a \in E_{\{\sigma_i\}}$.
Similarly, $\sigma_i(1) = 1$. Let $a \neq 0$.

$$1 = \sigma_i(1) = \sigma_i(aa^{-1}) = \sigma_i(a)\sigma_i(a^{-1}) = a\sigma_i(a^{-1})$$

Thus, $\sigma_i(a^{-1}) = a^{-1},\ \forall i \in I$. And, $\forall a \in E_{\{\sigma_i\}},\ (a \neq 0),\ a^{-1} \in E_{\{\sigma_i\}}$. Therefore, $E_{\{\sigma_i\}}$ is a subfield of E. □

Definitions 6.8.4. Let σ be an automorphism on field E. Then $E_{\{\sigma\}}$ is the subfield left fixed by σ.

For example, we have seen earlier that the subfield left fixed by the conjugation isomorphism

$$\psi_{\sqrt{3},-\sqrt{3}} : \mathbb{Q}\left(\sqrt{2},\sqrt{3}\right) \to \mathbb{Q}\left(\sqrt{2},\sqrt{3}\right)$$

is $\mathbb{Q}\left(\sqrt{2}\right)$. Therefore,

$$\mathbb{Q}\left(\sqrt{2},\sqrt{3}\right)_{\{\psi_{\sqrt{3},-\sqrt{3})}\}} = \mathbb{Q}\left(\sqrt{2}\right)$$

Here, $E = \mathbb{Q}\left(\sqrt{2},\sqrt{3}\right)$ and $\sigma = \psi_{\sqrt{3},-\sqrt{3}}$ and $E_{\{\sigma\}} = \mathbb{Q}\left(\sqrt{2}\right)$.

Theorem 6.8.3. *The set of all automorphisms of a field E is a group under function composition.*

Proof. Let E be a field. Let S_E be the set of all permutations on the field E. Clearly, the field permuations are associative under function composition.

We have, identity permutation $i : E \to E$ is an automorphism of E. Let σ be any automorphisms of E. Then, we have

$$i \circ \sigma = \sigma = \sigma \circ i$$

We have σ is a permutation on E. Let σ^{-1} be the inverse permutation of σ. Clearly, $\sigma^{-1} : E \to E$ is also an automorphism. And,

$$\sigma \circ \sigma^{-1} = i = \sigma^{-1} \circ \sigma$$

Let τ be an automorphism on E. Then $\sigma \circ \tau$ is an automorphism on E. Thus, set of all automorphisms on E is a subgroup of permuation group S_E. $\square$

Theorem 6.8.4. *Let E be a field and F be a subfield of E. Then the set of all automorphisms of E leaving F fixed, say $G(E/F)$ is a subgroup of the set of all automorphisms on E. Furthermore, $F \le E_{G(E/F)}$.*

Proof. Let $\sigma, \tau \in G(E/F)$. We have $\sigma(a) = a$ and $\tau(a) = a$ for every $a \in F$. Thus, $(\sigma \circ \tau)(a) = \sigma(\tau(a)) = \sigma(a) = a$. Therefore, $\sigma \circ \tau \in G(E/F)$.

We have, $\sigma(a) = a \implies a = \sigma^{-1}(a)$, $\forall a \in F$. Thus, $\sigma^{-1} \in G(E/F)$. Therefore, $G(E/F)$ is a subgroup of all automorphisms on E.

The subfield $E_{G(E/F)}$ is the set of all automorphims on E leaving F fixed. Let $\sigma \in G(E/F)$. Then,

$$a \in F \implies \sigma(a) = a \implies a \in E_{\{\sigma\}} \implies a \in E_{G(E/F)}$$

Thus, subfield F is contained in $E_{G(E/F)}$. $\square$

Definitions 6.8.5. Let F be a subfield of a field E. The group of E over F, $G(E/F)$ is the set of all automorphisms on E leaving F fixed.

Any automorphism on field E leaves the prime field $\mathbb{Q}$ or $\mathbb{Z}_p$ fixed.

For example, consider the set of all automorphisms on $\mathbb{Q}\left(\sqrt{2},\sqrt{3}\right)$ (leaving $\mathbb{Q}$ fixed.) We have,

$$i, \quad \psi_{\sqrt{2},-\sqrt{2}}, \quad \psi_{\sqrt{3},-\sqrt{3}}, \quad \psi_{\sqrt{2},-\sqrt{2}} \circ \psi_{\sqrt{3},-\sqrt{3}}$$

are the only automorphisms on $\mathbb{Q}\left(\sqrt{2},\sqrt{3}\right)$. Therefore, we have

$$G\left(\mathbb{Q}(\sqrt{2},\sqrt{3})/\mathbb{Q}\right) \simeq V \simeq \mathbb{Z}_2 \times \mathbb{Z}_2$$

since each automorphism is an inverse of itself. Also we have, $[\mathbb{Q}(\sqrt{2},\sqrt{3}):\mathbb{Q}] = 4$ is the order of the group $G(\mathbb{Q}(\sqrt{2},\sqrt{3})/\mathbb{Q})$.

Earlier we have seen, $\mathbb{Z}_2(\alpha)$ where $irr(\alpha,\mathbb{Z}_2) = x^3+x+1$. The conjugates of α over $\mathbb{Z}_2$ are β,γ. However, the distinct automorphisms are,

$$i, \quad \psi_{\alpha,\beta}, \quad \psi_{\beta,\alpha}$$

Therefore,

$$G\left(\mathbb{Z}_2(\alpha),\mathbb{Z}_2\right) \simeq \mathbb{Z}_3 \qquad \text{since } \psi_{\alpha,\beta}\circ\psi_{\beta,\alpha} = i$$

Once again, $[\mathbb{Z}_2(\alpha):\mathbb{Z}_2] = 3 = |G(\mathbb{Z}_2(\alpha)/\mathbb{Z}_2)|$ makes our task much easier.

6.8.3 Frobenius Automorphism

Theorem 6.8.5 (Frobeinus Automorphism)**.** *Let F be a finite field of characteristic p. Then the map $\sigma_p : F \to F$ defined by $\sigma_p(a) = a^p$, $\forall a \in F$ is an automorphism on F. Also $F_{\{\sigma_p\}} \simeq \mathbb{Z}_p$.*

Proof. Let $a, b \in F$. Then

$$\sigma_p(a+b) = (a+b)^p = a^p + b^p = \sigma_p(a) + \sigma_p(b)$$
$$\sigma_p(ab) = (ab)^p = a^p b^p = \sigma_p(a)\sigma_p(b)$$

Thus, σ_p is a homomorphism.

We have,

$$\sigma_p(a) = 0 \implies a^p = 0 \implies a = 0$$

Thus, $\ker(\sigma_p) = \{0\}$. Therefore, σ_p is injective.

Since F is finite and $\sigma_p : F \to F$, σ_p is surjective. Therefore, σ_p is an automorphism on F.

We have, F is of characteristics p. And $\sigma_p(1) = 1^p = 1$. Let $\mathbb{Z}_p$ be the prime field contained in F. Then, $\forall c \in \mathbb{Z}_p$, $\sigma_p(c) = c^p = c$. Therefore, $\mathbb{Z}_p$ is fixed by σ_p.

The polynomial $x^p - x$ has at most p zeroes. Thus, it doesn't have any zero other than elements of the prime subfield of F.

$$a \in F_{\{\sigma_p\}} \iff \sigma_p(a) = a \iff a^p = a \iff a^p - a = 0 \iff \phi_a(x^p - x) = 0$$

Therefore, $F_{\{\sigma_p\}} \simeq \mathbb{Z}_p$. $\square$

6.8.4 Exercises §48

1.

6.9 The Isomorphism Extension Theorem §49

6.9.1 Extension Theorem

Theorem 6.9.1 (isomorphism extension). *Let E be an algebraic extension of field F. Let σ be an isomorphism on F onto F'. Let $\bar{F}'$ be an algebraic closure of F'. Then σ can be extended to an isomorphism τ of E onto $\bar{F}'$ such that $\tau(a) = \sigma(a),\ \forall a \in F$.*

Proof. **Step 1 : Poset S**
Let E be an algebraic extension of F. Let $\sigma : F \to F'$ be an isomorphism. Let S be the set of all pairs (L, λ) such that $F \le L \le E$ and λ is an extension of σ onto a subfield of $\bar{F}$. S is non-empty, since $(F, \sigma) \in S$.

Consider the relation $\le$ on S defined by $(L_1, \lambda_1) \le (L_2, \lambda_2)$ if $L_1 \le L_2$ and $\lambda_1(a) = \lambda_2(a),\ \forall a \in L_1$. Then $\le$ is a partial order on S since $\le$ is reflexive, antisymmetric and transitive.

1. Reflexive
   Let $(L, \lambda) \in S$. Then $L \le L$ and $\lambda(a) = \lambda(a),\ \forall a \in L$.

2. Anti-symmetric
   Let $(L_1, \lambda_1), (L_2, \lambda_2) \in S$. Suppose $(L_1, \lambda_1) \le (L_2, \lambda_2)$ where $L_1 \not\approx L_2$. Then, $F \le L_1 < L_2 \le E$. Thus, $(L_2, \lambda_2) \not\le (L_1, \lambda_1)$.

3. Transitive
   Let $(L_1, \lambda_1) \le (L_2, \lambda_2)$ and $(L_2, \lambda_2) \le (L_3, \lambda_3)$. Then, $F \le L_1 \le L_2 \le L_3 \le E$. And $\lambda_3(a) = \lambda_2(a) = \lambda_1(a), \forall a \in L_1 \subseteq L_2$.
   Thus, $(L_1, \lambda_1) \le (L_3, \lambda_3)$.

Step 2 : Every chain in S has an upper bound (H, λ)
Let $\{(H_i, \lambda_i) : i \in I\}$ be a chain in S. Define $H = \cup\{H_i : i \in I\}$. We know that, if $c \in H$, then $c \in H_i$ for some $i \in I$. Define λ from H onto a subfield of $\bar{F}'$ such that $\lambda(c) = \lambda_i(c)$ where $c \in H_i$.

We need to prove thatλ is an isomorphism. Let $a, b \in H$. Then $a \in H_i$ and $b \in H_j$ for some $i, j \in I$. Then $H_i \le H_j$ or $H_j \le H_i$ since H_i, H_j belongs to a chain. Suppose $H_j \le H_i$, then $a, b \in H_i$. Thus, $a + b, ab \in H_i$. And, λ_i is an isomorphism on H_i. Thus,

$$\lambda(a + b) = \lambda_i(a + b) = \lambda_i(a) + \lambda_i(b) = \lambda(a) + \lambda(b)$$

Similarly,

$$\lambda(ab) = \lambda_i(ab) = \lambda_i(a)\lambda_i(b) = \lambda(a)\lambda(b)$$

If $\lambda(c) = 0$, then $\lambda_i(c) = 0$ for some $i \in I$. Since λ_i is an isomorphism on H_i, $c = 0$. Therefore, $\lambda(c) = 0 \iff c = 0$. Clearly, λ is an isomorphism.

From the defintion of H and λ, (H, λ) is an upper bounded for the chain since $H_i \leq H$ and $\lambda(c) = \lambda_i(c)$, $\forall c \in H_i$.

Step 3 : (K, τ) is maximal, but $K \neq E$
We have shown that, every chain in S has an upper bouned. By Zorn's lemma, S must have a maximal element, say (K, τ).

Suppose $K \neq E$. Let $\alpha \in E$ such that $\alpha \notin K$. Then α is algebraic over K with $p(x) = irr(\alpha, K)$. Let ψ_α be the canonical isomorphism corresponding to the evaluation homomorphism ϕ_α.

$$\psi_\alpha : K[x]/\langle p(x)\rangle \to K(\alpha), \quad \phi_\alpha : K[x] \to K(\alpha)$$

Consider the isomorphism $\tau : K \to K'$. Let $p(x) = a_0 + a_1x + \cdots + a_nx^n$. Then, $q(x) = \tau(p)(x) = \tau(a_0) + \tau(a_1)x + \cdots + \tau(a_n)x^n$ is an irreducible polynomial over K'.

Let α' be a zero of $q(x) \in K'[x]$. Consider $\psi_{\alpha'} : K'[x]/\langle q(x)\rangle \to K'(\alpha')$ which is analogous to ψ_α. Let $\bar{\tau} : K[x]/\langle p(x)\rangle \to K'[x]/\langle q(x)\rangle$ be an extension of τ such that $x + \langle p(x)\rangle \to x + \langle q(x)\rangle$. Then $\psi_{\alpha'} \circ \bar{\tau} \circ \psi_\alpha^{-1} : K(\alpha) \to K'(\alpha')$ is an isomorphism. Clearly, $(K, \tau) \leq \left(K(\alpha), \psi_{\alpha'} \circ \bar{\tau} \circ \psi_\alpha^{-1}\right)$ which contradicts (K, τ) being a maximal element of S. Therefore, $K = E$ and $\tau : K \to K'$ is an extension of $\sigma : F \to F'$ where $F' \leq K' \leq \bar{F}'$. □

Corollary 6.9.1.1. *If $E \leq \bar{F}$ is an algebraic extension of F and $\alpha, \beta \in E$ are conjugates over F. Then the conjugate isomorphism $\psi_{\alpha,\beta} : F(\alpha) \to F(\beta)$ can be extended to an isomorphism of E onto $\bar{F}$.*

Proof. Consider the conjugate isomorphism $\psi_{\alpha,\beta} : F(\alpha) \to F(\beta)$. Let E be an algebraic extension of $F(\alpha)$. Then, by isomorphims extension theorem $\psi_{\alpha,\beta}$ can be extended to an isomorphism from E onto a subfield of $\bar{F}'$. □

Corollary 6.9.1.2. *Let $\bar{F}$ and $\bar{F}'$ be two algebraic closure of F. Then $\bar{F}$ is isomorphic to $\bar{F}'$ under an isomorphism leaving each element of F fixed.*

Proof. Let $i : F \to F$ be the identity isomorphism. Let $\bar{F}, \bar{F}'$, both be algebraic closures of F. By isomorphism extension theorem, i can be extended to an isomorphism τ from the algebraic extension $\bar{F}$ onto a subfield of $\bar{F}'$ leaving F fixed since $i(a) = a$, $\forall a \in F$.

Since, τ is an isomorphism the range of τ, $\tau[\bar{F}]$ is a subfield of $\bar{F}'$. We have, $\tau^{-1} : \tau[\bar{F}] \to \bar{F}$ is also an isomorphism onto $\bar{F}$. Again by isomorphism extension theorem, τ can be extended to an algebraic extension $\bar{F}'$ onto $\bar{F}$. However, τ^{-1} is already onto $\bar{F}$. Thus, $\tau : \bar{F} \to \bar{F}'$ is an isomorphism leaving F fixed. □

In other words, for any field F, its algebraic closure $\bar{F}$ is unique upto an isomorphism leaving F fixed.

6.9.2 Index of field extension

Theorem 6.9.2. *Let E be a finite extension of F. Let σ be an isomorphism of F onto F'. Let $\bar{F}'$ be an algebraic closure of F'. Then, the number of extensions* *of σ to an isomorphism of E onto a subfield of $\bar{F}'$ is finite and is independent of $F', \bar{F}', \sigma$.*

In other words, the number of extensions depends only on F and E.

Proof. **Step 1 : Index of extension of F is independent of F_1**
Let $\sigma_1 : F \to F_1$ and $\sigma_2 : F \to F_2$ be two isomorphisms. Let E be an algebraic extension of F. And $\tau_1 : E \to \tau_1[E]$ and $\tau_2 : E \to \tau_2[E]$ be the extensions of σ_1, σ_2 from E to subfields $\tau_1[E] \leq \bar{F}_1$ and $\tau_2[E] \leq \bar{F}_2$ respectively. Then $\tau_2 \circ \tau_1^{-1} : \tau_1[E] \to \tau_2[E]$ is an isomorphism.

Let λ be the extension of $\tau_2 \circ \tau_1^{-1}$ where $\lambda : \bar{F}_1 \to \bar{F}_2$. Now, we have $\lambda : \bar{F}_1 \to \bar{F}_2$ such that for any isomorphism extension τ_1 from E onto a subfield of $\bar{F}_1$, we have an isomorphism extension τ_2 from E on to $\bar{F}_2$ is given by $\tau_2 = \lambda\tau_1$. Thus, the number of extensions of $\sigma : F \to F'$ is independent of the choice of σ and F' (and $\bar{F}'$).

Step 2 : Number of possible extensions is finite
Let $\sigma : F \to F'$ be an isomorphism. Let E be an algebraic extension of F. Let τ be an isomorphism extension of σ to E leaving F fixed. Then, $E = F(\alpha_1, \alpha_2, \ldots, \alpha_n)$ since every algebraic extensions is a finite extension.

We know that, isomorphisms preserve irreducibility. †[4]
Let $irr(\alpha_j, F) = p(x) \in F[x]$ with $\deg(\alpha_j, F) = k$. Then,

$$p(x) = a_0 + a_1x + \cdots + a_kx^k$$

We have, $\phi_{\alpha_j}(p(x)) = p(\alpha_j) = 0$. Let $q(x) \in F'[x]$ such that

$$q(x) = \sigma_x(p)(x) = \sigma(a_0) + \sigma(a_1)x + \cdots + \sigma(a_k)x^k$$

Then, τ must map α_j to a zero of $q(x)$, say α_j' such that $\sigma(p(\alpha_j)) = 0 = q(\alpha_j')$.

That is, for any isomorphism extension $\tau : E \to \bar{F}'$, we have $\tau(\alpha_j)$ is a zero of $q(x) \in F'[x]$. Clearly, any extension of σ will have to map α_j to a zero of $q(x) \in F'[x]$. However, $q(x)$ has at most k zeroes in $\bar{F}'$.

Therefore, the number of extension of σ is finite as E is a finite extension (number of α_js is finite) and for each α_j the number of possible candidates $\tau(\alpha_j)$ in $\bar{F}'$ is finite. Thus, the number of possible extensions of an isomorphism σ to an algebraic extension E leaving F fixed, is finite. □

Definitions 6.9.1 (index)**.** Let E be an extension of F. The index of E over F, $\{E : F\}$ is the number of isomorphisms of E onto a subfield of $\bar{F}$ leaving F fixed.

Corollary 6.9.2.1. *If $F \leq E \leq K$, where K is a finite extension of F. Then $\{K : F\} = \{K : E\}\{E : F\}$.*

[4]Let $p(x)$ be irreducible over F. Let $\sigma : F \to F'$ be an isomorphism. Then there exists a polynomial $q(x)$ which is irreducible over F'. And if $p(x) = a_0 + a_1x + \cdots + a_nx^n$, then $q(x) = \sigma(a_0) + \sigma(a_1)x + \cdots + \sigma(a_n)x^n$. We may write $q = \sigma_x(p)$ where $\sigma_x : F[x] \to F'[x]$.

Proof. Let E be an algebraic extension of F and K be an algebraic extension of E. Then K is an algebraic extension of F.

We know that, the number of isomorphisms from E onto a subfield of $\bar{F}$ leaving F fixed, $\{E : F\}$ is finite. Now, each of those isomorphisms have $\{K : E\}$ extensions, since there are $\{K : E\}$ different isomorphisms from K onto a subfield of $\bar{F}$ leaving E fixed. Therefore, there are $\{K : E\}\{E : F\}$ isomorphisms from K onto a subfield of $\bar{F}$ leaving F fixed. That is,

$$\{K : F\} = \{K : E\}\{E : F\}$$

□

For example, $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ has 4 different isomorphisms onto a subfield of $\mathbb{C}$.

$$\left\{\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}\right\} = \left\{\mathbb{Q}(\sqrt{2})(\sqrt{3}) : \mathbb{Q}(\sqrt{2})\right\}\left\{\mathbb{Q}(\sqrt{2}) : \mathbb{Q}\right\}$$

We have, $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2})(\sqrt{3})$ and

$$\left\{\mathbb{Q}(\sqrt{2})(\sqrt{3}) : \mathbb{Q}(\sqrt{2})\right\} = \left[\mathbb{Q}(\sqrt{2})(\sqrt{3}) : \mathbb{Q}(\sqrt{2})\right] = \deg\left(\sqrt{3}, \mathbb{Q}(\sqrt{2})\right) = 2$$

$$\left\{\mathbb{Q}(\sqrt{2}) : \mathbb{Q}\right\} = \left[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}\right] = \deg\left(\sqrt{2}, \mathbb{Q}\right) = 2$$

6.9.3 Exercise §49

1.

6.10 Splitting Fields §50

Definitions 6.10.1 (splitting field)**.** Let F be a field with algebraic closure $\bar{F}$. Let $\mathscr{F} = \{f_i(x) \in F[x] : i \in I\}$ be a collection of polynomials over F. Then $E \leq \bar{F}$ is a splitting field of $\mathscr{F}$ over F if E is the smallest subfield of $\bar{F}$ containing F and all the zeroes of polynomials in $\mathscr{F}$.

For example, $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ is the splitting field of $\{x^2 - 2, x^2 - 3\}$ over $\mathbb{Q}$.

Theorem 6.10.1. *Let $F \leq E \leq \bar{F}$. Then E is a splitting field of F if and only if every automorphism of $\bar{F}$ leaving F fixed maps E onto itself. And thus induces an automorphism of E leaving F fixed.*

Proof. Let E be a splitting field of F in $\bar{F}$ of $\{f_i(x) : i \in I\}$. Let σ be an automorphism of $\bar{F}$ leaving F fixed.

Step 1 : $\{\alpha_j : j \in J\}$ **generates** E
Let $\{\alpha_j : j \in J\}$ be the set of all zeroes of $\mathscr{F} = \{f_i(x) : i \in I\}$. We know that, the elements of $F(\alpha_j)$ are of the form

$$\phi_{\alpha_j} g(x) = g(\alpha_j) = a_0 + a_1\alpha_j + \cdots + a_{n_j-1}\alpha^{n_j-1}$$

where $a_k \in F$, $\deg(\alpha_j, F) = n_j$. Clearly, $g(x)$ is some polynomial in $F[x]$ of degree less than n_j.

Let S be the set of all finite sums of finite products of elemenets of the form $g(\alpha_j)$. Clearly, $S \subset E$ and S is closed under addition, closed under multiplication, contains identities $(0, 1 \in S)$ and additive inverses of its elements.

Let $\beta \in S$. Then β is a finite sum of finite products of elements of the form $g(\alpha_{j_k})$. Then, an element $\beta \in S$ belongs to some $F(\alpha_{j_1}, \alpha_{j_2}, \dots, \alpha_{j_r}) \subset S$. Thus, S has multiplicative inverse of each of its non-zero elements. Clearly, S is a subfield of E containing F and α_js. That is, S is the splitting field of F containing all zeroes of $\mathscr{F}$. Therefore, $S = E$. In other words, $\{\alpha_j : j \in J\}$ generates E over F. †[5]

Step 2 : σ is an automorphism of E
Since $\{\alpha_j : j \in J\}$ spans E over F, $\sigma : \bar{F} \to \bar{F}$ is completely determined on E by the images of α_js, say $\sigma(\alpha_j)$. And $\sigma(\alpha_j)$ is a zero of $irr(\alpha_j, F)$. If $f_i(\alpha_j) = 0$, then $irr(\alpha_j, F)$ divides $f_i(x)$. Similarly $irr(\sigma(\alpha_j), F)$ divides $f_i(x)$. Thus, σ maps E onto a subfield of E isomorphically. Similarly, σ^{-1} maps E onto a subfield of E isomorphically. Therefore, σ is an automorphism on E.

Converse Part

□

Definitions 6.10.2. Let E be an extension field of F. A polynomial $f(x) \in F[x]$ splits in $E[x]$ if it factors into linear components in $E[x]$.

For example, $x^4 - 5x^2 + 6 \in \mathbb{Q}$ splits in $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ since

$$x^4 - 5x^2 + 6 = (x - \sqrt{2})(x + \sqrt{2})(x - \sqrt{3})(x + \sqrt{3})$$

Corollary 6.10.1.1. *If $E \leq \bar{F}$ is a splitting field over F, then every irreducible polynomial in $F[x]$ having a zero in E splits in E.*

Proof. □

Corollary 6.10.1.2. *If $E \leq \bar{F}$ is a splitting field over F, then every isomorphic mapping of E onto a subfield of $\bar{F}$ leaving F fixed is an automorphism of E. In particular, if E is splitting field of finite degree over F, then*

$$\{E : F\} = |G(E/F)|$$

Proof. □

6.10.1 Exercise §50

1.

[5] E is a vector space over the field F spanned by $\{\alpha_j : j \in J\}$. Let $\beta \in E$. Then, $\beta = a_0\alpha_0 + a_1\alpha_1 + \cdots + a_r\alpha_r$ where $a_k \in F$.

Module 4 Galois Theory

6.11 Separable Extensions §51

6.11.1 Multplicities of zeroes of a polynomial

6.11.2 Separable extensions

6.11.3 Perfect fields

6.11.4 The primitive element theorem

6.11.5 Exercise §51

1.

6.12 Galois Theory §53

6.12.1 Context

6.12.2 Normal extensions

6.12.3 The main theorem

6.12.4 Galois groups over finite fields

6.12.5 Exercise §53

1.

6.13 Illustrations of Galois Theory §54

6.13.1 Symmetric functions

6.13.2 Exercise §54

1.

6.14 Cyclotomic Extensions §55

6.14.1 The Galois group of a cyclotomic extension

6.14.2 Constructible polygons

6.14.3 Exercise §54

1.

6.15 Insolvability of the quintic

There exists polynomials of degree 5 in $\mathbb{Q}[x]$ that are not solvable by radicals in $\mathbb{Q}$.

Subject 7

ME010202 Advanced Topology

7.1 Module I

Q : Why these axioms are called separation axioms ?
T_0, T_1, T_2 axioms separates points from points.
$T_3, T_{3\frac{1}{2}}$ axioms separates points from closed subsets.
T_4 axiom separates closed subsets from closed subsets.

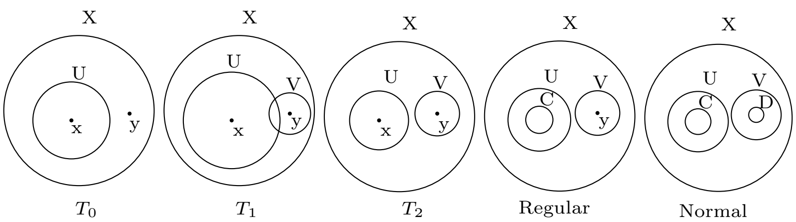


Figure 7.1: Separation Axioms

7.1.1 Compactness and Separation Axioms

[Joshi, 1983, chapter 7 §2.1-10]

Proposition 7.1.1. *Let X be a T_2 space, $x \in X$ and F is a compact subset of X not containing x. Then there exist open subsets U, V such that $x \in U$, $F \subset V$, $U \cap V = \phi$.*

T_2 spaces separates points from compact sets.

Proof. Let x be a points in X and F be a compact subset not containing x.
X is $T_2 \implies \forall y \in F,\ \exists U_y, V_y \in \mathcal{T},\ x \in U_y,\ y \in V_y, U - y \cap V_y = \phi$. †[1])
$\mathcal{C} = \{V_y : y \in F\}$ is an open cover of compact subset F.

[1] **Q :** Do we need compactness for this result ?
A : Given $x \in X$. For each point $y \in F$, we have two open sets namely U_y and V_y that separates these two points x and y in X. But, the intersection $\cap_{y \in F}\{U_y\}$ is not necessarily an

$\implies$ there exists a finite subcover, $\mathcal{C}' = \{V_{y_i} : y_i \in F, i = 1, 2, \cdots, n\}$.
Define $U = \cap_{i=1}^{n}\{U_{y_i}\}$ and $V = \cup_{i=1}^{n}\{V_{y_i}\}$.
$\implies x \in U,\ F \subset V,\ U \cap V = \phi$. □

Corollary 7.1.0.1. *A compact subset in a T_2 space is closed.*

Proof. Let $x \in X - F$.
By proposition, $x \in U$, $F \subset V$, $U \cap V = \phi \implies x \in U \subset X - V \subset X - F$
$X - F$ is a nbd of each of its points. Thus $X - F$ is open. ($\star$[2]) □

Corollary 7.1.0.2. *Every map from a compact space into a T_2 space is closed. And its range is a quotient space of the domain.*

Proof. Suppose X is compact, Y is T_2 and $f : X \to Y$ is continuous.
Let C be a closed subset of X.
$\implies C$ is compact, since compact is weakly hereditary. $\star$[3])
$\implies f(C)$ is compact, since compactness is preserved by continuous functions. ($\star$[4])
By corollary, $f(C)$ is compact, Y is $T_2 \implies f(C)$ is a closed subset of Y.
Thus $f : X \to Y$ is closed. †[5])
$f : X \to Y$ is closed $\implies f : X \to f(X)$ is a quotient map, since every closed, surjective map is a quotient map.. □

Corollary 7.1.0.3. *A continuous bijection from a compact space onto a T_2 space is a homeomorphism.*

Proof. Let G be an open subset of X.
Then $X - G$ is closed.
By corollary, f is closed, since f : compact $\to T_2$.
f is closed $\implies f(X - G)$ is closed.
$f(X - G) = f(X) - f(G)$ since f in injective.
$\implies f(X - G) = Y - f(G)$ since f surjective.
$\implies f(G)$ is open.
Thus f is an open map.
Every continuous bijective, open map is a homeomorphism. □

Corollary 7.1.0.4. *Every continuous, one-to-one function from a compact space into a T_2 space is an embedding.*

Proof. Let $f : X \to Y$ be continuous and injective.
$f : X \to f(X)$ is surjective.
$\implies f : X \to f(X)$ is continuous, surjective.
By corollary, $f : X \to f(X)$ is a homeomorphism.
Thus f is an embedding of X onto $f(X) \subset Y$. □

---

open subset of X, since F is not necessarily a finite subset of X and an **arbitary** intersection of open subsets is not necessarily an open subset. Therefore, we have restrict this family into a finite family. We use compactness for this part.

[2]Neighbourhood characterisation of open subsets : Let $G \subset X$ be a nbd of each of its points. Then G is an open subset of X.

[3]Compactness is weakly hereditary : Suppose X is compact. Then every closed subset of X is compact.

[4]Compactness is preserved by continuous functions : Suppose G be a compact subset of a topological space X. And function $f : X \to Y$ be a continuous function. Then $f(G)$ is a compact subset of the topological space Y.[Joshi, 1983, 6.1.8]

[5]Closed function is a function which maps closed subsets into closed subsets.

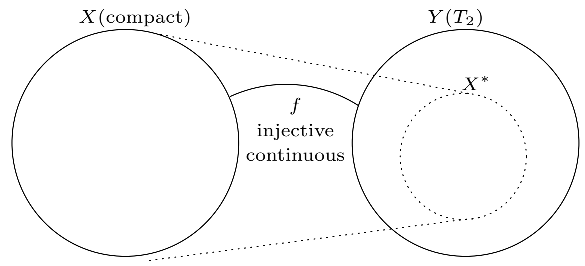


Figure 7.2: Embedding compact space into hausdorff space

Theorem 7.1.1. *Every compact T_2 space is a T_3 space.*

Proof. $T_2 \implies T_1$
It is enough to prove that compact, T_2 space is regular.
Let x be a point in X and C be a closed subset not containing x.
Then C is compact, compactness is weakly hereditary.
By proposition, T_2 space separates points from compact subsets.
Thus $T_2 + \text{compact} \implies T_3$ (†[6]) □

Proposition 7.1.2. *Let X be a regular space, C a closed subset of X and F a compact subset of X, such that $C \cap F = \phi$. Then there exist open subsets U, V such that $C \subset U$, $F \subset V$ and $U \cap V = \phi$.*

Regular spaces separates closed subsets from compact subsets.

Proof. proof technique is same as T_2 separates points from compact subsets.
Let C be a closed subset, F be a compact subset of X and $C \cap F = \phi$.
X is regular $\implies \forall y \in F,\ \exists U_y, V_y \in \mathcal{T},\ C \subset U_y,\ y \in V_y,\ U_y \cap V_y = \phi$
Define $\mathcal{C} = \{V_y : y \in F\}$ is cover of compact subset F.
$\implies \exists \mathcal{C}'$ such that $\mathcal{C}' = \{V_{y_i} : i = 1, 2, \cdots, n\}$ is a finite subcover.
Define $U = \cap_{i=1}^{n}\{U_{y_i}\}$ and $V = \cup_{i=1}^{n}\{V_{y_i}\}$.
$\implies C \subset U,\ F \subset V,\ U \cap V = \phi$. □

Theorem 7.1.2. *Every regular, Lindeloff space is normal.*

Proof. Let C, D be two disjoint, closed subsets of a regular, lindeloff space X.
X is regular $\implies \forall x \in C,\ \exists U_x, V_x$ such that $x \in U_x,\ D \subset V_x,\ U_x \cap V_x = \phi$
X is regular $\implies \forall y \in D,\ \exists U_y, V_y$ such that $C \subset U_y,\ y \in U_y,\ U_y \cap U_y = \phi$
Then $\{U_x : x \in C\}$ and $\{V_y : y \in D\}$ are open covers of C and D respectively. †[7])
Let $\{U_n : n = 1, 2, \cdots\}$ and $\{V_n : n = 1, 2, \cdots\}$ be their countable subcovers.
Define $G_n = U_n - \cup_{i=1}^{n}\overline{V_i}$ and $H_n = V_n - \cup_{i=1}^{n}\overline{U_i}$.
Define $G = \cup_{i=1}^{\infty} G_n$ and $H = \cup_{i=1}^{\infty} H_n$.
Claim : $C \subset G$ (similary $D \subset H$)
$x \in C \implies x \in U_n$ for some n, since $\{U_n : n \in \mathbb{N}\}$ is a cover of C
$\forall m,\ \overline{V_m} \subset X - C \implies \forall m,\ x \notin \overline{V_m} \implies x \in G$
Claim : $G \cap H = \phi$
$x \in G \cap H \implies x \in G_m \cap H_n$ for some m, n
With loss of generality, $n \geq m, x \in G_m \implies x \in U_m \implies x \notin H_n$ □

[6] $T_3 \not\Longrightarrow$ compact.
[7] Here, U_x, V_y are unrelated.

Corollary 7.1.2.1. *Every regular, second countable space is normal.*

Proof. Every second countable space is lindeloff.
By theorem, every regular, lindeloff space is normal. □

Corollary 7.1.2.2. *Every compact, T_2 space is T_4.*

Proof. By proposition, every compact, T_2 space is regular.
Every compact space is lindeloff since finite subcovers are countable.
By theorem, every regular, lindeloff space is normal. □

7.1.2 The Urysohn Characterisation of Normality

X is normal $\iff$ there exist Urysohn functions
Q : Significance of Urysohn's Lemma ?

1. There is no analog of Urysohn's Lemma for Regular spaces.
2. Constructs a nice real-valued function even on non-metrisable spaces.

Q : We know that, the completely regular space separates points from closed subsets using a real-valued function. Which space separates closed subsets from closed subsets using a real-valued function ?
A : Normal space(by Urysohn's Lemma). Not completely normal($\star$[8])

Proposition 7.1.3. *Let A B be subsets of a space X and suppose there exists a continuous function $f : X \to [0, 1]$, such that $f(x) = 0,\ \forall x \in A$ and $f(x) = 1,\ \forall x \in B$. Then there exists disjoint open subsets $U,\ V$ such that $A \subset U$ and $B \subset V$.*

Proof. Let f be a continuous function, $f(x) = 0,\ \forall x \in A$ and $f(x) = 1,\ \forall x \in B$.
Define $G = [0, \frac{1}{2})$, $H = (\frac{1}{2}, 1]$ and $U = f^{-1}(G)$, $V = f^{-1}(H)$.
$\implies A \subset U$, $B \subset V$ and $U \cap V = \phi$ since $G \cap H = \phi$. □

Corollary 7.1.2.3 (Urysohn's Lemma : Sufficient Condition)**.** *If X has the property that for any disjoint closed subsets $A,\ B$ of X, there exists a continuous function $f : X \to [0, 1]$ such that $f(x) = 0,\ \forall x \in A$ and $f(x) = 1,\ \forall x \in B$, then X is normal.*

There exists a Urysohn function $f \implies X$ is normal

Proof. Let A, B be two closed subsets of X.
By proposition, X normal. □

Theorem 7.1.3 (Urysohn's Lemma)**.** *A topological space X is normal iff it has the property that for every mutually disjoint, closed subsets $A,\ B$ of X, there exists a continuous function $f : X \to [0, 1]$ such that $f(x) = 0$ for all $x \in A$ and $f(x) = 1$ for all $x \in B$*

[8]Completely Normal Space separates any two subsets. [Joshi, 1983, chapter 7 Exer. 1.11] That is these subsets need not be closed. Thus every completely normal space is normal, but not the other way. Thus T_1 + Completely Normal $= T_5 \supset T_4$.

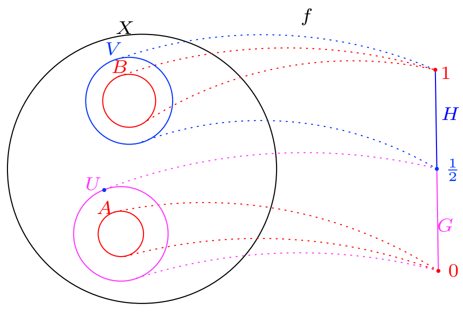


Figure 7.3: Urysohn's Lemma : Sufficient Condition

Proof. **Urysohn's Lemma : necessary condition**
X is normal $\implies$ there exist Urysohn functions

Let A, B be two closed subsets in a normal space X.
Enumerate rationals in the unit interval.
$\mathbb{Q} \cap [0,1] = \{0, 1, \cdots\} = \{q_0, q_1, q_2, \cdots\}$.
Define $F_1 = F_{q_1} = X - B$.
$A \subset X - B \implies \exists H$ such that $A \subset H \subset \overline{H} \subset X - B$. ($\star$[9])
Define $F_{q_0} = F_0 = H$.

$\overline{F_{q_0}} \subset F_{q_1} \implies$ true for $n = 1$.
Suppose : $\overline{F_{q_i}} \subset F_{q_j}, \ \forall j > i$ is true for $j = 1, 2, \cdots, n-1$.
Define $q_i = \sup\{q_k : q_k < q_n, k < n\}$ and $q_j = \inf\{q_k : q_k > q_n, k < n\}$ (†[10])
$\implies q_i < q_n < q_j$ and $\overline{F_{q_i}} \subset F_{q_j}$.
$\overline{F_{q_i}} \subset F_{q_j} \implies \exists H$ such that $\overline{F_{q_i}} \subset H \subset \overline{H} \subset F_{q_j}$. Define $F_{q_n} = H$.
Therefore, $\overline{F_{q_j}} \subset F_{q_i}, \ \forall j > i$.

Now, $\{F_t : t \in \mathbb{Q}\}$ has the properties required in lemma 2.
By lemma 2, there exists a function f. This is a Urysohn function. (†[11]) □

Lemma 7.1.4. *Let $f : X \to [0,1]$ be continuous. For each $t \in \mathbb{R}$ let $F_t\{x \in X : f(x) < t\}$. Then the indexed family $\{F_t : t \in \mathbb{R}\}$ has the following properties*

1. *F_t is an open subset of X for each $t \in \mathbb{R}$*
2. *$F_t = \phi$ for $t < 0$*
3. *$F_t = X$ for $t > 1$*
4. *For any $s, \ t \in \mathbb{R}, \ s < t \implies \overline{F_s} \subset F_t$.*

Moreover, for each $x \in X, \ f(x) = \inf\{t \in \mathbb{Q} : x \in F_t\}$.

Proof. Not needed. □

[9] Equivalent condition for normality : Let A be closed subset and G be an open subset containing A. Then there exists open subset H such that $A \subset H \subset \overline{H} \subset G$. cite[7.1.16(3)]joshi

[10] Let $\mathbb{Q} \cap [0,1] = \{0, 1, 0.3, 0.7, 0.8, \mathbf{0.5}, 0.6, \cdots\}$. Consider $n = 5$. Then $q_n = q_5 = 0.5$ and in the respective induction step, $q_i = \sup\{0, 0.3\} = 0.3$ and $q_j = \inf\{1, 0.7, 0.8\} = 0.7$

[11] You will have to replace "Urysohn function" with "There exists a continuous real-valued function, $f : X \to [0,1]$ such that $f(x) = 0, \ \forall x \in A, \ f(x) = 1, \ \forall x \in B$".

Lemma 7.1.5. *Let X be a topological space and suppose $\{F_t : t \in \mathbb{Q}\}$ is a family of sets in X such that*

1. *F_t is open in X for each $t \in \mathbb{Q}$*
2. *$F_t = \phi$ for $t \in \mathbb{Q},\ t < 0$*
3. *$F_t = X$ for $t \in \mathbb{Q},\ t > 1$*
4. *$\overline{F_s} \subset F_t$ for $s,\ t \in \mathbb{Q},\ s < t$*

For $x \in X$, let $f(x) = \inf\{t \in \mathbb{Q} : x \in F_t\}$. Then f is a continuous real-valued function on X and it takes values in the unit interval $[0, 1]$.

Proof. Function f is well-defined and Range$(f) = [0, 1]$.
Define $H = \{x \in X : f(x) < s\} = f^{-1}(-\infty, s)$ and
$K = \{x \in X : f(x) > s\} = f^{-1}(s, \infty)$.
$\mathcal{S} = \{(s, \infty), (-\infty, s) : s \in \mathbb{R}\}$ is a subbase for $\mathbb{R}$ with usual topology.
Claim : $H = \cup\{F_t : t \in \mathbb{Q}, t < s\}$
$x \in H \implies f(x) < s \implies \inf G_x < s \implies \exists q \in \mathbb{Q}\ (q < s),\ x \in F_q$
$\implies x \in \cup\{F_t : t \in \mathbb{Q}, t < s\} \implies H \subset \cup\{F_t : t \in \mathbb{Q},\ t < s\}$.
$x \in \cup\{F_t : t \in \mathbb{Q},\ t < s\} \implies \exists t \in \mathbb{Q}\ (t < s),\ x \in F_t \implies \inf G_x < s$
$\implies f(x) < s \implies x \in H \implies \cup\{F_t : t \in \mathbb{Q},\ t < s\} \subset H$.
Claim : $X - K = \cap\{\overline{F_t} : t \in \mathbb{Q}, t > s\}$.
$x \in X - K \implies x \notin K \implies f(x) \leq s \implies \inf G_x \leq s$
Let $t \in \mathbb{Q}\ (s < t) \implies \exists q \in G_x\ (s < q < t) \implies x \in \overline{F_q} \subset F_t \subset \overline{F_t}$.
$\forall t \in \mathbb{Q}\ (t > s),\ x \in \overline{F_t} \implies x \in \cap\{\overline{F_t} : t \in \mathbb{Q},\ t > s\}$.
Thus, $X - K \subset \cap\{\overline{F_t} : t \in \mathbb{Q},\ t > s\}$.
$x \in \cap\{\overline{F_t} : t \in \mathbb{Q},\ t > s\}$
Suppose(1) $x \in K \implies s < f(x) \implies \exists q, t \in \mathbb{Q}\ (s < q < t),\ t < f(x)$
Suppose(2) $x \in \overline{F_q} \implies x \in \overline{F_q} \subset F_t \implies \inf G_x < t \implies f(x) < t$
is a contradiction(2). Thus $x \notin \overline{F_q}$.
$x \notin \overline{F_q} \implies x \notin \cap\{\overline{F_t} : t \in \mathbb{Q},\ t > s\}$ is a contradition (1). Thus $x \notin K$
Thus, $\cap\{\overline{F_t} : t \in \mathbb{Q},\ s > t\} \subset X - K$.
Inverse images of subbase elements are open. Thus f is continuous. □

Corollary 7.1.5.1. *All T_4 spaces are completely regular and hence Tychonoff.*

Proof. Let $x \in X$ and D be closed subset not containing x. We have X is a T_4 space. Therefore X is T_1 as well as normal.

Now the singleton set, $\{x\}$ is closed, since X is a T_1 space. And by **Urysohn's lemma** for disjoint, closed subsets $\{x\}$, D there exists a Urysohn function which is a continuous, real-valued function $f : X \to [0, 1]$ such that $f(x) = 0$ and $f(y) = 1$ for all $y \in D$. Therefore the space X is completely regular and hence Tychonoff. □

Remark (Urysohn function). The function whose existence is asserted by Urysohn's lemma is called a Urysohn function

7.1.3 Tietze Characterisation of Normality

Tietze Characterisation of normality
X is normal $\iff$ there exist continuous extension of real-valued functions on closed subsets.

Proposition 7.1.4. *Let A be a subset of a space X and let $f : A \to \mathbb{R}$ be continuous. Then any two extensions of f to X agree on $\overline{A}$. In other words, if at all an extension of f exists its values on $\overline{A}$ are uniquely determined by values of f on A.*

Proposition 7.1.5 (Tietze : necessary condition)**.** *Suppose a topological space X has the property that for every closed subset A of X, every continuous real valued function on A has a continuous extension to X. Then X is normal.*

Definitions 7.1.1 (Pointwise Convergence)**.** Let X be a topological space and (Y, d) a metric space. Then a sequence of functions $\{f_n\}$ from X to Y converges pointwise to f if for every $x \in X$ the sequence $\{f_n(x)\}$ converges to $f(x)$ in Y.

In other words, given a very small value, $\epsilon > 0$, there exists some $\delta > 0$ such that for every $x \in X$ there exists $N_x \in \mathbb{N}$. This N_x may be different for different values of x and for every $n > N_x$, $d(f(x), f_n(x)) < \delta$.

Definitions 7.1.2 (Uniform Convergence)**.** Let X be a topological space and (Y, d) a metric space. Then a sequence of functions $\{f_n\}$ from X to Y converges uniformly to f if given a small $\epsilon > 0$, there exists $\delta > 0$ such that there exists $N \in \mathbb{N}$. This N is independent of the value of x and for every $n > N$, $d(f(x), f_n(x)) < \delta$.

Proposition 7.1.6. *Let X, (Y, d), $\{f_n\}$ and f be as above and suppose $\{f_n\}$ converges to f uniformly. If each f_n is continuous, then f is continuous.*

Definitions 7.1.3 (Uniform Convergence of Series)**.** Let X be a topological space and (Y, d) be a metric space. Then a series of function $\sum_{n=1}^{\infty} f_n$ converges uniformly to f if the sequence of partial sums converges uniformly to f.

In other words, let $g_m = \sum_{n=1}^{m} f_n$. Then $\sum_{n=1}^{\infty} f_n$ converges to f uniformly if the sequence of partial sums $\{g_m\}$ converges to f uniformly.

Proposition 7.1.7. *Let $\sum_{n=1}^{\infty} M_n$ be a convergent series of non-negative real numbers. Suppose $\{f_n\}$ is a sequence of real valued functions on a space X such that for each $x \in X$ and $n \in \mathbb{N}$, $|f_n(x)| \le M_n$. Then the series $\sum_{n=1}^{\infty} f_n$ converges uniformly to a real valued function on X.*

Theorem 7.1.6. *Let A be a closed subset of a normal space X and suppose function $f : A \to [-1, 1]$ is a continuous function. Then there exists a continuous function $F : X \to [-1, 1]$ such that $F(x) = f(x)$ for all $x \in A$.*

Theorem 7.1.7. *Let A be a closed subset of a normal space X and suppose function $f : A \to (-1, 1)$ is a continuous function. Then there exists a continuous function $F : X \to (-1, 1)$ such that $F(x) = f(x)$ for all $x \in A$.*

Corollary 7.1.7.1 (Tietze : sufficient condition)**.** *Any continuous real-valued function on a closed subset of a normal space can be extended continuously to the whole space.*

7.2 Products and Coproducts

7.2.1 Cartesian Products of Families of Sets

7.2.2 The Product Topology

7.2.3 Productive Properties

7.3 Embedding and Metrisation

7.3.1 Evaluation Functions into Products

7.3.2 Embedding Lemma and Tychonoff Embedding

7.3.3 The Urysohn Metrisation Theorem

7.4 Nets and Homotopy

7.4.1 Definition and Convergence of Nets

Definitions 7.4.1 (Directed Set)**.** [Joshi, 1983, 10.1.1]
A directed set D is a pair $(D, \geq)$ where D is a nonempty set and ge is a binary relation on D such that

1. The relation 'follows'($\geq$) is transitive. ie, $m \geq n,\ n \geq p \implies m \geq p$
2. The relation 'follows'($\geq$) is reflexive. ie, For every $m \in D,\ m \geq m$
3. For any $m, n \in D$, there exists $p \in D$ such that $p \geq m$ and $p \geq n$.

sequence in a set X is a function f from the set of all integers into X.

Definitions 7.4.2 (Net)**.** [Joshi, 1983, 10.1.2]
A net in a set X is a function S from a directed set D into the set X.

Remark. The set $\mathbb{N}$ together with the relation 'less than or equal to'($\leq$) is a directed set. Clearly, the relation 'less than or equal to' is reflexive and trasitive. And the third condition is true iff every finite subset E of D has an element $p \in E$ such that p follows each element of E. This is a weaker notion compared to the well ordering principle($\star$[12]of the set of all integers. Thus $\mathbb{N}$ is a directed set and every sequence in X is also a net in X.

Remark (Significance of Net)*.* A net on a set is a generalisation of 'a sequence on a set' obtained by simplifying the domain of the sequence into a directed set. The notion directed set is derived by assuming a few properties of $\mathbb{N}$.

The convergence of sequence is not strong enough to characterise topologies as the limit of convergent sequences are unique for both Hausdorff and Co-countable spaces. The notion of Net allows us to differentiate between Hausdorff spaces from Co-countable spaces in terms of convergence of nets. The limit of a convergent net on a topological space is unique iff it is a Hausdorff space. ie, We have removed a few restrictions, so that we will have some convergent nets (which are obviously not sequences) with multiple limit points for Co-countable spaces.

Remark. Examples of Directed Sets

1. Let X be a topological space and $x \in X$. Then the neighbourhood system $\mathcal{N}_x$ is a directed set with the binary relation $\subset$ (subset/inclusion).
   (a) Let $U,\ V,\ W$ be any three neighbourhoods of $x \in X$ such that $U \subset V$ and $V \subset W$. Then, clearly $U \subset W$. Therefore, $U \geq V,\ V \geq W \implies U \geq W$.
   (b) Let U be any neighbourhood of $x \in X$, then $U \subset U$.
   Therefore, $U \geq U$.
   (c) Let U, V be any two neighbourhoods of $x \in X$, then there exists their intersection $W = U \cap V$, which is a neighbourhood of x. Clearly $W \subset U$ and $W \subset V$.
   Therefore $\forall U, V \in \mathcal{N}_x, \exists W \in \mathcal{N}_x$ such that $W \geq U$ and $W \geq V$.

[12] Well-ordering principle : Every subset of $\mathbb{N}$ has a least element in it.

2. Let $\mathcal{P}$ be the set of all partitions on closed unit interval $[0,1]$. A partition $P \in \mathcal{P}$ is a refinement of $Q \in \mathcal{P}$ if every subinterval in P is contained in some subinterval of Q. Then $\mathcal{P}$ with the binary relation refinement is a directed set.

   (a) Suppose $P,\ Q,\ R$ are three partitions of $[0,1]$ such that P is a refinement of Q and Q is a refinement of R, then clearly P is a refinement of R since each subinterval of P is contained some subinterval of Q, which is contained in some subinterval of R.
   Therefore, $P \geq Q,\ Q \geq R \implies P \geq R$

   (b) Suppose P is a partition of $[0,1]$. Then trivialy, P is a refinement of itself since every subinterval of P is contained in the same subinterval of P.
   Therefore, $\forall P \in \mathcal{P},\ P \geq P$

   (c) Suppose $P,\ Q$ be any two partition of $[0,1]$. Then $R = P \cup Q$ is a refinement of both the partitions.
   Therefore $\forall P, Q \in \mathcal{P},\ \exists R \in \mathcal{P}$ such that $R \geq P$ and $R \geq Q$

For example, let $P = \{0,\ 0.3,\ 0.7,\ 1\}$. Then the subintervals in P are $[0, 0.3]$, $[0.3, 0.7]$ and $[0.7, 1]$. Let $Q = \{0,\ 0.3,\ 0.5,\ 1\}$ and $R = \{0,\ 0.3,\ 0.5,\ 0.7,\ 1\}$. Then R is a refinement of P, but Q is not a refinement of P since there is a subinterval $[0.5, 1]$ in Q which is not properly contained in any subinterval of P. However, R is a refinement of Q as well.

Remark. Examples of Nets

1. Let X be a topological space and $x \in X$. Let $\mathcal{N}_x$ be the set of all neighbourhoods of x. Let $D = (\mathcal{N}_x, X)$ be the directed set given by $(N, y) \in (\mathcal{N}_x, X)$ if $N \in \mathcal{N}_x$ and $y \in N$ and $(N, y) \geq (M, z)$ if $N \subset M$. Then the function $S : (\mathcal{N}_x, X) \to X$ given by $S(N, y) = y$ is a net on X.

   For example, let $X = \{a, b, c, d\}$ and $\mathcal{T} = \{\{a\},\ \{a, b\},\ \{a, b, c\},\ \{a, b, c, d\}\}$. Also let $S : (\mathcal{N}_b, X) \to X$ defined by $S(N, y) = y$. Suppose $C = \{a, b, c\}$. Then $C \in \mathcal{N}_b$. ie, C is a neighbourhood of b. Then $S(C, c) = c$.

2. Riemann Net - Let $D = (\mathcal{P}, \xi)$ where $\mathcal{P}$ is the set of all partitions on $[0,1]$ and ξ is a finite sequence in $[0,1]$ such that consecutive terms belongs to consecutive subintervals of the partition. The set $(\mathcal{P}, \xi)$ is directed set with $\geq$ given by $(P, \eta) \geq (Q, \psi)$ iff P is a refinement of Q.

   For example, let $P \in \mathcal{P}$ is given by $P = \{\ 0,\ 0.3,\ 0.7,\ 1\ \}$ and $\eta = \{\ 0.2,\ 0.6,\ 0.9\ \}$. Then $(P, \eta) \in (\mathcal{P}, \xi)$.

   Let $f : \mathbb{R} \to \mathbb{R}$ be any function, then the function,

$$S : (\mathcal{P}, \xi) \to \mathbb{R} \text{ defined by } S(P, \eta) = \sum_{j=1}^{k} f(\eta_k)(a_k - a_{k-1})$$

   where $P = \{a_0,\ a_1, \cdots,\ a_k\ \}$ is the Riemann Net with respect to the real function f.

For example, let $f(x) = x^2$ and P, η are same as above example, then $S(P, \eta) = 0.2^2(0.3 - 0) + 0.6^2(0.7 - 0.3) + 0.9(1 - 0.7) = 3.99$

Definitions 7.4.3 (Convergence of a Net)**.** [Joshi, 1983, 10.1.3]
A net $S : D \to X$ converges to a point $x \in X$ if for any nbd U of x, there exists $m \in D$ such that $n \in D,\ n \geq m \implies S(n) \in U$. And x is a limit of the net S.

Remark. The choice of m depends on the choice of neighbourhood U

$$S : D \to X,\ S \to x \iff (\forall U \in \mathcal{N}_x,\ \exists m_U \in D,\ \text{such that } n \geq m_U \implies S(n) \in U)$$

Theorem 7.4.1 (Net characterisation of Hausdorff space)**.** *[Joshi, 1983, 10.1.4] A topological space is Hausdorff iff limits of all nets in it are unique.*

Proof. Let X be a Hausdorff space. Suppose $S : D \to X$ is net on X such that S converges to two distinct points $x, y \in X$. Since X is a Hausdorff space and $x \neq y$, there exists open subsets U, V such that $x \in U,\ y \in V, U \cap V = \phi$.

The net S converges to $x \in X$, therefore $\exists m_x \in D$ such that $n \geq m_x \implies S(n) \in U$ And, the net S converges to $y \in X$, therefore $\exists m_y \in D$ such that $n \geq m_y \implies S(n) \in V$.

Since D is a directed set and $m_x, m_y \in D$, there exists $p \in D$ such that $p \geq m_x$ and $p \geq m_y$. Now, $n \geq p \implies n \geq m_x,\ n \geq m_y$, since $\geq$ is transitive. (ie, $n \geq p,\ p \geq m_x \implies n \geq m_x$, and $n \geq p,\ p \geq m_y \implies n \geq m_y$).

We have $n \geq p \implies n \geq m_x$ and $n \geq m_x \implies S(n) \in U$. Therefore, $n \geq p \implies S(n) \in U$. Similarly, $n \geq p \implies n \geq m_y \implies S(n) \in V$. Therefore $S(n) \in U \cap V$ which is a contradiction, since $U \cap V = \phi$. Therefore, if a net S converges to two points x, y, then $x = y$. That is, if a net S in a Hausdorff space X is convergent then its limit is unique.

Conversely, suppose that X is a topological space and every convergent net in X has a unique limit. Suppose X is not a Haudorff space. Then there exists at least two distinct points $x, y \in X$ such that every neighbourhood of x intersects with every neighbourhood of y. Now consider the set $D = \mathcal{N}_x \times \mathcal{N}_y$ and relation $\geq$ on D such that $(U_1, V_1) \geq (U_2, V_2)$ if $U_1 \subset U_2$ and $V_1 \subset V_2$.

By the axiom of choice, a function $S : D \to X$ such that $S(U, V) \in U \cap V$ is well defined, since every nbd of x intersects every nbd of y. Thus, S is a net in X. We claim that S converges to both x and y.

Let U be a nbd of x. Then $S(U', V') \in U' \cap V'$. We have $(U, X) \in D$ such that $(U', V') \geq (U, X) \implies U' \subset U$. Then, $S(U', V') \in U' \cap V' \subset U \cap X = U$. Thus, for any nbd U containing x, we have $(U, X) \in D$ such that $(U', V') \geq (U, X) \implies S(U', V') \in U$. Therefore, S converges to $x \in X$.

Similarly, Let V be a nbd of y. Then for any nbd V containing y, we have $(X, V) \in D$ such that $(U', V') \geq (X, V) \implies S(U', V') \in V$, since

$S(U', V') \in U' \cap V' \subset X \cap V = V$. Therefore, S converges to $y \in X$ as well, where $x \neq y$. This is a contradition to the assumption that every convergent net in X has a unique limit. Therefore, for any two points $x, y \in X$, there should be some nbd of x that doesn't intersect some nbd of y. Therefore, X is a Hausdorff space. $\square$

Definitions 7.4.4 (Eventual Subset)**.** [Joshi, 1983, 10.1.5]
A subset E of a directed set D is an eventual subset of D if there exists $m \in D$ such that $n \geq m \implies n \in E$.

Remark. Let E be an eventual subset of D such that $n \geq m \implies n \in E$. Then $p \in E \not\Longrightarrow p \geq m$. ie, Subset E may contain elements that doesn't follow the above m.

Remark. [Joshi, 1983, 10.1.6]
Let E be an eventual subset of D, then E is a directed set.

1. $m, n, p \in E,\ m \geq n,\ n \geq p \implies m, n, p \in D,\ m \geq n,\ n \geq p \implies m \geq p$
2. $m \in E \implies m \in D \implies m \geq m$
3. $m, n \in E \implies m, n \in D \implies \exists p \in D$ such that $p \geq m$ and $\geq n$.

   $\exists m' \in D$ such that $n' \geq m' \implies n' \in E$. (E is an eventual subset of D)

   $p, m' \in D \implies \exists p' \in D$ such that $p' \geq p$ and $p' \geq m'$. (D is a directed Set)

   $p' \geq m' \implies p' \in E$. (E is eventual subset of D with respect to m')

   $p' \geq p,\ p \geq m \implies p' \geq m$ and $p' \geq p,\ p \geq n \implies p' \geq n$.

   Therefore $\forall m, n \in E,\ \exists p' \in E$ such that $p' \geq m$ and $p' \geq n$.

Definitions 7.4.5 (Net eventually in A)**.** [Joshi, 1983, 10.1.5]
Let $S : D \to X$ be a net in a topological space X. Then S is eventually in subset A of X if $S^{-1}(A)$ is an eventual subset of D.

Remark. Let $S : D \to X$ be a net in X. Then S converges to $x \in X$ if S is eventually in each nbd U of x.

Definitions 7.4.6 (Cofinal subset)**.** [Joshi, 1983, 10.1.7]
A subset F of a directed D is a cofinal subset of D if for any $m \in D$, there exists $n \in F$ such that $n \geq m$.

Remark. Let X be a topological space and $x \in X$. Let $\mathcal{N}_x$ be the set of all neighbourhood of x and $\mathcal{L}$ be a local base of X at x. We have, $(\mathcal{N}_x, \geq)$ is a directed set where $\forall U, V \in \mathcal{N}_x,\ U \geq V \iff U \subset V$, then $\mathcal{L}$ is cofinal in $\mathcal{N}_x$.

Remark. [Joshi, 1983, 10.1.8]
Let F be a cofinal subset of D, then F is a directed set.

1. $m, n, p \in F,\ m \geq n,\ n \geq p \implies m, n, p \in D,\ m \geq n,\ n \geq p \implies m \geq p$
2. $m \in F \implies m \in D \implies m \geq m$

3. $m, n \in F \implies m, n \in D \implies \exists p \in D$ such that $p \geq m$ and $\geq n$.

   E is cofinal, $p \in D \implies \exists p' \in F$ such that $p' \geq p$.

   $p' \geq p,\ p \geq m \implies p' \geq m$ and $p' \geq p,\ p \geq n \implies p' \geq n$.

   Therefore $\forall m, n \in F,\ \exists p' \in F$ such that $p' \geq m$ and $p' \geq n$.

Definitions 7.4.7 (Net frequently in A)**.** [Joshi, 1983, 10.1.7]
Let $S : D \to X$ be a net in a topological space X. Then S is frequently in subset B of X if $S^{-1}(B)$ is a cofinal subset of D.

Proposition 7.4.1. *[Joshi, 1983, 10.1.6]*
Let $S : D \to X$ be a net in a topological space X. Let E be an eventual subset of D. Then, S converges to x iff $S_{/_E}$ converges to x. cite[10.1.6]joshi

Proof. Let $S : D \to X$ be a net in X, E be an eventual subset of D, and $x \in X$. Then, $S_{/_E} : E \to X$ is defined by $n \in E \implies S_{/_E}(n) = S(n)$

Suppose S converges to x. Let U be a nbd of x, then S is eventually in U. ie, $S^{-1}(U)$ is an eventual subset of D. Then $\exists m \in D$ such that $n \geq m \implies n \in S^{-1}(U) \implies S(n) \in U$. Since set E is eventual subset of D, $\exists m' \in D$ such that $n \geq m' \implies n \in E$.

Since E is a directed set, $S_{/_E} : E \to X$ is a net in X. And $m, m' \in D \implies \exists p \in D$ such that $p \geq m$ and $p \geq m'$. We have, $p \geq m' \implies p \in E$. And $n \geq' p \implies n \geq p,\ p \geq m \implies n \geq m \implies S(n) \in U \implies S_{/_E}(n) \in U$. Therefore, $n \geq' p \implies S_{/_E}(n) \in U$. Since U is arbitrary, $S_{/_E}$ converges to x.

Conversely, suppose that $S_{/_E}$ converges to x. Let U be a nbd of x, then $S_{/_E}$ is eventually in U. ie, $S_{/_E}^{-1}(U)$ is an eventual subset of D. ie, $\exists m \in D$ such that $n \geq m \implies n \in S_{/_E}^{-1}(U) \implies S_{/_E}(n) \in U \implies S(n) \in U$. Therefore, $n \geq m \implies S(n) \in U$. Since, U is arbitrary, S converges to every nbd of x. ie, S converges to x. □

Proposition 7.4.2. *[Joshi, 1983, 10.1.8]*
Let $S : D \to X$ be a net in a topological space X. Let F be a cofinal subset of D. If S converges to x, then $S_{/_F}$ converges to x.

Proof. Let $S : D \to X$ be a net in X and S converges to $x \in X$. Also let F be a cofinal subset of D. Then $S_{/_F}$ is also a net in X, since $(F, \geq')$ is a directed set where $\forall m, n \in F,\ m \geq n \implies m \geq' n$.

Since S converges to x, for any nbd U of x, $\exists m \in D$, such that $n \geq m \implies S(n) \in U$. Since F is cofinal, $\exists p \in F$ such that $p \geq m$. Thus $n \geq' p \implies n \geq p,\ p \geq m \implies n \geq m \implies S(n) \in U \implies S_{/_F}(n) \in U$. Therefore, $\exists p \in F$ such that $n \geq' p \implies S_{/_F}(n) \in U$. Since U is arbitrary, $S_{/_F}$ is eventually in every nbd of x. ie, $S_{/_F}$ converges to x. □

Remark. But converse of the above is not true. $S_{/_F}$ converges to x does not imply that S converges to x, since cofinal subset F not necessarily contain every element following a particular m.

Definitions 7.4.8 (Cluster point)**.** [Joshi, 1983, 10.1.9]
Let $S : D \to X$ be a net in a topological space X. Then $x \in X$ is a cluster point of S, if S is frequently in each nbd U of x in X.

Proposition 7.4.3. *[Joshi, 1983, 10.1.10]*
Let $S : D \to X$ be a net in a topological space X. Then $x \in X$ is a cluster points of X, if $S_{/_F}$ converges to x for some cofinal subset F of D.

Proof. Let $S : D \to X$ be a net in X and $(F, \geq')$ be a cofinal subset of $(D, \geq)$. Then $S_{/_F}$ is also a net in X. Suppose $S_{/_F}$ converges to $x \in X$. Let U be a nbd of x, then $\exists m \in F$ such that $n \geq' m \implies S_{/_F}(n) \in U$.

Let $m' \in D$. Then $\exists p' \in F$ such that $p' \geq m'$, since F is a cofinal subset of D. We have, $m, p' \in F$, then $\exists p \in F$ such that $p \geq' m$ and $p \geq' p'$. Since $F \subset D$, we have $p, m \in F \implies p, m \in D$ and $p \geq' m \implies p \geq m$.

Also $p \geq' m \implies S_{/_F}(p) \in U \implies S(p) \in U$. Therefore, $\forall m' \in D$, $\exists p \in D$ such that $p \geq m'$ and $S(p) \in U$. Since U, m' are arbitrary, S is frequently in every nbd of x. ie, x is a cluster point of S. ☐

Definitions 7.4.9 (Subnet)**.** [Joshi, 1983, 10.1.11]
Let $S : D \to X$ be a net in a topological space X. Then a net $T : E \to X$ in X, is a subnet of S if there exists a function $N : E \to D$ such that $S \circ N = T$ and $\forall m \in D$, $\exists p \in E$ such that $n \geq' p \implies N(n) \geq m$.

Remark. A net $T : E \to X$ is a subnet of $S : D \to X$ if $\exists N : E \to D$ such that $S \circ N = T$ and S is frequently in $T(E)$.
Let $T : E \to X$ be a subnet of $S : D \to X$ and A be a subset of X. If T eventually in A, then S is frequently in A.

Proposition 7.4.4. *[Joshi, 1983, 10.1.12]*
Let $S : D \to X$ be a net in a topological space X. Then $x \in X$ is a cluster point of S iff there exists a subnet of S which converges to x.

Synopsis. Let $(D, \geq)$, $(E, \geq')$ be two directed sets. And $T : E \to X$ be a subnet of $S : D \to X$.

If T converges to x, then T is eventually in each nbd U of x. And since T is a subnet of S, there exists $N : E \to D$ such that $N(E)$ is a cofinal subset of D. Therefore, S is frequently in each nbd U of x. Thus, x is a cluster point of S.

If x is a cluster point of X, then S is frequently in every nbd of x. Let $N : E \to D$ be $N(n, U) = n$. Construct a directed subset E of $D \times \mathcal{N}_x$ such that $(n, U) \in E \iff S(n) \in U$. Now T is eventually in every nbd U of x, as those points with images outside U are removed by construction. Therefore, it is sufficient to show that $T : E \to X$ is a subnet of the net $S : D \to X$. Clearly, E is a directed set and $N : E \to D$ defined by $N(n, U) = n$ satisfies both $S \circ N = T$ and $\forall n \in D$, $\exists p \in E$ such that $m \geq p \implies N(m) \geq n$.

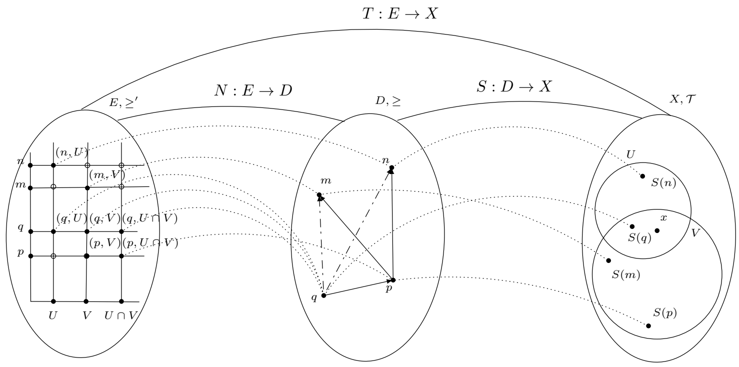


Figure 7.4: $\forall (n, U), (m, V) \in E,\ \exists (q, W) \in E$ such that $(q, W) \geq' (n, U),\ (q, W) \geq' (m, V)$

Proof. Let $S : D \to X$ be a net in X. Suppose there exists a subnet $T : E \to X$ that converges to $x \in X$. By the definition of subnet, we have $\exists N : E \to D$ such that $S \circ N = T$ and S is frequently in $T(E)$.

We have, T convergers to x, thus for any neighbourhood U of x, there exists $m' \in E$ such that $n' \geq' m' \implies T(n') \in U$.

Also we have, T is a subnet of S. Then $\exists N : E \to D$ such that $\forall m \in D$, $\exists p' \in E$ such that $n' \geq' p' \implies N(n') \geq m$.

Now, for any $m \in D$, we have $m', p' \in E$. Since E is a directed set, there exists $n' \in E$ such that $n' \geq' m'$ and $n' \geq' p'$.

Then, $n' \geq m' \implies T(n') \in U$ and $n' \geq' p' \implies N(n') \geq m$.

Thus for any $m \in D$, there exists $N(n') = n \in D$ such that $S(n) = S(N(n')) = T(n') \in U$.

Thus S is frequently in any neighbourhood U of x. Therefore, x is a cluster point of S.

Conversely, suppose that x is a cluster point of S. We have to construct a directed set $(E, \geq')$ and a function $N : E \to D$ such that T is a subnet of S and T converges to x. Let $\mathcal{N}_x$ be the family of all neighbourhood of x in X.

Consider $E = \{(n, U) \in D \times \mathcal{N}(x) : S(n) \in U\}$ and define $\geq'$ by $(n, U) \geq' (m, V)$ if $n \geq m$ and $U \subset V$. Trivially, $(n, U) \geq' (m, V) \geq' (p, W) \implies (n, U) \geq' (p, W)$ and $(n, U) \geq' (n, U)$. Also, for any $(n, U), (m, V) \in E$, we have $n, m \in D$ and $U, V \in \mathcal{N}(x)$. Since D is a directed set, $\exists p \in D$ such that $p \geq n$ and $p \geq m$. Also, $U \cap V \in \mathcal{N}_x$ and there exists $q \in D$ such that $S(q) \in U \cap V$ and $q \geq' p$, since S is frequently in every nbd of x. And $U \cap V \in \mathcal{N}(x)$ such that $U \cap V \subset U$ and $U \cap V \subset V$. Thus $\exists (q, U \cap V) \in E$ such that $(q, U \cap V) \geq' (n, U)$ and $(q, U \cap V) \geq' (m, V)$. Therefore, $(E, \geq')$ is a directed set.

Define $N : E \to D$ by $N(n, U) = n$. Again for any $(m, V) \in E$, there exists $m \in D$ such that $(n, U) \geq' (m, U)$ implies there exists $n \in D$ such that $N(n, U) = n$ and $n \geq m$.

It remains to prove that, T converges to x. Let $U \in \mathcal{N}_x$ be a nbd of x in X. We have, x is a cluster points of S. Therefore, $\forall m \in D$, $\exists p \in D$ such that $S(p) \in U$. By the constructon of E, we have $(p, U) \in E$. Suppose $(n, V) \geq' (p, U)$, then $n \geq p$, $V \subset U$, and $S(n) \in V$. Clearly $S(n) \in U$. Therefore, $\forall (n, V) \geq' (p, U)$, $T(n, V) \in U$. That is, T is eventually in every nbd of x. ie, subnet T is convergent to x. Therefore, for each cluster point of the net S, there exists some subset converging to it. □

Remark.
- Importance of Construction of E
  If x is a cluster point of a net S in X, then S is frequently in some cofinal subset of D. Thus, if we consider any cofinal subset D' of D which is a direct set with $\geq$ restricted to D'. Then $N : D' \to D$ defined by $N(n) = n$ gives a subnet $T : D' \to X$ of the net S. However, this subnet need not converge to x. The strongest statement, we can make on T is that 'x is a cluster point of T', since $N : D \times \mathcal{N}(x) \to D$, $N(n) = n$ is completely independent of U. This problem is overcame by constructing E which is dependent on each nbd U of x.
- Existence of $q \in D$ such that $q \geq' p$ and $S(q) \in U$. We have, p follows both $n\&m$ and $U \cap V$ is a subset of both $U\&V$. However, since S is only frequently in U, p not necessarily be in U. But there is always someone following p which has its image in U. This q follows both $n\&m$, since $\geq'$ is transitive.

7.5 Variations of Compactness

7.5.1 Variations of Compactness

In this chapter, we have two other notions of compactness - countable compactness and sequential compactness. ($\star$[13])

Compact A topological space is compact iff every open cover of it has a finite subcover. ([Joshi, 1983, 6.1.1]) [Heine-Borel]

Countably Compact A topological space is countably compact iff every countable, open cover of it has a finite subcover. [Joshi, 1983, 11.1.1]

Sequentially Compact A topological space is sequentially compact iff every sequence in it has a convergent subsequence. [Joshi, 1983, 11.1.8] [Bolzano-Weierstrass]

Countable compactness is a weaker notion compared to compactness. ($\star$[14]) However, sequentially compact and compact are not necessarily comparable. ($\star$[15])

[13] For $\mathbb{R}$, Compactness & Sequentially compactness are equivalent to the completeness axiom.
[14] Every compact space is countably compact.
[15] $\mathcal{T}_1, \mathcal{T}_2$ are non-comparable, if $\mathcal{T}_1 \not\subset \mathcal{T}_2$ and $\mathcal{T}_2 \not\subset \mathcal{T}_1$. [Joshi, 1983, 4.2.1]

We have seen earlier that compactness has the following properties 1. compactness is weakly hereditary. [Joshi, 1983, 6.1.10] 2. compactness is preserved under continuous functions. [Joshi, 1983, 6.1.8] 3. every continuous real functions on compact space is bounded and attains its extrema. [Joshi, 1983, 6.1.6] 4. every continuous real function on a compact, metric space is uniformly continuous by Lebesgue covering lemma. [Joshi, 1983, 6.1.7]

Countably compact spaces, Sequentially compact spaces have all the four properites listed above.

7.5.2 Countable compactness

Weakly hereditary property

A subspace $(A, \mathcal{T}_{/A})$ being countably compact doesn't imply that $(X, \mathcal{T})$ is countably compact. However, if $(X, \mathcal{T})$ is a countably compact space and A is a closed subset of X, then $(A, \mathcal{T}_{/A})$ is also a countably compact space. In other words, countably compactness is weakly hereditary.

Theorem 7.5.1. *Countable compactness is weakly hereditary. [Joshi, 1983, 11.1.3]*

Synopsis. Let A be a closed subset of countably compact space, X. If A has a countable open cover $\mathcal{U}$, then we can obtain a respective countable, open cover for X by attaching $X - A$ to the extensions of members of $\mathcal{U}$ to X. This cover has a finite subcover. Then restricting them to A, we get a finite subcover of $\mathcal{U}$.

Proof. Suppose X is a countably compact space. And A is a closed subset of X. We need to show that A is countably compact. Without loss of generality,($\star$[16]) assume that A is a proper subset of X. Then $X - A$ is a non-empty, open subset of X.

Let $\mathcal{U}$ be a countable open cover of A. Then $\mathcal{U} = \{U_1, U_2, \cdots\}$ where each element $U_k \in \mathcal{U}$ is an open subset of A. Since A is a subspace of X, every open subset U_k in A is of the form $G \cap A$ for some open subset G in X. Therefore, there exists open subsets $V(U_k)$ for each U_k such that $A \cap V(U_k) = U_k$. $\star$[17])

Define $\mathcal{V} = \{X - A, V(U_1), V(U_2), \cdots\}$. Clearly, $\mathcal{V}$ is a countable open cover ($\star$[18]) of X. We have X is countably compact, thus $\mathcal{V}$ has a finite subcover, say $\mathcal{V}'$. Without loss of generality assume ($\star$[19]) that $X - A \in \mathcal{V}'$. Suppose $X - A \notin \mathcal{V}'$, then we can define another finite subcover $\mathcal{V}' \cup \{X - A\}$. Thus $\mathcal{V}' = \{X - A,\ V(U_{n_1}),\ V(U_{n_2}), \cdots,\ V(U_{n_k})\}$.

Then the corresponding subcover $\mathcal{U}' = \{U_{n_1}, U_{n_2}, \cdots, U_{n_k}\}$ is a finite subcover of $\mathcal{U}$. Since countable open cover $\mathcal{U}$ and closed subset A are arbitrary, every closed subset of X with relative topology is countably compact. Therefore, countable compactness is weakly hereditary. □

[16]Suppose A is not a proper subset of X. Then $X = A$ and A is countably compact.
[17]Relative topology,$\mathcal{T}_{/A} = \{G \cap A : G \in \mathcal{T}\}$
[18]$X - A$ is open in X. If $y \notin A$, then $y \in X - A$. If $y \in A$, then $y \in U_k$ for some k.
[19]Otherwise, you will have to consider two cases: $X - A \in \mathcal{V}'$ and $X - A \notin \mathcal{V}'$

Remark. Proof depends on the following,

1. There is an extension map, $\psi : P(A) \to P(X)$ that preserve open subsets (and closed subsets). This ψ is an open map which not a true inverse of the restriction, $r : P(X) \to P(A)$, defined by $r(G) = G \cap A$ for every subset G of X.

2. Also we rely on the subset A being closed. Suppose X have many countable open covers, but X has only uncountable open covers corresponding to a particular countable open cover of A. In such a case, X being countably compact is insufficient for A to be countably compact.

The behaviour of countinous functions

We will now study the nature of continuous functions defined on countably compact spaces. Suppose X, Y are topological space and function $f : X \to Y$ is continuous. If X is countably compact, then $f(X)$ is also countably compact. Continuous images of countably compact spaces are countably compact. In other words, countable compactness is preserved under continuous functions.($\star$[20])

Theorem 7.5.2. *Countable compactness is preserved under continuous functions. [Joshi, 1983, 11.1.2]*

Synopsis. Let X be countably compact and $f : X \to Y$ be continuous. Suppose $\mathcal{U}$ is a countable cover of $f(X)$, then X has a countable cover $\mathcal{V}$ obtained by taking inverse images. Since X is countably compact, $\mathcal{V}$ has a finite subcover $\mathcal{V}'$. Now taking images of members of $\mathcal{V}'$, we get a finite subcover $\mathcal{U}'$ of $f(X)$.

Proof. Suppose X is a countably compact space, Y is a topological space and $f : X \to Y$ is a continuous function. Let $\mathcal{U} = \{U_1, U_2, \cdots\}$ be a countable cover of $f(X)$ by set open in $f(X)$. We have to show that $\mathcal{U}$ has a finite subcover.

Define $\mathcal{V} = \{f^{-1}(U_1), f^{-1}(U_2), \cdots\}$. Then $\mathcal{V}$ is a countable open cover of X, since $f^{-1}(U_k)$ are open subsets of X and,

$$\begin{aligned}\bigcup_{k=1}^{\infty} U_k = f(X) &\implies f^{-1}\left(\bigcup_{k=1}^{\infty} U_k\right) = X \\ &\implies \bigcup_{k=1}^{\infty} f^{-1}(U_k) = X\end{aligned}$$

We have, $\mathcal{V}$ is a countable open cover of X, which is a countably compact space. Therefore $\mathcal{V}$ has a finite subcover $\mathcal{V}' = \{f^{-1}(U_{n_1}),\ f^{-1}(U_{n_2}), \cdots,\ f^{-1}(U_{n_k})\}$.

$$\begin{aligned}\bigcup_{j=1}^{k} f^{-1}(U_{n_j}) = X &\implies f^{-1}\left(\bigcup_{j=1}^{k} U_{n_j}\right) = X \\ &\implies \bigcup_{j=1}^{k} U_{n_j} = f(X)\end{aligned}$$

[20] A topological property is preserved under continuous functions if whenever a space has that property so does every continuous image of it. [Joshi, 1983, 6.1.9]

Clearly $\mathcal{U}' = \{U_{n_1}, U_{n_2}, \cdots, U_{n_k}\}$ is a finite subcover of $\mathcal{U}$. Thus every countable open cover of $f(X)$ by sets open in $f(X)$ has a finite subcover. Therefore, continuous images of countably compact spaces are countably compact. □

Remark. 1. For a continuous function, $f : X \to Y$ the inverse images of open subsets are open in X. The relation $f^{-1} \subset f(X) \times X$ is not a function. However, we may consider a function, $\psi : P(Y) \to P(X)$ such that $\psi(U) = f^{-1}(U)$ for any subset U of Y. This ψ is an open map which maps open subsets of Y to open subsets of X.

Theorem 7.5.3. *Every continuous, real-valued function on a countably compact, metric space is bounded and attains its extrema. [Joshi, 1983, 11.1.7]*

Synopsis. Let X be a countably compact space and function $f : X \to \mathbb{R}$ be continuous. Then $f(X) \subset \mathbb{R}$ is countably compact. Real line $\mathbb{R}$ is metrisable($\star^{21}$). Then $f(X)$ is countably compact, metric space. Therefore $f(X)$ compact.($\star^{22}$). The subset $f(X)$ of $\mathbb{R}$ is bounded and closed, since every compact subset of $\mathbb{R}$ is bounded and closed. Thus $f(X)$ contains its supremum and infimum. Therefore, f is bounded and attains its extrema.

Proof. Let X be a countably compact space and $f : X \to \mathbb{R}$ be continuous, real-valued function on the countably compact space, X. We have to show that f is bounded and attains its extrema.

Since countable compactness is preserved under continuous functions, $f(X)$ is countably compact subset of $\mathbb{R}$. Since, $f(X)$ is a subset of the metric space, $\mathbb{R}$ and metrisability is hereditary, $f(X)$ is again metrisable. (suppose) We have, every countably compact, metric space is compact. Then $f(X)$ is a compact subset of $\mathbb{R}$.

Since every compact subset of $\mathbb{R}$ is bounded and closed, $f(X)$ is bounded and closed. Since every closed subset of $\mathbb{R}$ contains supremem and infimum, $f(X)$ contains its extrema. Therefore, every continuous, real-valued function on a countably compact space is bounded and attains its extrema.

We have assumed that every countably compact, metric space is compact. This result will be proved in the last section of this chapter. □

Remark. Since countably compact, metric spaces are compact. The above theorem can be used to prove that continuous, real-valued functions on a compact, metric space attains its extrema.

Due to the Lebesgue covering lemma, next result is quite simple.*

Theorem 7.5.4. *Every continuous, real-valued function on a countably compact, metric space is uniformly continuous.*

Proposition 7.5.1. *Let X be a first countable, Hausdorff space. Then every countably compact subset A of X is closed.[Joshi, 1983, Exercises 11.1.7]*

---

[21] [Joshi, 1983, 4.2 Example 4], $\mathbb{R}$ with usual metric $d : R \to R$, $d(x, y) = |x - y|$

[22] [Joshi, 1983, 11.1.11] On metric spaces, countable compactness $\implies$ compactness.

7.5.3 Sequential Compactness

Weakly hereditary property

Theorem 7.5.5. *Sequential compactness is weakly hereditary. [Joshi, 1983, Exercises 11.1.3]*

The behaviour of countinous functions

Theorem 7.5.6. *Sequential compactness is preserved under continuous functions. [Joshi, 1983, Exercises 11.1.4]*

Synopsis. Let X be sequentially compact and function $f : X \to Y$ be continuous. Then any sequence, $\{y_k\}$ in $f(X)$ has a sequence, $\{x_k\}$ in X such that $f(x_k) = y_k$. Sequence $\{x_k\}$ has a subsequence $\{x_{n_k}\}$ converging to x, then sequence $\{f(x_n)\}$ in $f(X)$ has the subsequence $\{f(x_{n_k})\}$ converging to $f(x)$.

Proof. Let X be a sequentially compact space, function $f : X \to Y$ be continuous and $\{y_n\}$ be a sequence in $f(X)$ subset of Y. Construct a sequence $\{x_n\}$ such that $f(x_k) = y_k$, $\forall k$.

Every sequence in X has a convergent subsequence. Thus $\{x_n\}$ has a subsequence $\{x_{n_k}\}$ converging to $x \in X$. The image of this subsequence $\{f(x_{n_k})\}$ is a subsequence of $\{y_k\}$. We claim that, $\{f(x_{n_k})\}$ converges to $f(x) \in f(X)$.

Let U be an open subset containing $f(x)$, then $f^{-1}(U)$ is an open subset containing x. Since $\{x_{n_k}\}$ converges to x. There exists an integer n such that for every $k \geq n$, $x_k \in f^{-1}(U)$. Clearly, for each $k \geq n$, $f(x_k) \in U$. Since U is arbitrary, $\{f(x_{n_k})\}$ converges to $f(x)$. Therefore, the image of any sequentially compact space is sequentially compact. In other words, sequentially compactness is preserved under continuous functions. □

Remark.
1. Given a sequence $\{y_n\}$ in $f(X)$, there is a sequence of subsets $\{U_n\}$ in $P(Y)$ such that $U_n = f^{-1}(y_n)$. Since each U_n is non-empty, we can construct a sequence $\{x_n\}$ in X using a choice function. The convergent subsequence of $\{y_n\}$ depends on the selection of this choice function.

Given every sequentially compact, metric space is countably compact. We may assert the properties of countably compact, metric spaces on sequentially compact, metric spaces.

Theorem 7.5.7. *Every continuous, real-valued function on a sequentially compact, metric space is bounded and attains its extrema.*

Theorem 7.5.8. *Every continuous, real-valued function on a sequentially compact, metric space is uniformly continuous. [Joshi, 1983, Exercises 11.1.6]*

7.5.4 Countable Compactness on T_1 spaces

In this section, we are going to see four different characterisations of countable compactness in T_1 spaces. The first two characterisations doesn't have anything to do with the T_1 axiom.

T_1 Space A topological space X satisfy T_1 axiom if for any two distinct points $x, y \in X$, there exists an open subset $U \subset X$ containing x but not y. [Joshi, 1983, 7.1.2]

countable compactness A topological space is countably compact if every countable open cover has a finite subcover. [Joshi, 1983, 11.1.1]

finite intersection property A family $\mathcal{F}$ of subsets of X has finite intersection property(f.i.p.) if every finite subfamily of $\mathcal{F}$ has a non-empty intersection. [Joshi, 1983, 10.2.6]

accumulation point A point $x \in X$ is accumulation point of a subset $A \subset X$ if every open subset containing x has atleast one point of A other than x. [Joshi, 1983, 5.2.7]

limit point A point $x \in X$ is a limit point of a sequence $< x_k >$ in X if for every open subset U containing x, there exists an integer $N \in \mathbb{N}$ such that $x_k \in U$ for every $k \geq N$. [Joshi, 1983, 4.1.7]

cluster point A point $x \in X$ is a cluster point of a sequence $< x_k >$ in X if for any neighbourhood V of x, the sequence $< x_k >$ assumes a point in V infinitely many times.($\star$[23])

Countable compactness in T_1 spaces

Theorem 7.5.9. *In a T_1 space X, following statements are equivalent,*

1. *X is countably compact*
2. *Every countably family of closed subsets of X with finite intersection property have non-empty intersection.*
3. *Every infinite subset $A \subset X$ has an accumulation point.($\star$[24])*
4. *Every sequence $< x_k >$ in X has a cluster point.*
5. *Every infinite open cover of X has a proper subcover.[Arens-Dugundji]*

Proof. $1 \implies 2$
Suppose X is countably compact. Let $\mathcal{C} = \{C_1, C_2, \cdots\}$ be a countable family of closed subsets of X with empty intersection. Define $\mathcal{U} = \{X - C_1, X - C_2, \cdots\}$ is a family of open subsets of X. By de Morgan's law, ($\star$[25])

$$\bigcap_{k=1}^{\infty} C_k = \phi, \text{ then } X = X - \left(\bigcap_{k=1}^{\infty} C_k\right) = \bigcup_{k=1}^{\infty} (X - C_k)$$

We have $\mathcal{U}$ is a countable cover of X and X is countably compact space. Thus $\mathcal{U}$ has a finite subcover $\mathcal{U}' = \{X - C_{n_1}, X - C_{n_2}, \cdots, X - C_{n_k}\}$.

$$\mathcal{U}' \text{ is a cover of } X, \text{ then } X = \bigcup_{j=1}^{k} \left(X - C_{n_j}\right)$$

[23] x is a cluster point of $< x_k >$ if for every integer N, there exists $k > N$ such that $x_k \in V$. In other words, $< x_k >$ is frequently in V. [Joshi, 1983, 10.1.9]

[24] Every infinite subset of $\mathbb{R}$ has a limit point is equivalent to the completeness axiom.

[25] Complement of Intersection = Union of complements, $X - (C \cap D) = (X - C) \cup (X - D)$,

$$X - \bigcup_{j=1}^{k} (X - C_{n_j}) = \bigcap_{j=1}^{k} (X - (X - C_{n_j})) = \bigcap_{j=1}^{k} C_{n_j} = \phi$$

Now $\mathcal{C}' = \{C_{n_1}, C_{n_2}, \cdots, C_{n_k}\}$ has empty intersection. This is a contradiction to the finite intersection property of $\mathcal{C}$. Thus $\mathcal{C}$ has non-empty intersection. Therefore, every countably family of closed subsets of X have non-empy intersection.

$2 \implies 1$

Let $\mathcal{U} = \{U_1, U_2, \cdots\}$ be a countable cover of X. Then $\mathcal{C} = \{X - U_1, X - U_2, \cdots\}$ is a countable family of closed subsets of X.

Let $\mathcal{U}' = \{U_{n_1}, U_{n_2}, \cdots, U_{n_k}\}$ be any finite subfamily of $\mathcal{U}$. Suppose X is not countably compact, then $\mathcal{U}$ doesn't have a finite subcover. Therefore, $\mathcal{U}'$ is not a cover of X. And $\mathcal{C}$ is a family of closed subsets with finite intersection property.

Therefore by assumption, the countable family of closed subsets $\mathcal{C}$ has a non-empty intersection.

$$\bigcap_{k=1}^{\infty} C_k \neq \phi, \text{ then } \bigcap_{k=1}^{\infty} C_k = \bigcap_{k=1}^{\infty} (X - U_k) = X - \left(\bigcup_{k=1}^{\infty} U_k \right) \neq \phi$$

Then $\mathcal{U}$ is not a cover of X as well. This is a contradiction, therefore X is countably compact.

$1 \implies 3$

Suppose X is countably compact. Let A be an infinite subset of X. Suppose A doesn't have an accumulation point.

Let B be a countably infinite subset of A. Then B also doesn't have any accumulation point. Therefore, the derived set B' is empty. Thus B is a closed subset of X. Since countable compactness is weakly hereditary, subspace B is again countably compact.

For each point $b \in B$, there is an open subset V_b such that $V_b \cap B = \{b\}$, since $b \in B$ is not an accumulation point. Thus $\mathcal{U} = \{V_b \cap B : b \in B\}$ is a countable open cover of B. Clearly, $\mathcal{U}$ doesn't have any finite subcover.

This is a contradiction to B being countably compact. Therefore, A has an accumulation point. □

7.5.5 Variations of Compactness on Metric Spaces

In this document, we will see that from metric space point of view these two notions were equivalent to the compactness and were used alternatively. For example : in functional analysis (semester 3), you will find definitions like 'a normed space is compact iff every sequence in it has a convergent subsequence', which is clearly sequential compactness for a topologist.

Lindeloff A topological space is Lindeloff iff every open cover has a countable subcover.

First countable A topological space is first countable iff every point in it has a countable local base.

Second countable A topological space is second countable iff it has a countable base.

Base A family of subsets $\mathcal{B}$ of X is a base of a topological space if every open subset can be expressed as union of some members of $\mathcal{B}$

Base Characterisation A family of subsets $\mathcal{B}$ of X is a base of a topological space iff for every $x \in X$, and for every neighbourhood U of x, there is a member $B \in \mathcal{B}$ such that $x \in B \subset U$.

Local Base A family of subsets $\mathcal{L}$ of X is a local base at point $x \in X$ if for every neighbourhood U of x, there is a member $L \in \mathcal{L}$ such that $x \in L \subset U$.

Equivalence

We are going to see when these three notions: compactness, countable compactness and sequentially compactness are equivalent.

Theorem 7.5.10. *Countably compact, metric spaces are second countable.*

Synopsis. For every positive real number r, there exists a non-empty maximal subsets A_r with every pair of points atleast r distance apart. A_r are finite. The union of maximal subsets $A_{\frac{1}{n}}$ for each natural number n is a countable, dense subset D of X. Thus countably compact, metric spaces are separable. The family $\mathcal{B}$ of all open balls with center at $d \in D$ and rational radius is a countable, base for X. Thus countably compact, metric spaces are second countable.

Proof. Let $(X; d)$ be a countably compact,, metric space. For each positive real number $r \in \mathbb{R},\ r > 0$ construct a family of subsets $A_r \subset X$ such that it is a maximal set of points which are atleast r distances apart.

Then A_r is finite for every positive real number r. Suppose A_r is infinite for some real number $r > 0$, then A_r has a accumulation point, say x by the Characterisation of countable compactness of X.

Then every neighbourhood of x must intersect A_r at infinitely many points, since every metric space is a T_1 space. Consider $B(x, \frac{r}{2})$. Since any two points of $B(x, \frac{r}{2})$ are less than r distances apart, the intersection $B(x, \frac{r}{2}) \cap A_r$ can have atmost one point in it. Thus for every positive real number r, A_r is finite.

Define $D = \cup_{n=1}^{\infty} A_{\frac{1}{n}}$. We claim that D is a countable, dense subset of X.

Let $x \in X$ and $B(x, r)$ be an open ball containing x, then there exists integer $n \in \mathbb{N}$ such that $\frac{1}{n} < r$. $\star$[26])

[26] By archimedean property of integers, we have $\forall r \in \mathbb{R},\ r > 0,\ \exists n \in \mathbb{N}$ such that $nr > 1$.

Then $B(x,r) \cap D \neq \phi$, since $B(x,r) \cap A_{\frac{1}{n}} \neq \phi$. Suppose $B(x,r) \cap A_{\frac{1}{n}} = \phi$, then $A_{\frac{1}{n}}$ is not maximal. Since, x is atleast $r > \frac{1}{n}$ distance apart from each points of $A_{\frac{1}{n}}$. Therefore, D intersects with every open subset and thus dense in X.

We have have a countable, dense subset D of X. Therefore, X is separable. Now define $\mathcal{B} = \{B(x,r) : r \in \mathbb{Q},\ x \in D\}$. Clearly, $\mathcal{B}$ is a countable base for X. By the construction of $\mathcal{B}$, X is second countable.($\star$[27]) □

Countable Compactness, Lindeloff $\iff$ Compactness

Theorem 7.5.11. *A topological space X is compact iff it is countably compact, Lindeloff space.*

Proof. Let X be a compact space. Let $\mathcal{U}$ be a countable open cover of X, then $\mathcal{U}$ has a finite subcover $\mathcal{U}'$. Therefore, every compact space is countably compact.($\star$[28])

Conversely, suppose X is a countably compact, Lindeloff space. Since X is Lindeloff, every open cover $\mathcal{U}$ has a countable subcover $\mathcal{U}'$. Since X countably compact, every countable open cover $\mathcal{U}'$ has a finite subcover $\mathcal{U}''$. Thus every open cover $\mathcal{U}$ has a finite subcover $\mathcal{U}''$. Therefore every countably compact, Lindeloff space is compact. □

Countable Compactness, First Countable $\implies$ Seq. Compactness

Theorem 7.5.12. *Every countably compact, first countable space is Sequentially compact.*

Proof. Let X be a countably compact, first countable space. Let $\{x_n\}$ be a sequence in X. By, equivalent conditions($\star$[29]) of countably compact spaces, every sequence in countably compact space X has a cluster point, say x. We have, X is first countable. Therefore, X has a countable local base $\mathcal{L}$ at $x \in X$. How to construct a subsequence of $\{x_n\}$ converging to x ?($\star$[30]) □

Remark. Every sequentially compact space is countably compact.$\star$

Theorem 7.5.13. *In a second countable space, all the three forms of compactness are equivalent. [Joshi, 1983, 11.1.10]*

Proof. Every second countable space is both first countable and Lindeloff. Every countably compact, Lindeloff space is countably compact. Therefore every countably compact, second countable space compact. Again, every countably compact, first countable space is sequentially compact. Therefore every countably compact, second countable space is sequentially compact. Conversely, every compact space is countably compact and every sequentially compact space is countably compact.($\star$[31]) □

[27] Every separable, metric space is second countable.
[28] Countable compactness is a weaker notion than compactness.
[29] [Joshi, 1983, 11.1] Conditions 1,2, and 4 are equivalent. 2 $\implies$ 4 without T_1 axiom is out of scope.
[30] [Joshi, 1983, Exercises 10.1.11]
[31] Countable compactness is a weaker notion than sequential compactness as well.

Theorem 7.5.14. *In a metric space, all the three forms of compactness are equivalent. [Joshi, 1983, 11.1.11]*

Proof. In a metric space each form of compactness implies second countability. And in second countable spaces, they are all equivalent. □

7.6 Homotopy of Paths

Definitions 7.6.1. Let X, Y be topological spaces and $f : X \to Y$, $f' : X \to Y$ be continuous functions. Then f, f' are homotopic if there exists a continuous function $F : X \times I \to Y$ such that for every $x \in X$, $F(x, 0) = f(x)$ and $F(x, 1) = f'(x)$. And we write, $f \simeq f'$.

Definitions 7.6.2. Let X be a topological space and $f : I \to X$ and $f' : I \to X$ be two paths. Then f, f' are path-homotopic if they have same initial point x_0 (ie, $x_0 = f(0) = f'(0)$), same final point x_1 (ie, $x_1 = f(1) = f'(1)$) and they are homotopic (ie, $\exists F : I \times I \to X$ such that $\forall x \in I$, $F(x, 0) = f(x)$ and $F(x, 1) = f'(x)$ also fixed at the end points x_0 and x_1(ie, $\forall t \in I$, $F(0, t) = x_0$ and $F(1, t) = x_1$). And we write $f \simeq_p f'$.

Remark. If two paths f, $f'$are homotopic, then they have the same end points and there exists a (topologically) continuous deformation from one path into another.

Proposition 7.6.1. *The relations $\simeq$, $\simeq_p$ are equivalence relations.*

Proof. Homotopy : Let f, f' be continuous functions from X into Y. Then f and f' are homotopic, $f \simeq f' \iff \exists F : X \times I \to Y$ such that F is continuous, $F(x, 0) = f(x)$, and $F(x, 1) = f'(x)$

1. $f \simeq f$
   We have $f : X \to Y$ is continuous. Define $F : X \times I \to Y$ such that $F(x, t) = f(x)$. Clearly, F is continuous, $F(x, 0) = f(x)$ and $F(x, 1) = f(x)$. And $\exists F : X \times I \to Y \implies f \simeq f$.

2. $f \simeq f' \implies f' \simeq f$
   We have, $f \simeq f'$. Thus there exists a continuous function $F : X \times I \to Y$ such that $F(x, 0) = f(x)$ and $F(x, 1) = f'(x)$.
   Consider $F' : X \times I \to Y$ defined by $F'(x, t) = F(x, 1 - t)$. Clearly, F' is continuous, $F'(x, 0) = F(x, 1) = f'(x)$, and $F'(x, 1) = F(x, 0) = f(x)$. Thus, $\exists F'(x, t) : X \times I \to Y \implies f' \simeq_p f$

3. $f \simeq f'$, $f' \simeq f'' \implies f \simeq f''$
   We have, $f \simeq f' \iff \exists F : X \times I \to Y$ such that F is continuous, $F(x, 0) = f(x)$ and $F(x, 1) = f'(x)$.

   Similarly, $f' \simeq f'' \iff \exists F' : X \times I \to Y$ such that F' is continuous, $F'(x, 0) = f'(x)$ and $F'(x, 1) = f''(x)$.

   Consider $G : X \times I \to Y$ defined by

$$G(x, t) = \begin{cases} F(x, 2t) & , t \in [0, \frac{1}{2}] \\ F'(x, 2t - 1) & , t \in [\frac{1}{2}, 1] \end{cases}$$

We have, $G(x, \frac{1}{2}) = F(x, 1) = F'(x, 0) = f'(x)$. Thus, G is continuous by pasting lemma since $[0, \frac{1}{2}] \cap [\frac{1}{2}, 1] = \{\frac{1}{2}\}$. Also $G(x, 0) = F(x, 0) = f(x)$ and $G(x, 1) = F'(x, 1) = f''(x)$. Thus, $\exists G : X \times I \to Y \implies f \simeq f''$

Path Homotopy : Let $f,\ f',\ f''$ be paths in X. Then f and f' are path homotopic, $f \simeq_p f' \iff \exists F : I \times I \to X$ such that F is continuous, $\forall s \in [0, 1],\ F(s, 0) = f(s),\ F(s, 1) = f'(s)$ and $\forall t \in [0, 1],\ F(0, t) = f(0) = f'(0),\ F(1, t) = f(1) = f'(1)$

1. $f \simeq_p f$
   We have $f : I \to X$ is continuous. Define $F : I \times I \to X$ such that $\forall s, t \in [0, 1],\ F(s, t) = f(s)$. Clearly, F is continuous, $\forall s \in [0, 1],\ F(s, 0) = f(s),\ F(s, 1) = f(s)$ and $\forall t \in [0, 1],\ F(0, t) = f(0),\ F(1, t) = f(1)$. Thus, $\exists F : I \times I \to X \implies f \simeq_p f$.

2. $f \simeq_p f' \implies f' \simeq_p f$
   We have, $f \simeq_p f'$. Thus there exists a continuous function $F : I \times I \to X$ such that $\forall s \in [0, 1],\ F(s, 0) = f(s),\ F(s, 1) = f'(s)$ and $\forall t \in [0, 1],\ F(0, t) = f(0) = f'(0),\ F(1, t) = f(1) = f'(1)$.

   Consider $F' : I \times I \to X$ defined by $F'(s, t) = F(s, 1 - t)$. Clearly, F' is continuous. And $F'(s, 0) = F(s, 1) = f'(s)$, and $F'(s, 1) = F(s, 0) = f(s)$. Also, $F'(0, t) = F(0, 1 - t) = f(0) = f'(0)$ and $F'(1, t) = F(1, 1 - t) = f(1) = f'(1)$. Thus, $\exists F'(s, t) : I \times I \to X \implies f' \simeq_p f$

3. $f \simeq f',\ f' \simeq f'' \implies f \simeq f''$
   We have, $f \simeq f' \iff \exists F : I \times I \to X$ such that F is continuous, $\forall s \in [0, 1],\ F(s, 0) = f(s),\ F(s, 1) = f'(s)$ and $\forall t \in [0, 1],\ F(0, t) = f(0) = f'(0),\ F(1, t) = f(1) = f'(1)$

   Similarly, $f' \simeq f'' \iff \exists F' : I \times I \to X$ such that F' is continuous, $\forall s \in [0, 1],\ F'(s, 0) = f'(s),\ F'(s, 1) = f''(s)$ and $\forall t \in [0, 1],\ F'(0, t) = f'(0) = f''(0),\ F'(1, t) = f'(1) = f''(1)$

   Consider $G : I \times I \to X$ defined by

$$G(s, t) = \begin{cases} F(s, 2t) & , t \in [0, \frac{1}{2}] \\ F'(s, 2t - 1) & , t \in [\frac{1}{2}, 1] \end{cases}$$

   We have, $G(s, \frac{1}{2}) = F(s, 1) = F'(s, 0) = f'(s)$. Thus, G is continuous by pasting lemma[Munkres, 2003, §18.3 pp. 106], since $[0, \frac{1}{2}] \cap [\frac{1}{2}, 1] = \{\frac{1}{2}\}$.

   Also $G(s, 0) = F(s, 0) = f(s)$ and $G(s, 1) = F'(s, 1) = f''(s)$.

   Again, $\forall t \in [0, \frac{1}{2}],\ G(0, t) = F(0, 2t) = f(0) = f'(0) = f''(0)$ and $\forall t \in [\frac{1}{2}, 1],\ G(0, t) = F'(0, 2t - 1) = f(0) = f'(0) = f''(0)$. Therefore, $\forall t \in [0, 1],\ G(0, t) = f(0) = f''(0)$.

Similarly, $\forall t \in [0, \frac{1}{2}],\ G(1,t) = F(1,2t) = f(1) = f'(1) = f''(1)$ and $\forall t \in [\frac{1}{2}, 1],\ G(1,t) = F'(1, 2t-1) = f(1) = f'(1) = f''(1)$. Therefore, $\forall t \in [0,1],\ G(1,t) = f(1) = f''(1)$. Thus, $\exists G : I \times I \to X \implies f \simeq_p f''$

$\square$

Definitions 7.6.3. Let f be a path in X (ie, $f : I \to X$), then $[f]$ is the equivalence class of all paths homotopic to f in X. (ie, $g \in [f] \iff f \simeq_p g$)

Remark. The set of homotopy classes of functions from X into Y is dentoed by $[X, Y]$. And, the set of all path-homotopic classes on X is denoted by $[I, X]$.

Remark (Straight-line homotopy). [Munkres, 2003, §51 Example 1 pp. 320] Let X be a topological space, and $f,\ g$ be continuous functions from X into a eucilidean space, say $\mathbb{R}^2$. Then $f,\ g$ are straight line homotopic if there exists a continuous function F from $X \times I$ such that F deforms f into g along straight line segments joining them.

For example, $F(x,t) = (1-t)f(x) + tg(x)$.

Remark. Let A be a convex subspace of $\mathbb{R}^n$. Then any two paths in A from x_0 to x_1 are path homotopic in A.

Proof. —continue page 321— $\square$

Remark. [Munkres, 2003, §51 Example 2 pp. 321] This demonstrates that the straight-line homotopy is very sensitive to the holes in the space.

Definitions 7.6.4. Let $f,\ g$ be two paths in X (ie, $f : I \to X$ and $g : I \to X$) such that $f(0) = x_0$, $f(1) = g(0) = x_1$ and $g(1) = x_2$. Then the prduct $h = f * g$ is given by $h : I \to X$ and

$$h(s) = \begin{cases} f(2s) & , s \in [0, \frac{1}{2}] \\ g(2s-1) & , s \in [\frac{1}{2}, 1] \end{cases}$$

This h is well-defined, and continuous by pasting lemma.($\star$[32])

Definitions 7.6.5. The product operation on path-homotopy classes is defined by $[f] * [g] = [f * g]$.

Remark. The product of path-homotopic classes is well-defined.

Proof. Let F be a path-homotopy between $f,\ f' \in [f]$ and G be a path-homotopy between $g,\ g' \in [g]$. Then $H : I \times I \to X$ defined by

$$H(s,t) = \begin{cases} F(2s,t) & s \in [0, \frac{1}{2}] \\ G(2s-1,t) & s \in [\frac{1}{2}, 1] \end{cases}$$

Then H is well-defined, and continuous by pasting lemma.

[32]Pasting Lemma : Let $X = A \cup B$, where A and B are closed in A. Let $f : A \to Y$ and $g : B \to Y$ be continuous. If $f(x) = g(x)$ for every $x \in A \cup B$, then f and g combine to give a continuous function $h : X \to Y$, defined by setting $h(x) = f(x)$ if $x \in A$ and $h(x) = g(x)$ if $x \in B$.

$\forall s \in [0, \frac{1}{2}],\ H(s, 0) = F(2s, 0) = f(2s)$ and
$\forall s \in [\frac{1}{2}, 1],\ H(s, 0) = G(2s - 1, 0) = g(2s - 1)$.
$\implies H(s, 0) = (f * g)(s)$, by the definition of $f * g$

$\forall s \in [0, \frac{1}{2}],\ H(s, 1) = F(2s, 1) = f'(2s)$ and
$\forall s \in [\frac{1}{2}, 1],\ H(s, 1) = G(2s - 1, 1) = g'(2s - 1)$.
$\implies H(s, 1) = (f' * g')(s)$, by the definition of $f' * g'$

$H(0, t) = F(0, t) = f(0) = x_0 = (f * g)(0)$, and
$H(1, t) = G(1, t) = g'(1) = x_2 = (f' * g')(1)$

Then $H : I \times I \to X$ is a path-homotopy between $f * g$ and $f' * g'$. □

Definitions 7.6.6 (Groupoid). Let G be a set and $*$ be a binary operation on G. Then $(G, *)$ is a groupoid if it satisfies the following axioms

g1 Associativity - $\forall x, y, z \in G,\ (x * y) * z = x * (y * z)$

g2 Existence of left and right identies - There exist unique elements e_L and e_R such that $\forall x \in G,\ x * e_R = x$ and $e_L * x = x$.

g3 Existence of inverse
$\forall x \in G,\ \exists x^{-1} \in G$ such that $x * x^{-1} = e_L$ and $x^{-1} * x = e_R$

Definitions 7.6.7 (Positive Linear Map). A positive liear map $p : [a, b] \to [c, d]$ is the unique map of the form $p(x) = mx + k$ such that $p(a) = c$ and $p(b) = d$. Clearly, scaling factor, $m = \frac{d-c}{b-a}$ as we want to transform an interval of length $b - a$ into an interval of length $d - c$. And offset k is given by,

$$p(a) = \frac{d - c}{b - a} a + k = c \implies k = c - \frac{a(d - c)}{b - a} = \frac{bc - ad}{b - a}$$

But, we won't fix m and k in $p(x) = mx+k$, instead we will focus on the unique map with graph of positive slope and passing through required end points. The graph of a positive linear map from $[a, b]$ to $[c, d]$ is always a straight-line with positive slope.

Remark. The inverse of a positive linear map is also a positive linear map. Given $p : [a, b] \to [c, d], p(x) = mx + k,$ where $m = \frac{d-c}{b-a},\ k = \frac{bc-ad}{b-a}$. Then it's inverse, $\bar{p} : [c, d] \to [a, b]$ is given by $p(y) : m'y+k',$ where $m' = \frac{b-a}{d-c} = \frac{1}{m},\ k' = \frac{ad-bc}{d-c} = \frac{-k}{m}$. Clearly $m > 0 \implies m' = \frac{1}{m} > 0$.

Remark. The composite of two positive linear maps is also a (piece-wise) positive linear map. Let $f,\ g$ be two positive linear maps. Then their composite map $f * g$ is given by

$$(f * g)(x) = \begin{cases} f(2x) & x \in [a, \frac{a+b}{2}] \\ g(2\left(x - \frac{b-a}{2}\right)) & x \in [\frac{a+b}{2}, b] \end{cases}$$

Remember $f * g$ exists only if $f(b) = g(a)$. Therefore, $f * g$ is a well-defined, continuous (by pasting lemma) and (piecewise) positive linear map.

Lemma 7.6.1. *Let $f,\ f'$ be two paths in X and $k : X \to Y$ be a continuous function. Let F be the path homotopy in X between the paths f and f'. Then $k \circ F$ is a path homotopy in Y between that paths $k \circ f$ and $k \circ f'$ That is, path homotopy is preserved under a continuous function.*

Lemma 7.6.2. *Let $f,\ g$ be two paths in X with $f(1) = g(0)$ and $k : X \to Y$ be a continous function. Then $k \circ (f * g) = (k * f) \circ (k * g)$*

Theorem 7.6.3. *Let f, g, h be paths in a topological space X, and $[f], [g], [h]$ be respective path-homotopy classes. Suppose the operation product, $*$ is defined by*

$$[f] * [g] = [f *' g] \text{ where } (f *' g)(s) = \begin{cases} f(2s) & s \in [0, \frac{1}{2}] \\ g(2s-1) & s \in [\frac{1}{2}, 1] \end{cases}$$

Then the product, $*$ has the following properties :

1. *Associativity*
$$\forall [f], [g], [h] \in [I, X],\ ([f] * [g]) * [h] = [f] * ([g] * [h])$$
2. *Existence of left and right identies*
Let $e_x : I \to X$ defined by $\forall s \in [0, 1],\ e_x(s) = x$. Let f be a path from x_0 to x_1, then there exist unique paths e_{x_0} and e_{x_1} such that $[f] * [e_{x_1}] = [f]$ and $[e_{x_0}] * [f] = [f]$. That is, $e_{x_0},\ e_{x_1}$ are respectively the left and right path-homotopy-identities.
3. *Existence of inverse*
Let f be a path in X. The path, $\bar{f}$, defined by $\bar{f}(s) = f(1-s)$ is the reverse path of f. Then $[f] * [\bar{f}] = [e_{x_0}]$ and $[\bar{f}] * [f] = [e_{x_1}]$. That is, the inverse of class of f is the class of reverse path of f.

In other words, Set of all path-homotopy classes together with binary operation product, $*$ is a groupoid. ie, $([I, X], *)$ is a groupoid.

Proof. Step 1 : Properties 2&3

Let $e_0 : I \to I$ such that $e_0(t) = 0,\ \forall t \in I$. And $i : I \to I$ such that $i(t) = t,\ \forall t \in I$. Then $e_0 * i$ is also a path. Since I is convex, there is a path homotopy($\star$[33]) G between i and $e_0 * i$. Let $f : I \to X$ be continuous path in X from x_0 to x_1. Then $f \circ G$ is a path homotopy (by Lemma 2) in X between $f \circ i$ and $f \circ e_0 * i$ where $f \circ i$ and $f \circ e_0$ are paths from x_0 to x_1 in X.

$$\begin{aligned} f \circ (e_0 * i) &= (f \circ e_0) * (f \circ i), \text{ by Lemma 1} \\ &= e_{x_0} * f, \text{ since } \forall s \in I,\ f(e_0(s)) = x_0 = e_{x_0}(s) \text{ and } f * i \simeq_p f \end{aligned}$$

Therefore $[e_{x_0}] * [f] \simeq_p [f]$, since $e_0 * i \simeq_p i$, and $f \circ (e_0 * i) \simeq_p f \circ i = f$.

Similarly, $e_1 : I \to I$ such that $e_1(t) = 1$. Let H be a path homotopy($\star$[34]) between $i * e_1$ and i. Thus, $f \circ H$ is a path homotopy in X from $f \circ (i * e_1)$ and $f \circ i$.

$$\begin{aligned} f \circ (i * e_1) &= (f \circ i) * (f \circ e_1), \text{ by Lemma 1} \\ &= f * e_{x_1}, \text{ since } f * i \simeq_p f,\ i * e_1 \simeq_p e_1,\ \forall s \in I,\ (f(e_1(s)) = x_1 = e_{x_1}(s) \end{aligned}$$

[33] $G : I \times I \to I,\ G(s, 0) = i(s),\ G(s, 1) = (e_0 * i)(s),\ G(0, t) = 0,\ G(1, t) = 1.$
[34] $H : I \times I \to I,\ H(s, 0) = (i * e_1)(s),\ H(s, 1) = i(s),\ H(0, t) = 0,\ H(1, t) = 1$

Since $i * e_1 \simeq_p i$, we have $f \circ (i * e_1) \simeq_p f \circ i = f$. Therefore $[f] * [e_{x_1}] \simeq_p [f]$. Thus, $[f] * [e_{x_1}] \simeq_p [f] \simeq_p [e_{x_0}] * [f]$. Therefore, $[f]$ has left and right inverses ie, property 2 holds.

Consider inverse path $\bar{i} : I \to I,\ \bar{i}(s) = 1 - s$. Then $i * \bar{i}$ is a path in I with both end points at 0. We have, $e_0 : I \to I,\ e_0(s) = 0$ is also a path with both end points at 0. Since I is convex, there is a path homotopy H in I between e_0 and $i * \bar{i}$. Then $f \circ H$ is a path homotopy between $f \circ e_0 = e_{x_0}$ and $f \circ (i * \bar{i}) = (f \circ i) * (f \circ \bar{i}) = f * \bar{f}$. Therefore, $[e_{x_0}] \simeq_p [f] * [\bar{f}]$.

Similarly $\bar{i} * i$ and e_1 are paths with both end points at 1. Since I is convex, there is a path homotopy G in I between $\bar{i} * i$ and e_1. Then $f \circ G$ is a path homotopy between $f \circ (\bar{i} * i) = (f \circ \bar{i}) * (f \circ i) = \bar{f} * f$ and $f \circ e_1 = e_{x_1}$. Therefore, $[\bar{f}] * [f] \simeq_p [e_{x_1}]$. Thus the path $\bar{f} : I \to X,\ \bar{f}(s) = f(1 - s),\ \forall s \in I$ is reverse of f. Also $[f] * [\bar{f}] = [e_{x_0}]$ and $[\bar{f}] * [f] = [e_{x_1}]$. ie, property 3 holds.

Step 2 : Property 1

Let f, g, h be three paths in X and $f(1) = g(0) = x_1$ and $g(1) = h(0) = x_2$. Then $f * (g * h)$ is defined by

$$(g * h)(s) = \begin{cases} g(2s) & s \in [0, \frac{1}{2}] \\ h(2s - 1) & s \in [\frac{1}{2}, 1] \end{cases}$$

$$\begin{aligned}(f * (g * h))(s) &= \begin{cases} f(2s) & s \in [0, \frac{1}{2}] \\ (g * h)(2s - 1) & s \in [\frac{1}{2}, 1] \end{cases} \\ &= \begin{cases} f(2s) & s \in [0, \frac{1}{2}] \\ g(2(2s - 1)) & s \in [\frac{1}{2}, \frac{3}{4}] \\ h(2(2s - 1) - 1) & s \in [\frac{3}{4}, 1] \end{cases}\end{aligned}$$

Similarly, $(f * g) * h$ is defined by,

$$(f * g)(s) = \begin{cases} f(2s) & s \in [0, \frac{1}{2}] \\ g(2s - 1) & s \in [\frac{1}{2}, 1] \end{cases}$$

$$\begin{aligned}((f * g) * h)(s) &= \begin{cases} (f * g)(2s) & s \in [0, \frac{1}{2}] \\ h(2s - 1) & s \in [\frac{1}{2}, 1] \end{cases} \\ &= \begin{cases} f(2(2s)) & s \in [0, \frac{1}{4} \\ g(2(2s - 1)) & s \in [\frac{1}{4}, \frac{1}{2}] \\ h(2s - 1) & s \in [\frac{1}{2}, \frac{3}{4}] \end{cases}\end{aligned}$$

Clearly, $f*(g*h)$ and $(f*g)*h$ are distinct path with common endpoints. ie, $(f*(g*h))(0) = f(0) = ((f*g)*h)(0)$. And $(f*(g*h)(1) = h(1) = ((f*g)*h)(1)$.

We need to define a path homotopy G between $f * (g * h)$ and $(f * g) * h$. Let $[a, b], [c, d] \subset I$. Consider the path $p : I \to I$ defined by the following three unique($\star$[35]) positive linear maps $p_1 : [0, a] \to [0, c]$, $p_2 : [a, b] \to [c, d]$ and $p_3 : [b, 1] \to [d, 1]$.

$$p(t) = \begin{cases} p_1(t) & t \in [0, a] \\ p_2(t) & t \in [a, b] \\ p_3(t) & t \in [b, 1] \end{cases}$$

We can easily construct, a path homotopy P between identity map $i : I \to I,\ i(s) = s$ and p as follows :

$$P(s, t) = \begin{cases} t + (p_1(t) - t)\frac{s}{a} & s \in [0, a] \\ t + (p_2(t) - t)\frac{(s-a)}{(b-a)} & s \in [a, b] \\ t + (p_3(t) - t)\frac{(s-b)}{(1-b)} & s \in [b, 1] \end{cases}$$

Therefore, we have $f * (g * h) \simeq_p i$ since there exists a path homotopy P corresponding to $[a, b] = [\frac{1}{2}, \frac{3}{4}]$ and $[c, d] = [x_1, x_2]$. Similarly $(f * g) * h \simeq_p i$ since there exists a path homotopy P where $[a, b] = [\frac{1}{4}, \frac{1}{2}]$ and $[c, d] = [x_1, x_2]$. ie, $[f * (g * h)] \simeq_p [(f * g) * h]$. □

Theorem 7.6.4. *Let f be a path in X, and $a_0, a_1, \cdots, a_n$ be numbers such that $0 = a_0 < a_1 < \cdots < a_n = 1$. Let $f_i : I \to X$ be the path that equals the positive linear map of I onto $[a_{i-1}, a_i]$ followed by f. Then $[f] = [f_1] * [f_2] * \cdots [f_n]$. In other words, every path is path-homotopic to a piecewise-linear path.*

Proof. Let f be a piece-wise positive linear map such that

$$f(t) = \begin{cases} f_1(t) & t \in [0 = a_0, a_1] \\ f_2(t) & t \in [a_1, a_2] \\ \vdots & \vdots \\ f_n(t) & t \in [a_{n-1}, a_n] \end{cases}$$

where $f_i : I \to [a_{i-1}, a_i]$ such that $f_i(t)$ are a positive linear maps.

Consider the path $p : I \to I$ defined by the unique positive linear maps on the subintervals of any partition $\{0 = x_0, x_1, \cdots, x_n\}$ of I.
ie, $0 = x_0 < x_1 < \cdots < x_n = 1$

$$p_1 : [x_0, x_1] \to [a_0, a_1]$$
$$p_2 : [x_1, x_2] \to [a_1, a_2]$$
$$\vdots$$
$$p_n : [x_{n-1}, x_n] \to [a_{n-1}, a_n]$$

$$\text{Define, } p(t) = \begin{cases} p_1(t) & t \in [x_0, x_1] \\ p_2(t) & t \in [x_1, x_2] \\ \vdots & \vdots \\ p_n(t) & t \in [x_{n-1}, x_n] \end{cases}$$

[35] $p_1(t) = \frac{ct}{a}$, $p_2(t) = \frac{(d-c)t}{b-a} + \frac{bc-ad}{b-a}$, $p_3(t) = \frac{(1-d)t}{1-b} + \frac{d-b}{1-b}$

Then there exists a path homotopy P between identity map $i : I \to I,\ i(t) = t$ and p given by

$$P(s,t) = \begin{cases} t + (p_1(t) - t)\frac{a_1}{x_1} & s \in [0, x_1] \\ t + (p_2(t) - t)\frac{s - x_1}{x_2 - x_1} & s \in [x_1, x_2] \\ \vdots & \\ t + (p_n(t) - t)\frac{s - x_{n-1}}{x_n - x_{n-1}} & s \in [x_{n-1}, x_n] \end{cases}$$

Since any product of $f_1,\ f_2,\ \cdots,\ f_n$ is a path p for some partition decided by the order of associativity. This partition can be constructed as follows :
Let the last product operation (by associtivity) corresponds to $\frac{1}{2}$. The expression on its left corresponds to $[0, \frac{1}{2}]$ and expression on the right corresponds to $[\frac{1}{2}, 1]$. If there are any operations on any of these parts, the last operation (by associtivity) in them corresponds to the midpoint the respective subinterval and so on.

For examples : Consider, $(f_1 * (f_2 * f_3)) * (f_4 * f_5)$. Suppose we number the operations, $(f_1 *_1 (f_2 *_2 f_3)) *_3 (f_4 *_4 f_5)$. Then we have, $*_3 \to \frac{1}{2} \implies *_1 \to \frac{1}{4} \implies *_2 \to \frac{3}{8}$. Again $*_3 \to \frac{1}{2} \implies *_4 \to \frac{3}{4}$. Thus, we have $\{0, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{3}{4}, 1\}$.

Thus given two paths f and f' with distinct order of associativity of n paths : $f_1,\ f_2,\ \cdots,\ f_n$. We have path homotopy $P,\ P'$ given by the $P(s,t)$ for the respective partition constructed according to the order of associativity. Then, we have $f \simeq_p i$ and $f' \simeq_p i$. Thus, irrespective of the order of associativity all these paths are path homotopic. ie, $[f] = [f_1] * [f_2] \cdots [f_n]$ $\square$

Subject 8

ME010203 Numerical Analysis with Python

8.1 Interpolation and Curve Fitting

Definitions 8.1.1. Given $(n+1)$ data points $(x_k, y_k),\ k = 0, 1, \cdots, n$, the problem of estimating $y(x)$ using a function $y : \mathbb{R} \to \mathbb{R}$ that satisfy the data points is the interpolation problem. ie, $y(x_k) = y_k,\ k = 0, 1, \cdots, n$.

Definitions 8.1.2. Given $(n+1)$ data points $(x_k, y_k),\ k = 0, 1, \cdots, n$, the problem of estimating $y(x)$ using a function $y : \mathbb{R} \to \mathbb{R}$ that is sufficiently close to the data points is the curve-fitting problem.
ie, Given $\epsilon > 0,\ |y(x_k) - y_k| < \epsilon,\ k = 0, 1, \cdots, n$.

Remark. The data could be from scientific experiments or computations on mathematical models. The interpolation problem assumes that the data is accurate. But, curve-fitting problem assumes that there are some errors involved which are sufficiently small.

Definitions 8.1.3. Given $(n+1)$ data points $(x_k, y_k),\ k = 0, 1, \cdots, n$, the problem of estimating $y(x)$ using a polynomial function of degree n that satisfy the data points is the polynomial interpolation problem.

Remark. Polynomial is the 'simplest' interpolant. [Kiusalaas, 2013, 3.2]

8.2 Polynomial Interpolation

There exists a unique polynomial of degree n that satisfy $(n+1)$ distinct data points. There are a few methods to find this polynomial : 1. Lagrange's method 2. Newton's method.

8.2.1 Lagrange's Method

Interpolation polynomial [1] is given by,

$$P(x) = \sum_{i=0}^{n} y_i l_i(x), \text{ where } l_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_i}{x_j - x_i} \tag{8.1}$$

Remark. Lagrange's cardinal functions l_i, are polynomials of degree n and

$$l_i(x_j) = \delta_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}$$

Proposition 8.2.1. *Error in polynomial interpolation is given by*

$$f(x) - P(x) = \frac{(x - x_0)(x - x_1)\cdots(x - x_n)}{(n+1)!} f^{(n+1)}(\xi) \tag{8.2}$$

where $\xi \in (x_0, x_n)$

Remark. The error increases as x moves away from the unknown value ξ.

8.2.2 Newton's Method

The interpolation polynomial is given by,

$$P(x) = a_0 + a_1(x - x_0) + \cdots + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \tag{8.3}$$

where $a_i = \nabla^i y_i,\ i = 0, 1, \cdots, n$.

Remark. For Newton's Method, usually it is assumed that $x_0 < x_1 < \cdots < x_n$.

Remark. Lagrange's method is conceptually simple. But, Newton's method is computationaly more efficient than Lagrange's method.

Computing coefficients a_i of the polynomial

The coefficients are given by,

$$a_0 = y_0,\ a_1 = \nabla y_1,\ a_2 = \nabla^2 y_2,\ a_3 = \nabla^3 y_3, \cdots, a_n = \nabla^n y_n \tag{8.4}$$

Remark. The divided difference $\nabla^i y_i$ are computed as follows:

$$\begin{array}{lll}
\nabla y_1 = \dfrac{y_1 - y_0}{x_1 - x_0} & & \\
\nabla y_2 = \dfrac{y_2 - y_1}{x_2 - x_1} & \nabla^2 y_2 = \dfrac{\nabla y_2 - \nabla y_1}{x_2 - x_1} & \\
\nabla y_3 = \dfrac{y_3 - y_2}{x_3 - x_2} & \nabla^2 y_3 = \dfrac{\nabla y_3 - \nabla y_2}{x_3 - x_2} & \nabla^3 y_3 = \dfrac{\nabla^2 y_3 - \nabla^2 y_2}{x_3 - x_2}
\end{array}$$

Remark. Practise Problems
Find interpolation polynomial for the following data points :

[1] Using P_n to represent some polynomial of degree n. It is quite a confusing a notation when it comes to Newton's method as author construct a psuedo-recursive definition.

1. $\{(0,7),\ (2,11),\ (3,28)\}$
   Ans : $5x^2 - 8x + 7$
   [Kiusalaas, 2013, Example 3.1]

2. $\{(-2,-1),\ (1,2),\ (4,59),\ (-1,4),\ (3,24),\ (-4,-53)\}$
   Ans : $x^3 - 2x + 3$
   [Kiusalaas, 2013, Example 3.2]

3. $\{(-1.2,-5.76),\ (0.3,-5.61),\ (1.1,-3.69)\}$
   Ans : $x^2 + x - 6$
   [Kiusalaas, 2013, Problem Set 3.1.1]

4. $\{(-3,0),\ (2,5),\ (-1,-4),\ (3,12),\ (1,0)\}$
   Ans : $x^2 + 2x - 3$
   [Kiusalaas, 2013, Problem Set 3.1.7]

5. $\{(0,1.225),(3,0.905),(6,0.652)\}$
   Ans : $0.0037x^2 - 0.1178x + 1.225$
   [Kiusalaas, 2013, Problem Set 3.1.9]

Remark. In Lagrange's Method, we can interpolate at the given point even without computing the polynomial. In Newton's method, we have to compute polynomial and then interpolate for the given point.

That is, evaluate the value of cardinal polynomials at the point and substitute in Equation 8.1 as shown in Section 3.2. [Kiusalaas, 2013, Example 3.1]

8.2.3 Implementation of Newton's Method

Program 8.2.1. *Computing Coefficients*

```
def coefficients(xData,yData):
  m = len(xData)
  a = yData.copy()
  for k in range(1,m):
    a[k:m] = (a[k:m]-a[k-1])/(xData[k:m]-xData[k-1])
  return a
```

Line 1 `def coefficients(xData,yData):`
Defines a function which takes two arguments/parameters, named $xData$ and $yData$. In [Kiusalaas, 2013, 3.2], you will find coeffts which I have changed to coefficients. $xData$,$yData$ are numpy array objects. $xData$

x_0	y_0				
x_1	y_1	∇y_1			
x_2	y_2	∇y_2	$\nabla^2 y_2$		
$\dots$	$\dots$	$\dots$	$\dots$	$\ddots$	
x_n	y_n	∇y_n	$\nabla^2 y_n$	$\dots$	$\nabla^n y_n$

Table 8.1: The $\nabla^i y_i$ Computation Table

is a array with values $x_0, x_1, \cdots, x_n$. And $yData$ is array with values $y_0, y_1, \cdots, y_n$. For example, the value of x_3 can be accessed as $xData[3]$.

Line 2 `m = len(xData)`
The functon $len()$ is extended by numpy to give the length of array objects. In this context, $len(xData)$ will return the value $n+1$, since there are $n+1$ values in $xData$ array.

Line 3 `a = yData.copy()`
We need a copy of $yData$ to work with. Unlike other programming languages like java, in python $a = yData$ will assign a new label a to the same memory location and manipulating a will corrupt the original data in $yData$ as well. In order to avoid this, we are **making a copy of the array object using the array method provided by the numpy library.**

Line 4 `for k in range(1,m):`
This is a python loop statement. This ask python interpreter to repeat the following sub-block $m-1$ times. [2] In this context, Line 5 will be executed n times, since the $range(1, m)$ object is a list-type object with values $1, 2, \cdots, m-1$. And intepreter executes Line 5 for each values in the $range()$ object, ie, $k = 1, 2, \cdots, m-1$ before interpreting Line 6.

Line 5 `a[k:m] = (a[k:m]-a[k-1])/(xData[k:m]-xData[k-1])`
This is very nice feature available in python. **This statement, evaluates $m-k$ values in a single step. ie, $a[k], a[k+1], \cdots, a[m]$. This calculation corresponds to subsequent columns of the divided difference table, that we are familiar with.** For example, executing Line 5 with $k = 3$ is same as evaluating the $\nabla^3 y_j$ column. Note that the value $a[0]$ is never updated and similarly $a[2]$ changes when Line 5 is executed with $k = 1, 2$. From column 3 onward, $a[2]$ is not updated. Therefore, **after completing nth executing of the Line 5, we have $a[0] = y_0,\ a[1] = \nabla y_1,\ a[2] = \nabla^2 y_2, \cdots,\ a[n] = \nabla^n y_n$.**

Line 6 `return a`
This returns the array a which is the array of coefficients.

The logic of this program is in Line 4 and Line 5. So they need more explanation/understanding than anything else.

Program 8.2.2. *Interpolating using Newton's Method*

```
def interpolate(a,xData,x):
  n = len(xData)-1
  p = a[n]
  for k in range(1,n+1):
    p = a[n-k]+(x-xData[n-k])*p
  return p
```

The logic this program is in Line 3, Line 4 and Line 5.

[2] Python block is a group of statement with same level of indentation. A sub-block is a block with an additional indentation.

Line 3 : We initialize the polynomial with the coefficient $a[n] = \nabla^n y_n = a_n$.

Line 4 : We are going define the polynomial recursively. This takes exactly n *steps further. So we use a loop which repeats* n *times.*

Line 5 : The value of p *and* k *changes each time Line 5 is executed. Let* P_j *be the value in* p *after executing Line 5 with* $k = j$*. Then,*
$P_0 = p = a[n]$
$P_1 = a[n-1] + (x - x_{n-1})P_0$
$P_2 = a[n-2] + (x - x_{n-2})P_1$
$\vdots$
$P_n = a[0] + (x - x_0)P_{n-1}$. *Clearly,* P_n *is the unique* n *degree polynomial given by the Newton's method.*

Program 8.2.3. *How to interpolate ?*

```
from numpy import array
xData = array([-2,1,4,-1,3,-4])
yData = array([-1,2,59,4,24,-53])
a = coefficients(xData,yData)
print(interpolate(a,xData,2))
```

You will have to define both the functions (coefficients, interpolate) before doing this.

Line 1 `from numpy import array`
For defining array objects, we need to import them from numpy library.

Line 2 `xData = array([-2,1,4,-1,3,-4])`
You can change this line according to the first component of the given data points.

Line 3 `yData = array([-1,2,59,4,24,-53])`
You can change this line according to the second component of the given data points.

Line 4 `a = coefficients(xData,yData)`
Call function `coefficients` and store the array returned into a

Line 5 `print(interpolate(a,xData,2))`
Call function `interpolate` to interpolate at $x = 2$ and print the value $P(2)$

Program 8.2.4 (Just for Fun)**.** *We can do more using sympy !*

```
from numpy import array}
from sympy import Symbol
xData = array([-2,1,4,-1,3,-4])
yData = array([-1,2,59,4,24,-53])
a = coefficients(xData,yData)
x = Symbol('x')
p = interpolate(a,xData,x)
p.subs({x:2})
```

Remark. Programming Problems

1. $\{(0.15, 4.79867),\ (2.30, 4.49013),\ (3.15, 4.2243),\ (4, 85, 3.47313),$
   $(6.25, 2.66674),\ (7.95, 1.51909)\}$ [Kiusalaas, 2013, Example 3.4]

2. $\{(0, -0.7854),\ (0.5, 0.6529),\ (1, 1.7390),\ (1.5, 2.2071),\ (2, 1.9425)\}$
   [Kiusalaas, 2013, Problem Set 3.1.5]

8.2.4 Limitations of Polynomial Interpolation

1. Inaccuracy - The error in interpolation increases as the point moves away from most of the data points.

2. Oscilation - As the number of data points considered for polynomial interpolation increases, the degree of the polynomial increases. And the graph of the interpolant tend to oscilate excessively. In such cases, the error in interpolation is quite high.

3. The best practice is to consider four to six data points nearest to the point of interest and ignore the rest of them.

Remark. The interpolant obtained by joining cubic polynomials corresponding to four nearest data points each, is a cubic spline[3].

8.3 Roots of a Function

Definitions 8.3.1. Let $f : \mathbb{R} \to \mathbb{R}$, then $x \in \mathbb{R}$ is a root of f if $f(x) = 0$.

Remark. Suppose $a < b$ and $f(a), f(b)$ are nonzero and are of different signs. If f is continuous in $[a, b]$, then there is a point $c \in [a, b]$ such that $f(c) = 0$.

Thus given $a < b$ and $f(a), f(b)$ are nonzero values of different sign, then there may be a bracketed root in $[a, b]$.

Note : There is no guarantee that there exists a root in $[a, b]$ as we are not sure about the continuity of f.

Remark. Given a bracketed root, we can find it using

1. Bisection Method or

2. Newton-Raphson Method

8.3.1 Bisection Method

Suppose $a < b$ and $f(a), f(b)$ are nonzero values of different signs. We evaluate $f(c)$ where $c = \frac{a+b}{2}$. If $f(c)$ is a nonzero value, then at least one of the pairs $f(a), f(c)$ or $f(c), f(b)$ are of different signs. WLOG suppose that $f(a), f(c)$ are of different signs, then set $b = c$ and $c = \frac{a+b}{2}$. And continue this process until we get sufficiently accurate value of a root.

[3]Cubic spline is a function, the graph of which is piece-wise cubic

Remark. Suppose $f(x) = x^5 - 2$. Then $f(0) = -2,\ f(1) = -1,\ f(2) = 30$. Since f is known to be continuous, there is a bracketed root in $[1, 2]$. Now
$f(1.5) > 0 \implies [1, 1.5]$
$f(1.25) > 0 \implies [1, 1.25]$
$f(1.125) < 0 \implies [1.125, 1.25]$
$f(1.1825) > 0 \implies [1.125, 1.1825]$
$f(1.15375) > 0 \implies [1.125, 1.15375]$
$f(1.139375) < 0 \implies [1.139375, 1.15375]$
$f(1.1465625) < 0 \implies [1.1465625, 1.15375]$
$f(1.150156250) > 0 \implies [1.1465625, 1.15015625]$
$f(1.148359375) < 0 \implies [1.1483594, 1.15015625]$
$f(1.149257825) > 0 \implies [1.1483594, 1.14925783]$

Thus, we have 1.14 is a root of f with accuracy upto two decimal points.

8.3.2 Newton-Raphson Method

Suppose f is differentiable at $x \in \mathbb{R}$ and $f(x) \neq 0$. Then compute $x = x - \frac{f(x)}{df(x)}$ and evaluate $f(x)$. Repeat this process to get more accurate value of a root near x.

Remark. Suppose $f(x) = x^5 - 2$. Then $df(x) = 5x^4$. Let $x = 2$. Then

$$\begin{aligned}
x = 2 - \frac{30}{80} &\implies f(1.625) = 9.330 \\
x = 1.625 - \frac{9.330}{34.86} &\implies f(1.35735) = 2.6074 \\
x = 1.35735 - \frac{2.6074}{16.9721} &\implies f(1.20373) = 0.52733 \\
x = 1.20373 - \frac{0.52733}{10.4975} &\implies f(1.15351) = 0.04224 \\
x = 1.15351 - \frac{0.042245}{8.85225} &\implies f(1.148738) = 0.00034312
\end{aligned}$$

Thus we have 1.1487 is quite close to a root of f.

8.4 Matrix Operations

Consider a system of n linear, simultaneous equations in n unknowns,

$$\begin{aligned}
A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n &= b_1 \\
A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n &= b_2 \\
&\vdots \\
A_{n1}x_1 + A_{n2}x_2 + \cdots + A_{nn}x_n &= b_n
\end{aligned}$$

We may represent them using matrices as $Ax = b$. That is,

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Gauss Elimination and Doolittle Decomposition are two methods to solve sytem of equaitons, $Ax = b$.

8.5 Gauss elimination method

Gauss elimination method has of two phases 1. elimination and 2. back substitution. In elimination phase, system $Ax = b$ is transformed into an equivalent system $Ux = c$ where U is an upper-triangular[4] matrix. And in back substitution phase, $Ux = c$ is solved. Since $Ax = b$ and $Ux = c$ are equivalent, they have the same solution x.

8.5.1 Elimination Phase

We can eliminate unknowns from an equation by adding a scalar multiple of an equation to another equation of the system. In matrices, this is equivalent to adding a scalar multiple of one row to another row, say $R_i \leftarrow R_i + \lambda R_k$.

$$\begin{array}{c} A_{k1}x_1 + A_{k2}x_2 + \cdots + A_{kn}x_n = b_k + \\ \lambda\left(A_{i1}x_1 + A_{i2}x_2 + \cdots + A_{in}x_n = b_i\right) \\ \hline (A_{k1} + \lambda A_{i1})x_1 + (A_{k2} + \lambda A_{i2})x_2 + \cdots + (A_{kn} + \lambda A_{in})x_n = b_k + \lambda b_i \end{array}$$

$$\begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kn} \\ \vdots & \vdots & \ddots & \vdots \\ A_{i1} & A_{i2} & \cdots & A_{in} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \xrightarrow{R_i \leftarrow R_i + \lambda R_k} \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kn} \\ \vdots & \vdots & \ddots & \vdots \\ A_{i1} + \lambda A_{k1} & A_{i2} + \lambda A_{k2} & \cdots & A_{in} + \lambda A_{kn} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$$\begin{bmatrix} \vdots \\ b_k \\ \vdots \\ b_i \\ \vdots \end{bmatrix} \xrightarrow{R_i \leftarrow R_i + \lambda R_k} \begin{bmatrix} \vdots \\ b_k \\ \vdots \\ b_i + \lambda b_k \\ \vdots \end{bmatrix}$$

[4]upper triangular - all the entries below the main diagonal are zero. ie $U_{ij} = 0, \quad \text{if } i < j$

8.5.2 Back substitution

Let $Ux = c$ be a system of n linear equations in n unknowns and U is an upper triangular matrix. Then we can solve the system of equations from the back.

$$\begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,n-1} & U_{1,n} \\ 0 & U_{2,2} & \cdots & U_{2,n-1} & U_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & U_{n-1,n-1} & U_{n-1,n} \\ 0 & 0 & \cdots & 0 & U_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{n-1} \\ c_n \end{bmatrix}$$

$$U_{n,n}\ x_n = c_n \implies x_n = \frac{c_n}{U_{n,n}}$$
$$\sum_{i=n-1}^{n} U_{n-1,i}\ x_i = c_{n-1} \implies x_{n-1} = \frac{c_{n-1} - U_{n-1,n}x_n}{U_{n-1,n-1}}$$
$$\cdots$$
$$\sum_{i=1}^{n} U_{1,i}\ x_i = c_1 \implies x_1 = \frac{c_1 - \sum_{i=2}^{n} U_{1,i}x_i}{U_{1,1}}$$

8.5.3 Illustrative example

Consider the following system of linear equations,

$$4x_1 - 2x_2 + x_3 = 11$$
$$-2x_1 + 4x_2 - 2x_3 = -16$$
$$x_1 - 2x_2 + 4x_3 = 17$$

We may represent the above system of linear equations using matrices,

$$\begin{bmatrix} 4 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 11 \\ -16 \\ 17 \end{bmatrix}$$

Phase 1 : Elimination Process

Using eq.1, the unknown x_1 is eliminated from all subsequent equations. An equivalent operation can be performed on both the matrices A and b by adding a suitable scalar multiples of row R_1 to row R_2 and R_3.

$$\begin{bmatrix} 4 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 4 \end{bmatrix} \xrightarrow[R_3 \leftarrow R_3 - 0.25R_1]{R_2 \leftarrow R_2 + 0.5R_1} \begin{bmatrix} 4 & -2 & 1 \\ 0 & 3 & -1.5 \\ 0 & -1.5 & 3.75 \end{bmatrix}$$
$$\begin{bmatrix} 11 \\ -16 \\ 17 \end{bmatrix} \xrightarrow[R_3 \leftarrow R_3 - 0.25R_1]{R_2 \leftarrow R_2 + 0.5R_1} \begin{bmatrix} 11 \\ -10.5 \\ 14.25 \end{bmatrix}$$

And using eq.2, x_2 is eliminated from all subsequent equations(only those rows below it). Again, we perform this by adding suitable scalar multiples of row 2 to row R_3.

$$\begin{bmatrix} 4 & -2 & 1 \\ 0 & 3 & -1.5 \\ 0 & -1.5 & 3.75 \end{bmatrix} \xrightarrow{R_3 \leftarrow R_3 + 0.5R_2} \begin{bmatrix} 4 & -2 & 1 \\ 0 & 3 & -1.5 \\ 0 & 0 & 3 \end{bmatrix}$$

$$\begin{bmatrix} 11 \\ -16 \\ 17 \end{bmatrix} \xrightarrow{R_3 \leftarrow R_3 + 0.5R_2} \begin{bmatrix} 11 \\ -10.5 \\ 9 \end{bmatrix}$$

The elimination process is complete when all entries below the diagonal elements are reduced to zero. ie, upper triangular matrix.

Phase 2 : Back substitution Process

The unknowns are easily found from the equations by solving them in the reverse order. The unknowns are solved from the bottom and solved variables are used to solve the remain unknowns.

$$\begin{bmatrix} 4 & -2 & 1 & 11 \\ 0 & 3 & -1.5 & -10.5 \\ 0 & 0 & 3 & 9 \end{bmatrix} \to \begin{cases} 4x_1 - 2x_2 + \quad x_3 = 11 \\ \qquad 3x_2 - \quad 1.5x_3 = -10.5 \\ \qquad\qquad 3x_3 = 9 \end{cases}$$

$$\begin{aligned} x_3 &= \frac{9}{3} = 3 \\ x_2 &= \frac{-10.5 + 1.5x_3}{3} = -2 \\ x_1 &= \frac{11 - x_3 + 2x_2}{4} = 1 \end{aligned}$$

Remark. Why don't they use row-reduced echelon matrix of A to simplify the back substitution phase ?

This doesn't have much advantage from algorithmic point of view. That is, the time complexity (number of steps for computation) is unaffected. And algorithms always prefer methods even with slight advantage in time or memory. And they won't consider complications in the manual execution of the method. Therefore, programmers won't consider alternate algorithm for the sake of computational simplicity.

8.5.4 Python : Gauss elimination method

Program 8.5.1 (Gauss elimination).

```
from numpy import dot
def gaussElimination(a,b):
  n = len(b)
  for k in range(0,n-1):
    for i in range(k+1,n):
      if a[i,k] != 0.0:
        lam = a[i,k]/a[k,k]
        a[i,k+1:n] = a[i,k+1:n]-lam*a[k,k+1:n]
        b[i] = b[i]-lam*b[k]
  for k in range(n-1,-1,-1):
    x[k] = (b[k]-dot(a[k,k+1:n],x[k+1:n]))/a[k,k]
```

```
    return b
```

Line 1 `from numpy import dot`
Imports the "dot()" function for numpy arrays which takes two 'numpy arrays' as input arguments and returns the dot product of them.

Line 2 `def gaussElimination(a,b):`
it defines "gaussElimination()" as a function which takes two arguments (inputs). First argument is the coefficient matrix A and second argument is the constant matrix b of the linear system of the form $Ax = b$.

Line 3 $n = len(b)$
it assigns the length of the list b into variable n which is obviously the number of equations.

Line 4 `for` k `in range(`0`,`$n-1$`):`
it is a loop construct. Five instructions following it are part of this loop body, which are executed for each values of k ie, $k = 0, 1, \cdots, n-1$. For each value of k, the unknown x_{k+1} is selected for elimination process.

Line 5 `for i in range(`$k+1$`,`n`):`
it is a loop inside another loop. Four instructions following it are part of this loop body, which are executed for each values of i, ie, $i = k+1, k+2, \cdots, n$. This eliminates x_{k+1} from all the equations after the $k+1$th equation of the system. Value of $i+1$ is the equation[5] from which x_{k+1} is eliminated.

Line 6 `if` $a[i,k]! = 0.0$`:`
If $a[i,k] = A_{i+1,k+1} \neq 0$, then those three instruction following it are executed. Otherwise, it skips the execution of those three statements. If $x_{k+1} = x[k]$ is not there in the ith equation, it doesn't need to be eliminated.

Line 7 $lam = a[i,k]/a[k,k]$
In this step, λ is computed so that equ.$(i+1)$ - λ equ.$(k+1)$ doesn't have x_{k+1} term in it.

Line 8 $a[i, k+1:n] = a[i, k+1:n] - lam \times a[k, k+1:n]$
Coefficients of $(i+1)$th equation are updated.
Equivalent to $a[i, 0:n] = a[i, 0:n] - lam \times a[k, 0:n]$, since zeroes need not be substracted. This is same as equ.$(i+1) \leftarrow$ equ.$(i+1)$ - λ equ.$(k+1)$

Line 9 $b[i] = b[i] - lam * b[k]$
The same row operations are performed on the matrix b instead of using an augmented matrix.

Line 10 `for` k `in range(`$n-1$`,`-1`,`-1`):`
This is another loop construct. The following statement is executed n times for values of $k = n-1, n-2, \cdots, 0$. Value of $k+1$ gives the unknown x_{k+1} which is solved by the back substitution process.

[5]Python starts counting from zero. For example : $A_{11} = a[0,0]$, $x_1 = x[0]$ and $b_1 = b[0]$

Line 11 $x[k] = (b[k] - dot(a[k, k+1:n], x[k+1:n]))/a[k, k]$
This is the back substitution process. After elimination phase we have k equation in the form $A_{k,k}x_k + A_{k,k+1}x_{k+1} + \cdots + A_{k,n}x_n = b_k$. And we already have values of $x_{k+1},\ x_{k+2},\ \cdots,\ x_n$. Then

$$x_k = \frac{b_k - (A_{k,k+1}x_{k+1} + A_{k,k+2}x_{k+2} + \cdots)}{A_{k,k}}$$

This is equivalent to

$$b_k \leftarrow \frac{b_k - \begin{bmatrix} A_{k,k+1} & A_{k,k+2} & \cdots A_{k,n} \end{bmatrix} \begin{bmatrix} x_{k+1} \\ x_{k+2} \\ \vdots \\ x_n \end{bmatrix}}{A_{k,k}}$$

Remember : The values of x_k are updated into b_k as they are computed. Thus $x_k, x_{k+1}, \cdots, x_n$ are stored in b for next back substitution ie, for evaluating x_{k-1}. We start with x_{n-1}, as $x_n = b_n$ is already solved.

Line 12 `return b`
It returns the new b matrix as output of the "gaussElimination()" function where $x_k = b_k,\ \forall k$.

8.6 LU Decomposition Method : Doolittle

Let $Ax = b$ be a linear system of n equations in n unknowns and let $A = LU$ for some lower triangular matrix L and upper triangular matrix U. Then we have $LUx = Ly = b$ where $y = Ux$. There are two phases for this method : 1. LU decomposition and 2. substitution.

First, we compute L and U such that $A = LU$ using Gauss elimination. Then We can solve $Ly = b$ using forward substitution process and then solve $Ux = y$ using back substitution process.

For Doolittle decomposition, we prefer to write A as a product LU as shown below:

$$A = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ L_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{n,1} & L_{n,2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,n} \\ 0 & U_{2,2} & \cdots & U_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & U_{n,n} \end{bmatrix}$$

$$A = \begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,n} \\ L_{2,1}U_{1,1} & L_{2,1}U_{1,2} + U_{2,2} & \cdots & L_{2,1}U_{1,n} + U_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n,1}U_{1,1} & L_{n,1}U_{1,2} + L_{n,2}U_{2,2} & \cdots & \sum_{k=1}^{n-1} L_{n,k}U_{k,n} + U_{n,n} \end{bmatrix}$$

Note that in Doolittle's decomposition method, the diagonal entries of the lower triangular matrix L are all 1. ie, $L_{ii} = 1,\ \forall i$. Thus, we can use an $n \times n$ matrix to represent both L and U by overwriting trivial entries(zeroes and ones) of both the matrices. And this matrix is reprensented by $[L\backslash U]$. [6]

$$[L\backslash U] = \begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,n} \\ L_{2,1} & U_{2,2} & \cdots & U_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n,1} & L_{n,2} & \cdots & U_{n,n} \end{bmatrix}$$

is the combined matrix made from both the triangular matrices L and U.

The triangular matrices L and U such that $LU = A$ can be computed the variables in the Gauss elimination method.

$$A = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ L_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{n,1} & L_{n,2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,n} \\ 0 & U_{2,2} & \cdots & U_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & U_{n,n} \end{bmatrix}$$

We can break down this matrix multiplication into the following row operations on the rows of the upper triangular matrix[7]

$$\begin{aligned} &U_{R1} \leftarrow U_{R1} \\ &U_{R2} \leftarrow L_{2,1} \cdot U_{R1} + U_{R2} \\ &\quad \cdots \\ &U_{Rn} \leftarrow L_{n,1} \cdot U_{R1} + L_{n,2} \cdot U_{R2} + \cdots + L_{n,n-1} \cdot U_{R(n-1)} + U_{Rn} \end{aligned}$$

Clearly, λ we use to eliminate x_k from row i are $L_{i,k}$. And the matrix obtained after Gauss elimination is the upper triangular matrix U.

8.6.1 Illustrative example

$$\text{Solve} \begin{bmatrix} -3 & 6 & -4 \\ 9 & -8 & 24 \\ -12 & 24 & -26 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -3 \\ 65 \\ -42 \end{bmatrix}$$

Phase 1 : LU Decomposition

Suppose, we have a system of three linear equations, then

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{bmatrix} \begin{bmatrix} U_{1,1} & U_{1,2} & U_{1,3} \\ 0 & U_{2,2} & U_{2,3} \\ 0 & 0 & U_{3,3} \end{bmatrix}$$

$$A = \begin{bmatrix} U_{1,1} & U_{1,2} & U_{1,3} \\ L_{2,1}U_{1,1} & L_{2,1}U_{1,2} + U_{2,2} & L_{2,1}U_{1,3} + U_{2,3} \\ L_{3,1}U_{1,1} & L_{3,1}U_{1,2} + L_{3,2}U_{2,2} & L_{3,1}U_{1,3} + L_{3,1}U_{2,3} + U_{3,3} \end{bmatrix}$$

[6] algorithmic implementation all decomposition algorithms prefer to use a combined matrix

[7] U_{Rk} : kth row of the matrix U

We can compute L and U using the Gauss elimination process [8] The matrix obtained after Gauss elimination on A is U and the values of the variable lam used in Gauss elimination are the entries in L. That is, in order to eliminate x_k from row i, we use $lam = L_{i,k}$.

$$\text{Given, } A = \begin{bmatrix} -3 & 6 & -4 \\ 9 & -8 & 24 \\ -12 & 24 & -26 \end{bmatrix}$$

$$\begin{bmatrix} -3 & 6 & -4 \\ 9 & -8 & 24 \\ -12 & 24 & -26 \end{bmatrix} \xrightarrow[R_3 \leftarrow R_3 - 4R_1]{R_2 \leftarrow R_2 + 3R_1} \begin{bmatrix} -3 & 6 & -4 \\ 0 & 10 & 12 \\ 0 & 0 & -10 \end{bmatrix} \implies L_{2,1} = 3, \; L_{3,1} = -4$$

We store these non-trivial entries of L into A itself.
That is, $A_{2,1} = L_{2,1}, \; A_{3,1} = L_{3,1}$.

" In this case, $A_{3,2}$ became zero (this is not a trivial zero yet), and we won't eliminate x_2 from row 3 to save computation time. Thus, we are not computing $L_{3,2} = 0$ or storing it. However, the variable representing $L_{3,2}$ is $A_{3,2}$, which is already zero after Gauss elimination and we are quite happy with that. "

Clearly, $L_{3,2} = 0$. Therefore, we have

$$U = \begin{bmatrix} -3 & 6 & -4 \\ 0 & 10 & 12 \\ 0 & 0 & -10 \end{bmatrix}, \; L = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix}$$

Since we are already stored those two non-trivial entries of L into A. We get,

$$[L \backslash U] = \begin{bmatrix} -3 & 6 & -4 \\ 3 & 10 & 12 \\ -4 & 0 & -10 \end{bmatrix}$$

Phase 2 : Substitution

Suppose $Ly = b$,

$$\begin{bmatrix} 1 & 0 & 0 \\ L_{2,1} & 1 & 0 \\ L_{3,1} & L_{3,2} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \rightarrow \begin{cases} y_1 & = b_1 \\ L_{2,1}y_1 + y_2 & = b_2 \\ L_{3,1}y_1 + L_{3,2}y_2 + y_3 & = b_3 \end{cases}$$

Now, we can find the values of y_k and store them into the matrix b itself.

$$\begin{aligned} & b_1 \leftarrow y_1 = b_1 \\ & b_2 \leftarrow y_2 = b_2 - \begin{bmatrix} L_{2,1} \end{bmatrix} \begin{bmatrix} b_1 \end{bmatrix}, \text{ since } b_1 = y_1 \\ & b_3 \leftarrow y_3 = b_3 - \begin{bmatrix} L_{3,1} & L_{3,2} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \text{ since } b_1 = y_1, \; b_2 = y_2 \end{aligned}$$

[8] We usually need a proof for such a strong statement. In this paper, they are more focussed on the application side and therefore we will don't present any vigorous proof.

In general,

$$b_k \leftarrow y_k = b_k - \begin{bmatrix} L_{k,1} & L_{k,2} & \cdots & L_{k,k-1} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k - 1 \end{bmatrix}, \text{ since } b_j = y_j,\ j = 1, 2, \cdots, (k-1)$$

We have $A = LU \implies LUx = b$. Suppose $Ux = y$, then we get $Ly = b$. First of all, we will solve $Ly = b$ using forward substitution.

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -3 \\ 65 \\ -42 \end{bmatrix} \to \begin{cases} y_1 & = -3 \\ 3y_1 + y_2 & = 65 \\ -4y_1 + y_3 & = -42 \end{cases}$$

$$\begin{aligned} y_1 &= -3 \\ y_2 &= 65 - 3y_1 = 74 \\ y_3 &= -42 + 4y_1 = 54 \end{aligned}$$

Suppose $Ux = y$,

$$\begin{bmatrix} U_{1,1} & U_{1,2} & U_{1,3} \\ 0 & U_{2,2} & U_{2,3} \\ 0 & 0 & U_{3,3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Now, we can find the values of x_k and store them into the matrix y itself.

$$\begin{aligned} y_3 &\leftarrow x_3 = \frac{y_3}{U_{3,3}} \\ y_2 &\leftarrow x_2 = \frac{y_2 - \begin{bmatrix} y_3 \end{bmatrix} \begin{bmatrix} U_{2,3} \end{bmatrix}}{U_{2,2}}, \text{ since } y_3 = x_3 \\ y_1 &\leftarrow x_1 = \frac{y_1 - \begin{bmatrix} y_2 \\ y_3 \end{bmatrix} \begin{bmatrix} U_{1,2} & U_{1,3} \end{bmatrix}}{U_{1,1}}, \text{ since } y_2 = x_2,\ y_3 = x_3 \end{aligned}$$

In general,

$$x_k = \frac{y_k - \begin{bmatrix} y_{k+1} \\ y_{k+2} \\ \vdots \\ y_n \end{bmatrix} \begin{bmatrix} U_{k,k+1} & U_{k,k+2} & \cdots & U_{k,n} \end{bmatrix}}{U_{k,k}}, \text{ since } y_j = x_j,\ j = k+1, k+2, \cdots, n$$

8.6.2 Python : Doolittle's LU Decomposition method

Program 8.6.1.

```
from numpy import dot
def LUdecomposition(a):
  n = len(a)
  for k in range(0,n-1) :
    for i in range(k+1,n):
```

```
        if a[i,k] != 0.0:
          lam = a[i,k]/a[k,k]
          a[i,k+1:n] = a[i,k+1:n] - lam*a[k,k+1:n]
          a[i,k] = lam
    return a
def LUsolve(a,b):
    n = len(a)
    for k in range(1,n):
      b[k] = b[k]-dot(a[k,0:k],b[0:k])
    b[n-1] = b[n-1]/a[n-1,n-1]
    for k in range(n-2,-1,-1):
      b[k] = (b[k] - dot(a[k,k+1:n],b[k+1:n]))/a[k,k]
    return b
```

This program mainly uses the Gauss elimination algorithm. Thus, the explanation for Lines 3-8 are not repeated here.

But remember the loop at Line 4 has inner loop at Line 5 and Line 7-9 are at same level of indentation which means they all are either executed or skipped depending on the truthness of the condition in Line 6. And Line 6-9 are executed for each instance of inner loop. Again, Line 5-9 are executed for each instance of the outer loop.

This time the gaussElimination() function which you have seen earlier is split into two functions 1. LUdecomposition() and 2. LUsolve(). And forward substituion is also added to LUsolve().

Line 2 `def LUdecomposition(a):`
: LUdecomposition(A) computes L and U such that $A = LU$ and combine both triangular matrices into a single matrix $[L/U]$, by over-writting their trivial entries. And returns this combined matrix.

Line 9 $a[i,k] = lam$
: Clearly, lam used for eliminating x_k from row i, $\lambda_{i,k} = a[i,k]/a[k,k] = L[i,k]$, $\forall k, \forall i$, $(i > k)$. Also $a[i,k]$ which is reduced zero by Gauss elimination process is not used anymore [9] in Gauss elimination process. Thus $L[i,k]$ can stored at $a[i,k]$ straight away. And $U[i,j]$, $j \le i$ are already the entries of the matrix obained from Gauss elimination. Thus for each iterations of k, the matrix a is updated ($k+1$th row and $k+1$th column) with respective entries of the combined matrix $[L\backslash U]$.

Line 10 `return` a
: Matrix a is already $[L\backslash U]$, and thus LUdecomposition(A) returns $[L\backslash U]$ such that $A = LU$.

Line 11 `def LUsolve(a,b):`
: LUsolve() function does both forward substitution and back substitution. Suppose $Ax = b$ is the system to be solved. Then the inputs of LUsolve() are $a = [L\backslash U]$ where $A = LU$.

[9] $A[i,k]$ is not used after elimination of x_k from row i - It turns out that the trivial zeroes which are ignored on the row operations in Gauss elimination not only save time, but also provide a variable to store our intermediate result $L_{i,k}$ in Doolittle method.

Line 12 $n = len(a)$
We have to compute this again as this function starts fresh and thus value of the variable n from LUdecomposition() is lost.

Line 13 `for` k `in range(1,`n`):`
This is the loop for forward substitution.

Line 14 $b[k] = b[k] - dot(a[k, 0 : k], b[0 : k])$
updating b_{k^*} with y_{k^*} such that $Ly = b$ where $k^* = k - 1$.

$$b_k \leftarrow b_k - \begin{bmatrix} L_{k,1} & L_{k,2} & \cdots & L_{k,k-1} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k - 1 \end{bmatrix}$$

Line 15 $b[n-1] = b[n-1]/a[n-1, n-1]$
Computing [10] y_n and storing it into b_n.

$$b_n \leftarrow \frac{b_n}{U_{n,n}}$$

Line 16 `for` k `in range(`$n-2$`,`-1`,`-1`):`
This is the loop for back substitution.

Line 17 $b[k] = (b[k] - dot(a[k, k+1 : n], b[k+1 : n]))/a[k, k]$
updating b_{k^*} with x_{k^*} such that $Ux = y$ where $k^* = k - 1$.

$$b_k \leftarrow \frac{b_k - \begin{bmatrix} x_{k+1} \\ x_{k+2} \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} U_{k,k+1} & U_{k,k+2} & \cdots & U_{k,n} \end{bmatrix}}{U_{k,k}}$$

Line 18 `return` b
LUsolve($[L\backslash U]$,b) returns b where $b[i+1] = x_i$.

Programmer's Tip : There are few things to remember when spliting a function into two functions.

1. These functions are completely independent of one another.
2. Variables defined inside a function are not available outside.
3. The best way to give/take data to/from a function is through arguements/return-value

Beginner's Tip : In any programming language, we reuse variable. Thus, same variable may represent different values at different points of time. In Doolittle LU Decomposition, the variable 'a' initially represent matrix A, this variable is passed into LUdecomposition() function. In that function, A is changed to $[L\backslash U]$ in a step-by-step fashion. The value of 'a' in updated in step 7, 8 and 9. This is bit hard to imagine this transition of 'a' from A to $[L\backslash U]$ for a beginner at programming. Simiarly, in LUsolve() function, the variable 'b' changes from matrix b to matrix y, and then to matrix x.

[10] Mathematically, you can define dot product of empty matrices as zero, but nympy dot function can't handle such a situation. Therefore, we have to do this step separately.

8.7 Numerical Integraion

Numerical integration/Quadrature is the numerical approximation of $\int_a^b f(x)dx$ by $\sum_{i=0}^n A_i f(x_i)$ where x_i are nodal abscissas, and A_i are weights. There are two methods to determine these nodal abscissas and suitable weights so that the sum is sufficiently accurate to the value of the integral.

1. Newton-Cotes forumulas
2. Gauss quadrature

Newton-Cotes formulas are useful when $f(x)$ can be evaluated without much computation. And using those values $f(x)$ can be interpolated to a piecewise-polynomial function. Then using equally spaces nodal abscissas and suitable weights $\int_a^b f(x)dx$ can be numerically approximated.

Gauss quadrature rules require lesser evalutions of f. And therefore are quite useful when evaluation of $f(x)$ has much computational complexity. Also, this method can manage integrable singularities where as Newton-Cote formulas can't numerically integrate function with singularities.

8.7.1 Newton-Cotes formulas

We divide the interval of integral (a, b) into n subintervals of equal length, ie, $h = (b - a)/n$. Let $x_0 = a, x_1, x_2, \cdots, x_{n-1}, x_n = b$ be the end points of these subintervals. Then we can find an n degree polymonial interpolant satisfying f at those points, using Lagrange's method.

$$\text{Polynomial, } P(x) = \sum_{i=0}^{n} f(x_i) l_i(x) \text{ where } l_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_i}{x_i - x_j}$$

Thus the integral $I = \int_a^b f(x)dx$ can be numerically evaluated as follows:

$$\begin{aligned} I = \int_a^b P_n(x)dx &= \sum_{i=0}^{n} \left(f(x_i) \int_a^b l_i(x)dx \right) \\ &= \sum_{i=0}^{n} A_i f(x_i), \text{ where } A_i = \int_a^b l_i(x)dx \end{aligned}$$

The simplest cases of Newton-Cotes formulas are when $n = 1, 2, and 3$

Trapezoidal rule $n = 1 \implies A_0 = \frac{h}{2},\ A_1 = \frac{h}{2}$ and

$$\int_a^b f(x)dx = A_0 f(x_0) + A_1 f(x_1) = \frac{h}{2}(f(a) + f(b))$$

Simpson's $1/3$ **rule** $n = 2 \implies A_0 = \frac{h}{3},\ A_1 = \frac{4h}{3},\ A_2 = \frac{h}{3}$ and

$$\int_a^b f(x)dx = \sum_{i=0}^{2} A_i f(x_i) = \frac{h}{3} \left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right)$$

Simpson's 3/8 **rule** $n = 3 \implies A_0 = \frac{3h}{8},\ A_1 = \frac{9h}{8},\ A_2 = \frac{9h}{8},\ A_3 = \frac{3h}{8}$ and

$$\int_a^b f(x)dx = \sum_{i=0}^{3} A_i f(x_i) = \frac{3h}{8}(f(a) + 3f(a+h) + 3f(a+2h) + f(b))$$

Trapezoidal Rule : $n = 1$

Consider interval (a, b). Since $n = 1$, we have $x_0 = a$ and $x_1 = b$.

$$l_0(x) = \frac{x - x_1}{x_0 - x_1}$$

$$l_1(x) = \frac{x - x_0}{x_1 - x_0}$$

$$\begin{aligned}
A_0 &= \int_a^b l_0(x)dx = \int_a^b \frac{x - x_1}{x_0 - x_1}dx = \frac{-1}{h}\int_a^b (x - b)dx \\
&= \frac{-1}{h}\left(\frac{(x-b)^2}{2}\right)_a^b = \frac{-1}{h}\left(\frac{0 - (a-b)^2}{2}\right) = \frac{h}{2} \\
A_1 &= \int_a^b l_1(x)dx = \int_a^b \frac{x - x_0}{x_1 - x_0}dx = \frac{1}{h}\int_a^b (x - a)dx \\
&= \frac{1}{h}\left(\frac{(x-a)^2}{2}\right)_a^b = \frac{1}{h}\left(\frac{(b-a)^2 - 0}{2}\right) = \frac{h}{2}
\end{aligned}$$

Therefore,

$$\int_a^b f(x)dx = A_0 f(x_0) + A_1 f(x_1) = \frac{h}{2}(f(a) + f(b))$$

Simpon's 1/3 **Rule :** $n = 2$

Consider interval (a, b) divided into two subintervals of equal length $h = \frac{a+b}{2}$. We have $x_0 = a,\ x_1 = \frac{a+b}{2},\ x_2 = b$.

$$l_0(x) = \frac{x - x_1}{x_0 - x_1}\frac{x - x_2}{x_0 - x_2}$$

$$l_1(x) = \frac{x - x_0}{x_1 - x_0}\frac{x - x_2}{x_1 - x_2}$$

$$l_2(x) = \frac{x - x_0}{x_2 - x_0}\frac{x - x_1}{x_2 - x_1}$$

$$\begin{aligned}
A_0 &= \int_a^b l_0(x)dx \\
&= \int_a^b \frac{x - x_1}{x_0 - x_1}\frac{x - x_2}{x_0 - x_2}dx \\
&= \int_a^b \left(\frac{x - \frac{a+b}{2}}{a - \frac{a+b}{2}}\right)\left(\frac{x - b}{a - b}\right)dx
\end{aligned}$$

Changing variable of integration $y = x - \frac{a+b}{2}$

$$\begin{aligned} y = x - \frac{a+b}{2} &\implies dy = dx \\ x = a &\implies y = -h \\ x = b &\implies y = h \end{aligned}$$

Continuing with the value of A_0,

$$\begin{aligned} A_0 &= \int_{-h}^{h} \left(\frac{y}{-h}\right)\left(\frac{y-h}{-2h}\right) dy \\ &= \frac{1}{2h^2}\int_{-h}^{h} y^2 - \frac{1}{2h}\int_{-h}^{h} ydy \\ &= \frac{1}{2h^2}\left(\frac{h^3}{3} - \frac{(-h)^3}{3}\right) - \frac{1}{2h}\left(\frac{h^2}{2} - \frac{(-h)^2}{2}\right) \\ &= \frac{h}{3} \end{aligned}$$

$$\begin{aligned} A_1 &= \int_a^b l_1(x)dx \\ &= \int_a^b \frac{x-x_0}{x_1-x_0}\frac{x-x_2}{x_1-x_2}dx \\ &= \int_a^b \left(\frac{x-a}{h}\right)\left(\frac{x-b}{-h}\right) dx \end{aligned}$$

Applying change of variable, $y = x - \frac{a+b}{2}$

$$\begin{aligned} &= \int_{-h}^{h} \left(\frac{y+h}{h}\right)\left(\frac{y-h}{-h}\right) dy \\ &= \frac{-1}{h^2}\int_{-h}^{h} y^2 dy + \int_{-h}^{h} 1dy \\ &= \frac{-1}{h^2}\left(\frac{h^3}{3} - \frac{(-h)^3}{3}\right) + (h - (-h)) \\ &= \frac{-2h^3}{3h^2} + 2h \\ &= \frac{4h}{3} \end{aligned}$$

$$\begin{aligned} A_2 &= \int_a^b l_2(x)dx \\ &= \int_a^b \left(\frac{x-x_0}{x_2-x_0}\right)\left(\frac{x-x_1}{x_2-x_1}\right) dx \\ &= \int_a^b \left(\frac{x-a}{2h}\right)\left(\frac{x-\frac{a+b}{2}}{h}\right) dx \end{aligned}$$

Applying change of variable, $y = x - \frac{a+b}{2}$

$$
\begin{aligned}
&= \int_{-h}^{h} \left(\frac{y+h}{2h}\right)\left(\frac{y}{h}\right) dy \\
&= \frac{1}{2h^2}\int_{-h}^{h} y^2 dy + \frac{1}{2h}\int_{-h}^{h} y dy \\
&= \frac{1}{2h^2}\left(\frac{h^3}{3} - \frac{(-h)^3}{3}\right) + \frac{1}{2h}\left(\frac{h^2}{2} - \frac{(-h)^2}{2}\right) \\
&= \frac{2h^3}{6h^2} \\
&= \frac{h}{3}
\end{aligned}
$$

Therefore,

$$\int_a^b f(x)dx = \sum_{i=0}^{2} A_i f(x_i) = \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$$

8.7.2 Composite Trapezoidal Rule

Suppose an interval (a, b) is divided into n subintervals. In Composite Trapezoidal Rule, Trapezoidal Rule is applied to each subinterval. Thus we have,

$$
\begin{aligned}
I &= I_0 + I_1 + \cdots + I_{n-1} \text{ where } I_k \text{ is the integral over } (x_k, x_{k+1}) \\
&= \frac{h(f(x_0) + f(x_1))}{2} + \frac{h(f(x_1) + f(x_2))}{2} + \cdots + \frac{h(f(x_{n-1} + f(x_n))}{2} \\
&= \frac{h}{2}(f(x_0) + 2f(x_1) + \cdots + 2f(x_{n-1}) + f(x_n))
\end{aligned}
$$

8.7.3 Recursive Trapezoidal Rule

Suppose an interval (a, b) is divided into 2^{k-1} subintervals. In Recursive Trapezoidal Rule, we apply Trapezoidal Rule on each subinterval. And there is a recursive forumula since the number of intervals doubles as value of k increases by one. Let $H = b - a$

$$
\begin{aligned}
k = 1 &\implies 2^0 = 1 \text{ and} \\
I_1 =& \frac{H}{2}(f(a) + f(b))
\end{aligned}
$$

$$
\begin{aligned}
k = 2 &\implies 2^1 = 2 \text{ and} \\
I_2 =& \frac{H}{4}(f(a) + 2f\left(a + \frac{H}{2}\right) + f(b) \\
=& \frac{H}{4}(f(a) + f(b)) + \frac{H}{2}f\left(a + \frac{H}{2}\right) \\
=& \frac{I_1}{2} + \frac{H}{2}f\left(a + \frac{H}{2}\right)
\end{aligned}
$$

$k = 3 \implies 2^2 = 4$ and

$$
\begin{aligned}
I_3 =& \frac{H}{8}(f(a) + 2f\left(a + \frac{H}{4}\right) + 2f\left(a + \frac{2H}{4}\right) + 2f\left(a + \frac{3H}{4}\right) + f(b)) \\
=& \frac{H}{8}(f(a) + 2f\left(a + \frac{2H}{4}\right) + 2f\left(a + \frac{4H}{4}\right) + 2f\left(a + \frac{6H}{4}\right) + f(b)) \\
&+ \frac{H}{4}(f\left(a + \frac{H}{4}\right) + f\left(a + \frac{3H}{4}\right)) \\
=& \frac{I_2}{2} + \frac{H}{4}(f\left(a + \frac{H}{4}\right) + f\left(a + \frac{3H}{4}\right))
\end{aligned}
$$

Consider interval $(0, 64)$. We have $b - a = H = 64$.
$I_1 = 32(f(0) + f(64))$
$I_2 = 16(f(0) + 2f(32) + f(64))$
$I_3 = 8(f(0) + 2f(16) + 2f(32) + 2f(48) + f(64))$
$I_4 = 4(f(0) + 2f(8) + 2f(16) + \cdots + 2f(56) + f(64)$
$I_5 = 2(f(0) + 2f(4) + 2f(8) + \cdots + 2f(60) + f(64)$
$I_6 = f(0) + 2f(2) + 2f(4) + \cdots + 2f(62) + f(64)$
$\vdots$
The values corresponding to the intervals in I_{k-1} appear in I_k as alternate terms. Other terms, corresponds to the odd multiples of $\frac{H}{2^k}$. We separate them into two sums and represent the first sum as $\frac{I_{k-1}}{2}$.

$$
\begin{aligned}
\text{Clearly, } I_k =& \frac{H}{2^k}\sum_{i=0}^{2^{k-2}} f\left(a + \frac{2iH}{2^k}\right) + \frac{2H}{2^k}\sum_{i=1}^{2^{k-2}} f\left(a + \frac{(2i-1)H}{2^{k-1}}\right) \\
=& \frac{I_{k-1}}{2} + \frac{H}{2^{k-1}}\sum_{i=1}^{2^{k-2}} f\left(a + \frac{(2i-1)H}{2^{k-1}}\right)
\end{aligned}
$$

8.7.4 Python : Recursive Trapezoidal Rule

Program 8.7.1.

```
def recursiveTrapezoidalRule(f,a,b,Iold,k):
    if k == 1 :
        Inew = (f(a)+f(b))*(b-a)/2.0
    else :
        n = 2**(k-2)
        h = (b-a) * 1.0 / n
        x = a + h/2.0
        sum = 0.0
        for i in range(n):
            sum = sum + f(x)
            x = x + h
        Inew = (Iold + h * sum ) / 2.0
    return Inew
```

Line 1 `def recursiveTrapezoidalRule(f,a,b,Iold,k):`
This function has five input arguments. (a) f is a real function (b) a,b are start and end of interval in which f is going to be integrated (c) $Iold$ is

the value of the integral for 2^{k-1} subintevals using recursive trapezoidal method (d) k is a variable such that 2^k is the number of subintervals considered for Integration.

Line 2 `if` $k == 1$ `:`
If $k = 1$, we proceed to Line 3, otherwise we go to Line 4.

Line 3 $Inew = (f(a) + f(b)) * (b - a)/2.0$
For $k = 1$, we use trapezoidal rule $I_1 = \frac{b-a}{2}(f(a) + f(b))$. When writing this in python, we are use 2.0 so that the python won't ignore the decimal part of this fraction. In python, $5/2 = 2$. And $5/2.0 = 2.5$

Line 4 `else :`
If Line 2 is false, (ie $k \neq 1$) python executes Line 5-8. These line implements the recursive formula for I_k.

Line 5 $n = 2 * *(k - 2)$
Equivalent to $n \leftarrow 2^{k-2}$.

Line 6 $h = (b - a) * 1.0/n$
This the length of a subinterval when we divide (a, b) into 2^k subintervals/panels. Equivalent to $h \leftarrow \frac{b-a}{2^{k-2}}$

Line 7 $x = a + h/2.0$
This the parameter of f in the first term in the sum $\sum_{i=1}^{2^{k-2}} f\left(a + \frac{(2i-1)H}{2^{k-1}}\right)$ in the recursive formula for I_k. Equivalent to $x \leftarrow a + \frac{h}{2}$.

Line 8 $sum = 0.0$
We are going to use this variable to find that sum. To start with, we will make it 0 and will add each term to it one-by-one. Equivalent to $sum \leftarrow 0$.

Line 9 `for` i `in range(`n`):`
This the variable i in the recursive forumula for I_k. For each value of $i = 1, 2, \cdots, 2^{k-2}$, Lines 10 and 11 are executed. That is, for each value of i, the corresponding term in the sum is computed and added to the variable sum.

Line 10 $sum = sum + f(x)$
Value of f at x is computed and added to the partial sum. Equivalent to $sum \leftarrow sum + f(x)$. However, the value of x is changed for each i in Line 11. Thus, for next value of i, x and sum have the new values to use.

Line 11 $x = x + h$
Equivalent to $x \leftarrow x + h$. For $i = 1$, $x = a + \frac{h}{2}$ before Line 11. At Line 11, $x \leftarrow a + \frac{3h}{2}$. And this is the value of x for $i = 2$ before Line 11 next time. Thus, x iterates through $a + \frac{h}{2}, a + \frac{3h}{2}, \cdots, a + \frac{(2n-1)h}{2}$. This x is updated and used for next execution of Line 10 and 11.

Line 12 $Inew = (Iold + h * sum)/2.0$
We reach here only after executing Line 10-11 for all values of i. That is, sum in the recursive forumula is already computed. This line, implements the recursive formula and stores that value into the variable $Inew$. Equivalent to $Inew \leftarrow \frac{Iold + h \times sum}{2}$.

Line 13 `return Inew`

It returns the value of I_k, the intergral of f over (a, b) using recursive trapezoidal rule for 2^k subintervals.

Subject 9

ME010204 Complex Analysis

9.1 Module 2

9.1.1 Arcs & Closed Curves

An arc γ in a complex plane is defined as the set of points given by $\gamma = \{z : z = z(t), a \leq t \leq b\}$ and $z(t)$ is a continuous function of the real variable t. Thus, every arc in the complex plane is the continuous image of closed interval and $z \in \gamma$ means $z = z(t) = x(t) + iy(t)$. That is, points on γ are images of a complex function of a real variable.

This representation of an arc $z = z(t)$ is called a parameteric representation and t is called the parameter and $[a, b]$ is called the parametric interval. The point $z = z(a)$ is called origin or initial point of γ and $z = z(b)$ is called the terminus or terminal point of γ. If $z(a) = z(b)$, then γ is called a closed curve, otherwise it is open.

If in γ, $z(t_1) = z(t_2) = z$ for $t_1 \neq t_2$, then z is called a multiple point on γ. Geometrically, a multiple point z in γ is a point where γ crosses itself.

An arc having no multiple points is called a simple arc or Jordan arc.

9.1.2 Differentiable Arc

differentiable $z'(t)$ exist and is continuous at all points.

regular differentiable and $z'(t) \neq 0$ at points on γ.

piecewise differentiable differentiable except for finitely many points

piecewise regular regular except for finitely many points.

opposite arc set of points $z = z(-t),\ -b \leq t \leq -a$.

9.1.3 Complex Integration

1. If $f(z)$ has an antiderivative $F(z)$, then $\int_\gamma f(z)\ dz$ depends only on the end points and is independent of the path γ.

2. If $f(z)$ has an antiderivative, then $\int_\gamma f(z)\ dz = 0$ for all closed curves γ.

3. If $f(z)$ has no antiderivatives, $\int_\gamma f(z)\ dz$ depends on the path γ.

For different choices of γ, the integral may have different values even though the end points are the same.

9.1.4 Exercise

Evaluate $\int_c f(z)\ dz$ where $f(z) = y - x - i3x^2$ and contour is c_1: the line segment 0 and $i + 1$ c_2: the polygon joining $(0, 0),\ (0, 1),\ (1, 1)$ In the above example, $\int_{c_1} f(z)\ dz \neq \int_{c_2} f(z)\ dz$ even though c_1 and c_2 have the same end points.

9.1.5 Evaluating Line Integral : Method 1

Integrals of the form $\int_a^b f(t)\ dt$ where $f(t)$ is a complex valued function of a real variable t. Then, we can write $f(t) = u(t) + iv(t)$.

$$\int_a^b f(t)\ dt = \int_a^b u(t) + iv(t)\ dt = \int_a^b u(t)\ dt + i \int_a^b v(t)\ dt$$

That is, $\Re \int f(t)\ dt = \int \Re f(t)\ dt$ and $\Im \int f(t)\ dt = \int \Im f(t)\ dt$.
For example, $f(t) = e^{it},\ 0 \leq t \leq \frac{\pi}{2}$. Then

$$\int_0^{\frac{\pi}{2}} f(t)\ dt = \int_0^{\frac{\pi}{2}} \cos t\ dt + i \int_0^{\frac{\pi}{2}} \sin t\ dt = 1 + i$$

9.1.6 Evaluating Line Integral : Method 2

Let γ be a piecewise differentiable arc in the complex plane defined by the equation $z = z(t),\ a \leq t \leq b$ and $f(z)$ be defined and continuous in γ. Then the line integral $\int_\gamma f(z)\ dz$ is defined by

$$\int_\gamma f(z)\ dz = \int_a^b f(z(t))\ z'(t)\ dt$$

For example, let $f(z) = u + iv$. Then $f(z)\ dz = (u + iv)(dx + idy)$

$$\int_\gamma f(z)\ dz = \int_\gamma udx - vdy + i \int_\gamma vdx + udy$$

That is, real and imaginary part of $\int_\gamma f(z)\ dz$ can be written in the form $\int_\gamma pdx + qdy$ where p, q are real valued functions of x and y. That is, real valued functions of two real variables x and y. Therefore, the study of line integral $\int_\gamma f(z)\ dz$ can be restricted to the study of line integrals of the form $\int_\gamma pdx + qdy$ and line integrals can be defined as $\int_\gamma pdx + qdy$ where γ is a piecewise differentiable arc.

Properties

1. Scalar multiplication,

$$\int_\gamma cf(z)\ dz = c\int_\gamma f(z)\ dz$$

2. Modulus Inequality,

$$\left|\int_a^b f(t)\ dt\right| \le \int_a^b |f(t)\ dt|$$

-more–

3. Change of variable,

$$\int_\gamma f(z)\ dz = \int_a^b f(z(t))\ z'(t)\ dt$$

4. Inverse arc,

$$\int_{-\gamma} f(z)\ dz = -\int_\gamma f(z)\ dz$$

5. Integration by parts,

$$\int_\gamma f(z)\ dz = \int_{\gamma_1} f(z)\ dz + \int_{\gamma_2} f(z)\ dz + \cdots + \int_{\gamma_n} f(z)\ dz$$

9.1.7 Line Integral : Type 3

Line integrals with respect to $\bar{z}$ are denoted by

$$\int_\gamma f(z)\ d\bar{z} = \overline{\int_\gamma \bar{f}\ dz}$$

Proof. We have, $x = \frac{z+\bar{z}}{2}$ and $y = \frac{z-\bar{z}}{2i}$.

$$dx = \frac{dz + d\bar{z}}{2} \qquad dy = \frac{dz - d\bar{z}}{2i}$$

$$\int_\gamma f(z)\ dx = \int_\gamma f(z)\frac{dz + d\bar{z}}{2} \qquad \int_\gamma f(z)\ dy = \int_\gamma f(z)\frac{dz - d\bar{z}}{2i}$$

$$\int_\gamma f(z)\ dx - idy = \int_\gamma f(z)\ d\bar{z}$$

–more– □

9.1.8 Line Integral : Type 4

Line integrals with respect to arc length, s is denoted by

$$\int_\gamma f(z)\ ds = \int_\gamma f(z)\ |dz| \text{ where } ds = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\ dx$$

When $f = 1$, it gives the length of the arc.

9.1.9 Rectifiable Arc

Length of an arc γ is defined as $L(\gamma) = \int |dz|$. If $L(\gamma) < \infty$, then γ is a rectifiable arc and the process of determining the length of an arc is called rectification.

Subject 10

ME010205 Measure & Integration

10.2 Lebesgue Measure

10.2.1 Introduction

set function A function which maps sets into (extended) real numbers.

σ-algebra A family $\mathcal{A}$ of subsets of a nonempty set X such that

1. $\mathcal{A}$ contains the empty set,
2. $\mathcal{A}$ contains complement of each of its members and
3. $\mathcal{A}$ is closed under countable unions.

From these 3 axioms, we can deduce the following,

4. $\mathcal{A}$ is closed under countable intersections (by de Morgan's laws).

$$\left(\bigcup_{k=1}^{\infty} E_k^c\right)^c = \bigcap_{k=1}^{\infty} E_k \in \mathcal{A}$$

5. $E, F \in \mathcal{A} \implies E - F \in \mathcal{A}$ since $E - F = E \cap F^c$

Definitions 10.2.1 (Length of an interval)**.** Length is a real valued set function. Let I be a bounded interval say $[a, b)$. Then its length $l(I) = b - a$ is the difference between endpoints. If an interval I is unbounded say (a, ∞), then its length, $l(I) = \infty$.

Exercise

Techniques in Measure Theory Let $\mathcal{A}$ be a σ-algebra. Let Lebesgue Measure $m : \mathcal{A} \to [0, \infty]$ be countably additive over disjoint collection of sets in $\mathcal{A}$.

- Lebesgue Measure m has monotonicity.
  $A \subseteq B \implies B = A \cup (B - A)$ is a disjoint union
  $\implies m(B) = m(A) + m(B - A) \geq m(A)$

- If exists $E \in \mathcal{A}$ such that $m(E) < \infty$, then $m(\phi) = 0$
  Suppose $m(\phi) = c$ and $m(E) = k$ where $k < \infty$. If $c \neq 0$, then $m(E \cup \phi) = m(E) + m(\phi) = c + k > k = m(E)$ is a contradiction.

- $m\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty} m(E_k)$
  Define $\{F_k : k \in \mathbb{N}\}$ by $F_k = E_k - \cup_{j=1}^{k-1} E_j$
  Then $F_1 = E_1,\ F_2 = E_2 - E_1,\ F_3 = E_3 - (E_1 \cup E_2),\ \ldots$
  Also $F_k \in \mathcal{A}$ and $F_k \subseteq E_k,\ \forall k \in \mathbb{N}$. Thus $m(F_k) \leq m(E_k),\ \forall k$ However, $\cup_{k=1}^{\infty} E_k = \cup_{k=1}^{\infty} F_k$
  $\implies m\left(\cup_{k=1}^{\infty} E_k\right) = m\left(\cup_{k=1}^{\infty} F_k\right) = \sum_{k=1}^{\infty} m(F_k) \leq \sum_{k=1}^{\infty} m(E_k)$

Counting Measure The counting measure $c : \mathcal{A} \to [0, \infty]$ is a set function which maps sets to their cardinality. For example, if $E = \{2, 3, 4\}$, then $c(E) = 3$.

- The counting measure is **translation invariant** since translation never increases the cardinality of the set.
  For example, $5 + E = \{7, 8, 9\}$. And $m(5 + E) = 3 = m(E)$.

- The counting measure is **countably additive** over disjoint collections since the cardinality of disjoint union of two sets is the sum of their cardinalities.

- However, counting measure of (non-degenerate) intervals are ∞ which is **not the same as their length** for bounded intervals.

10.2.2 Lebesgue Outer Measure

G_δ A set which is countable intersection of open subsets.

F_σ A set which is countable union of closed subsets.

Caratheodory Construction of Lebesgue Measure

1. Construct Lebesgue Outer Measure m^* (with Axiom 3 relaxed)
   ie, Obtain the underlying relation of the set function

2. Restrict m^* to the σ-algebra of our interest
   ie, Choose a domain so that set function is well defined.

Definitions 10.2.2 (Lebesgue Outer Measure)**.** Let $A \subset \mathbb{R}$. Let $\mathcal{C} = \{I_k : k \in \mathbb{N}\}$ be an open cover of A such that I_k are non-empty, bounded, open intervals. Consider the sum of length of intervals for such covers of A. (Lebesgue) Outer Measure $m^*(A)$ is the infimum of all such sums.

$$m^*(A) = \inf\left\{\sum_{k=1}^{\infty} l(I_k) : A \subset \bigcup_{k=1}^{\infty} I_k\right\} \tag{10.1}$$

Properties of Lebesgue Outer Measure

1. Outer Measure of the empty set is zero
   Let $\varepsilon > 0$. Then $\mathcal{C}_\varepsilon = \{(0, \frac{\varepsilon}{2^n}) : n \in \mathbb{N}\}$ is an open cover of ϕ containing nonempty, bounded, open intervals. Clearly, sum of length of intervals in $C_\varepsilon = \varepsilon$. Suppose $m^*(\phi) = \delta$ and $\delta > 0$. There exists ε such that $0 < \varepsilon < \delta$. The sum of intervals of $\mathcal{C}_\varepsilon$ is less than δ, which is a contradiction by the definition of Outer Measure.

2. Outer Measure is monotone
   Suppose $A \subset B$. Then every cover of B is also an cover of A. Let $\mathcal{U}$ be the set of all open covers of A with nonempty, bounded intervals and $\mathcal{V}$ be the set of all such open covers of A. Clearly, $\mathcal{V} \subset \mathcal{U}$. We know that, if $A \subset B$, then $\inf B \leq \inf A$. Therefore,
   $$A \subset B \implies m^*(A) \leq m^*(B) \tag{10.2}$$

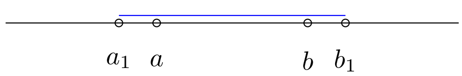
3. Outer Measure of Countable Sets is zero
   Let C be a countable set. That is, $C = \{c_k\}_{k=1}^\infty$.
   Then $\{(c_k - \frac{\varepsilon}{2^k}, c_k + \frac{\varepsilon}{2^k})\}_{k=1}^\infty$ is cover of C with sum of length of intervals ε. Thus, for any $\varepsilon > 0$, we have $m^*(C) \leq \varepsilon$. Thus, $m^*(C) = 0$.

4. Outer Measure of an Interval is its length

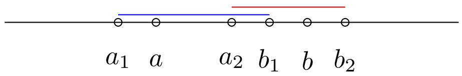
   Proof. **Case 1 : Closed, Bounded Interval** Let $[a, b]$ be a closed, bounded interval. Then for any $\varepsilon > 0$, $(a - \varepsilon, b + \varepsilon)$ is a cover of $[a, b]$. Thus, by the definition of Lebesgue outer measure $m^*([a, b]) \leq b - a + 2\varepsilon$ since $[a, b] \subset (a - \varepsilon, b + \varepsilon)$ and m^* is monotonic. Therefore,
   $$m^*([a, b]) \leq b - a \tag{10.3}$$

   Since $[a, b]$ is closed and bounded, $[a, b]$ is compact. And by Heine-Borel theorem, every open cover of $[a, b]$ has a finite subcover. Thus, it is sufficient to prove the theorem for finite covers of $[a, b]$.

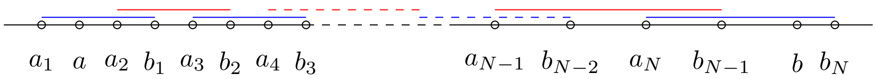
   Let $\mathcal{C}$ be a finite cover of $[a, b]$ with n open intervals. Let (a_1, b_1) be an open interval containing a in $\mathcal{C}$. Then $a_1 < a < b_1$. If $b_1 > b$ then $l(a_1, b_1) > l(a, b)$. And $\displaystyle\sum_{i=1}^{k} l(I_k) \geq l(a_1, b_1) \geq b - a$.

   Suppose $b_1 < b$. Clearly, $a < b_1$. And the cover $\mathcal{C}$ must have an open interval containing b_1. Otherwise $\mathcal{C}$ is not a cover of $[a, b]$. That is, there exists (a_2, b_2) containing $b_1 \in (a, b)$ such that $a_2 < b_1 < b_2$. If $b_2 > b$, then $\displaystyle\sum_{i=1}^{k} l(I_k) \geq l(a_1, b_1) + l(a_2, b_2) \geq l(a_1, b_2) \geq b - a$.



Suppose $b_2 < b$. Continuing like this we get, N open intervals in $\mathcal{C}$, $\{(a_k, b_k) : k = 1, 2, \cdots, N\}$ such that $a_1 < a < b_1$ and $a_N < b < b_N$ and $a_k < b_{k-1} < b_k$ for all k. The process should terminate in finite steps as $\mathcal{C}$ is a finite cover of $[a, b]$. Then $\sum_{k=1}^{N} l(I_k) \geq \sum_{k=1}^{N} l(a_k, b_k) \geq l(a_1, b_N) \geq b - a$.



Clearly, every open cover of $[a, b]$ contains a finite subcover $\mathcal{C}$, which contains a finite subcover of the form $\{(a_k, b_k) : k = 1, 2, \cdots, N\}$ such that $\sum_{k=1}^{N} l(I_k) \geq b - a$. Thus, for any open cover $\sum_{k=1}^{\infty} l(I_k) \geq b - a$. And thus,

$$m^*([a, b]) \geq b - a \tag{10.4}$$

Case 2 : Bounded Interval Let I be a bounded interval. Then there exists bounded closed intervals J_1 and J_2 such that $J_1 \subsetneq I \subsetneq J_2$ such that $l(I) - \varepsilon < l(J_1)$ and $l(J_2) < l(I) + \varepsilon$. Suppose $I = (a, b]$, then $J_1 = [a + \frac{\varepsilon}{2}, b - \frac{\varepsilon}{2}]$ and $J_2 = [a - \frac{\varepsilon}{2}, b + \frac{\varepsilon}{2}]$.

By monotonicity of Lebesgue outer measure, we have $m^*(J_1) \leq m^*(I) \leq m^*(J_2)$. However $m^*(J_1) = l(I) - \varepsilon$ and $m^*(J_2) = l(I) + \varepsilon$. Thus, $l(I) - \varepsilon \leq m^*(I) \leq l(I) + \varepsilon$. Therefore, $m^*(I) = l(I)$.

Case 3 : Unbounded Interval Let I be an unbounded interval. Then for any natural number n, there exists a closed bounded interval J such that $J \subset I$ and $l(J) = n$. And $n = m^*(J) \leq m^*(I)$, $\forall n \in \mathbb{N}$. Therefore, $m^*(I) = \infty = l(I)$. □

5. Outer Measure is translation invariant

   Proof. Let A be any set and $y \in \mathbb{R}$. Let $\{I_k : k = 1, 2, \dots\}$ be a cover of A. Then $\{I_k + y : k = 1, 2, \dots\}$ is a cover of $A + y$. And $l(I_k) = l(I_k + y)$ for every natural number k and real number y. Thus, $\sum_{k=1}^{\infty} l(I_k) = \sum_{k=1}^{\infty} l(I_k + y)$. Clearly, for each cover $\{I_k\}_{k=1}^{\infty}$ of A, there exists a cover $\{I_k + y\}_{k=1}^{\infty}$ of $A + y$ containing intevals of same length. Therefore, $m^*(A) = m^*(A + y)$. □

6. Outer Measure is countably subadditive

   Proof. Let $\{E_k\}_{k=1}^{\infty}$ be a countable collection of sets. It is enough to prove that

$$m^*\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty} m^*(E_k) \tag{10.5}$$

For each natural number k, we have a cover of E_k, say $\{I_{k,i}\}_{i=1}^{\infty}$ such that $\sum_{i=1}^{\infty} l(I_{k,i}) < m^*(E_k) + \frac{\varepsilon}{2^k}$. Suppose that, for some $\varepsilon > 0$, E_k doesn't have such a cover, then $m^*(E_k) + \frac{\varepsilon}{2^k}$ is an upper bound contradicting the assumption that $m^*(E_k)$ is the least upper bound.

Clearly, '

$$m^*\left(\bigcup_{i,k=0}^{\infty} I_{k,i}\right) \leq \sum_{k=1}^{\infty}\sum_{i=1}^{\infty} l(I_{k,i})$$
$$= \sum_{k=1}^{\infty}\left(m^*(E_k) + \frac{\varepsilon}{2^k}\right)$$
$$= \sum_{k=1}^{\infty} m^*(E_k) + \varepsilon$$

□

Note : Finite subadditivity is a weaker notion than countable subadditivity. Since every finite collection is a countable collection.

Exercise

5. Closed Interval $[0, 1]$ is uncountable.
   Suppose $[0, 1]$ is countable, then Lebesgue outer measure of any countable set is zero, $m^*([0, 1]) = 0$. But, $[0, 1]$ is an interval and Lebesgue outer measure of an interval is its length, $m^*([0, 1]) = l([0, 1]) = 1$ which is a contradiction.

6. $m^*([0, 1] - \mathbb{Q}) = 1$

   $$[0, 1] = ([0, 1] \cap \mathbb{Q}) \ \cup \ ([0, 1] \cap \mathbb{Q}^c)$$

   Clearly, $m^*([0, 1]) = 1$. And $[0, 1] \cap \mathbb{Q}$ is a countable set since $\mathbb{Q}$ is countable. And thus has Lebesgue outer measure zero. Thus by countable subadditivity, we have

   $$1 = m^*([0, 1]) \leq m^*([0, 1] \cap \mathbb{Q}^c) + 0$$

   Thus, $m^*([0, 1] \cap \mathbb{Q}^c) \geq 1$. And $[0, 1] \cap \mathbb{Q}^c \subset [0, 1]$. By monotonicity, $m^*([0, 1] \cap \mathbb{Q}^c) \leq m^*([0, 1]) = 1$. Therefore, $m^*([0, 1] \cap \mathbb{Q}^c) = 1$.

7. Construction of a G_δ set containing E

8. hint : if sum of interval is less than 1. Then it is not a cover of $[0, 1]$.

9. hint : $A \cup B = A \cup (B - A) = A \cup (B \cap A^c)$

10. hint : A and B are separated by distance α, thus are disjont.

10.2.3 σ-algebra of Lebesgue Measurable Sets

Lebesgue Outer Measure is defined for any subset of real numbers and Lebesgue outer measure of an interval is its length. However, it isn't countable additive.

There exists disjoint sets A, B such that $m^*(A \cup B) < m^*(A) + m^*(B)$.

Since countable additivity is a favourable property over countable subadditivity. We restrict the family of subsets of real numbers to those subsets that allow countable additivity.

Lebesgue Measurable Set

Definitions 10.2.3 (Measurable Set). Let E be a subset of $\mathbb{R}$. Then E is Lebesgue measurable if

$$m^*(A) = m^*(A \cap E) + m^*(A \cap E^c) \tag{10.6}$$

for any subset A of $\mathbb{R}$.

In other words, E is Lebesgue measurable if E doesn't affect countable additivity of Lebesgue Outer Measure.

We will consider only those subset of real numbers, which won't affect countable additivity. These subsets are **Lebesgue Measurable**. And we could show that the collection of all Lebesgue Measurable sets forms a σ-algebra. Clearly, intervals allow countable additivity, thus the Borel Algebra is contained in this σ-algebra of Lebesgue measurable sets.

Simplified Condition for Lebesgue Measurability

We know that Lebesgue Outer Measure has countable subadditivity.

$$m^*(A) \leq m^*(A \cap E) + m^*(A \cap E^c)$$

Thus, for condition (10.6), it is sufficient to check the following condition,

$$m^*(A) \geq m^*(A \cap E) + m^*(A \cap E^c) \tag{10.7}$$

Properties of Lebesgue Measure

1. Any set of Lebesgue outer measure zero is Lebesgue measurable.

   Proof. Let E be a subset of real numbers with Lebesgue outer measure zero. Let A be any subset of real numbers. Then $A = (A \cap E) \cup (A \cap E^c)$. By countable additivity, $m^*(A) \leq m^*(A \cap E) + m^*(A \cap E^c)$. Since $A \cap E \subset E$, we have by monotonicity $m^*(A \cap E) \leq m^*(E) = 0$.

   Again, $A \cap E^c \subset A$ and by monotonicity, $m^*(A) \geq m^*(A \cap E^c) = 0 + m^*(A \cap E^c) = m^*(A \cap E) + m^*(A \cap E^c)$. Thus, E is Lebesgue measurable by the simplified condition for Lebesgue measurability. □

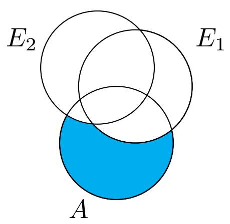
2. Countable sets are Lebesgue measurable.

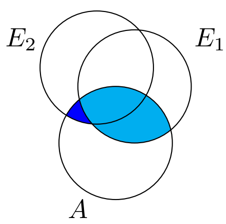
   Proof. Countable sets are of Lebesgue outer measure zero. And sets of Lebesgue outer measure zero are Lebesgue measurable. Thus, they are Lebesgue measurable. □

3. Finite union of Lebesgue measurable sets is Lebesgue measurable.

   Proof. It is enough to prove that if E_1 and E_2 are Lebesgue measurable, then their union is also Lebesgue measurable. Then, by finite mathematical induction, we can prove that the result if true for any finite collection of Lebesgue measurable sets.

   Suppose E_1, E_2 are Lebesgue measurable sets. Since E_1 is Lebesgue measurable,
$$m^*(A) = m^*(A \cap E_1) + m^*(A \cap E_1^c) \tag{10.8}$$
   And consider $A \cap E_1^c$ instead of A. Since E_2 is Lebesgue measurable, we get
$$m^*(A \cap E_1^c) = m^*(A \cap E_1^c \cap E_2) + m^*(A \cap E_1^c \cap E_2^c) \tag{10.9}$$
   We have $(A \cap E_1^c) \cap E_2^c = A \cap (E_1^c \cap E_2^c) = A \cap (E_1 \cup E_2)^c$. And $(A \cap E_1) \cup (A \cap E_1^c \cap E_2) = (A \cap E_1) \cup [A \cap (E_2 \cap E_1^c)] = A \cap (E_1 \cup E_2)$.





$$\begin{aligned} m^*(A) =& m^*(A \cap E_1) + m^*(A \cap E_1^c) \\ =& m^*(A \cap E_1) + m^*(A \cap E_1^c \cap E_2) + m^*(A \cap E_1^c \cap E_2^c) \\ \geq& m^*[A \cap (E_1 \cup E_2)] + m^*[A \cap (E_1 \cup E_2)^c] \end{aligned}$$

   Therefore $E_1 \cup E_2$ is Lebesgue measurable. And by finite induction, finite union of Lebesgue measurable sets is also Lebesgue measurable. □

4. Lebesgue Measure is finitely additive.
   In other words, Suppose $\{E_k\}_{k=1}^n$ be a finite collection of disjoint, Lebesgue measurable sets. Then Lebesgue measure of their union is the sum of Lebesgue measures.

   Proof. Let A be any subset of $\mathbb{R}$ and $\{E_k\}_{k=1}^n$ be a finite collection of disjoint, Lebesgue measurable subsets of $\mathbb{R}$.
$$\text{Claim : } m^*\left(A \cap \left[\bigcup_{k=1}^{\infty} E_k\right]\right) = \sum_{k=1}^{\infty} m^*(A \cap E_k) \tag{10.10}$$

Trivially, the claim is true for $n = 1$. Suppose the claim is true for $n-1$. That is,

$$m^*\left(A \cap \left[\bigcup_{k=1}^{n-1} E_k\right]\right) = \sum_{k=1}^{n-1} m^*(A \cap E_k) \tag{10.11}$$

From set theory we have,

$$A \cap \left[\bigcup_{k=1}^{n} E_k\right] \cap E_n = A \cap E_n \tag{10.12}$$

$$A \cap \left[\bigcup_{k=1}^{n} E_k\right] \cap E_n^c = A \cap \left[\bigcup_{k=1}^{n-1} E_n\right] \tag{10.13}$$

By Lebesgue measurability of E_n, we have

$$\begin{aligned} m^*\left(A \cap \left[\bigcup_{k=1}^{n} E_k\right]\right) =& m^*\left(A \cap \left[\bigcup_{k=1}^{n} E_k\right] \cap E_n\right) + m^*\left(A \cap \left[\bigcup_{k=1}^{n} E_k\right] \cap E_n^c\right) \\ &= m^*\left(A \cap E_n\right) + m^*\left(A \cap \left[\bigcup_{k=1}^{n-1} E_n\right]\right) \\ &= \sum_{k=1}^{n} m^*\left(A \cap E_k\right), \text{ by mathematical induction} \end{aligned}$$

Taking $A = \mathbb{R}$, we get Lebesgue measure is finitely additive. That is,

$$m^*\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{k=1}^{n} m^*\left(E_k\right) \tag{10.14}$$

□

5. Countable union of Lebesgue measurable sets is Lebesgue measurable

Proof. Let A be any subset of $\mathbb{R}$. And $\{E_k\}_{k=1}^{\infty}$ be a countable collection of disjoint, Lebesgue measurable subsets of $\mathbb{R}$. Define $F_n = \cup_{k=1}^{n} E_k$ and $E = \cup_{k=1}^{\infty} E_k$. Clearly, $F \subset E$ and $F^c \supset E^c$. Thus, $m^*(A \cap F_n^c) \geq m^*(A \cap E^c)$.

$$\begin{aligned} m^*(A) &= m^*(A \cap F_n) + m^*(A \cap F_n^c) \\ &\geq m^*\left(A \cap \left[\bigcup_{k=1}^{n} E_k\right]\right) + m^*(A \cap E^c) \\ &\geq \sum_{k=1}^{n} m^*(A \cap E_k) + m^*(A \cap E^c) \end{aligned}$$

$$\lim_{n\to\infty} m^*(A) \geq \lim_{n\to\infty} \sum_{k=1}^{n} m^*(A\cap E_k) + m^*(A\cap E^c)$$
$$m^*(A) \geq \sum_{k=1}^{\infty} m^*(A\cap E_k) + m^*(A\cap E^c)$$
$$\geq m^*\left(A\cap\left[\bigcup_{k=1}^{\infty} E_k\right]\right) + m^*(A\cap E^c)$$
$$\implies m^*(A) \geq m^*(A\cap E) + m^*(A\cap E^c)$$

By above inequality, $E = \cup_{k=1}^{\infty} E_k$ is Lebesgue measurable.

And, any countable collection of Lebesgue measurable sets can be expressed as a countable collection of disjoint, Lebesgue measurable sets. Let $\{E_k\}_{k=1}^{\infty}$ be a collection of Lebesgue measurable sets. Then, the countable collection, $\{E'_k\}_{k=1}^{\infty}$ defined by $E'_k = E_k - \cup_{j=1}^{k-1} E_k$ contains disjoint, Lebesgue measurable†[1] subsets of $\mathbb{R}$. Therefore, countable union of Lebesgue measurable sets is Lebesgue measurable. $\square$

6. Every interval is Lebesgue measurable.

Proof. It is sufficient to prove that (a,∞) is Lebesgue measurable. Suppose (a,∞) is Lebesgue measurable for every $a\in\mathbb{R}$. Then interval (a,b) is Lebesgue measurable, since $(a,b) = [(a,\infty)\cap(\mathbb{R}-(b,\infty))] - \{b\}$.

Let A be any subset of $\mathbb{R}$. Define $A_1 = A\cap(-\infty,a)$ and $A_2 = A\cap(a,\infty)$ such that $A - \{a\} = A_1\cup A_2$ and $A_1\cap A_2 = \phi$. And the interval (a,∞) is Lebesuge measurable only if

$$m^*(A) \geq m^*(A\cap(a,\infty)) + m^*(A\cap(a,\infty)^c) = m^*(A_1) + m^*(A_2)\dagger^2 \quad (10.15)$$

Let collection $\{I_k\}_{k=1}^{\infty}$ be a countable cover of A. Then collection $\{I'_k\}_{k=1}^{\infty}$ defined by $I'_k = I_k\cap(a,\infty)$ is a cover of A_1. And collection $\{I''_k\}_{k=1}^{\infty}$ defined by $I''_k = I_k\cap(-\infty,a)$ is a cover of A_2.

Clearly, $m^*(A_1) \leq \sum_{k=1}^{\infty} l(I'_k)$ and $m^*(A_2) \leq \sum_{k=1}^{\infty} l(I''_k)$.

$$m^*(A_1) + m^*(A_2) \leq \sum_{k=1}^{\infty} l(I'_k) + \sum_{k=1}^{\infty} l(I''_k)$$
$$\leq \sum_{k=1}^{\infty} l(I'_k) + l(I''_k)$$
$$\leq \sum_{k=1}^{\infty} l(I_k) \leq m^*(A)$$

[1] Suppose E_1, E_2 are measurable, then $E_2^c = \mathbb{R} - E_2$ is measurable by the duality of measurability condition. And $E_1\cap E_2^c = E_1 - E_2$ is Lebesgue measurable since countable intersection of Lebesgue measurable sets is Lebesgue measurable (by Property 4 and de Morgan's Law).

[2] We have, $m^*(A\cap(-\infty,a]) = m^*(A\cap(-\infty,a)$ since removing finite number of points from a subset of $\mathbb{R}$ won't affect its Lebesgue measure.

Thus, (a, ∞) is Lebesgue measurable. Therefore, every interval is Lebesgue measurable. $\square$

7. σ-algebra of Lebesgue measurable sets $\mathcal{M}$ contains Borel Sets $\mathcal{B}$.

$$\mathcal{B} \subset \mathcal{M} \tag{10.16}$$

Proof. The Borel algebra $\mathcal{B}$ is the σ-algebra containing all intervals. We have proved that, every intervals are Lebesgue measurable. Also, we have proved that the set of all Lebesgue measurable subsets of $\mathbb{R}$ is a σ-algebra as complements of Lebesgue measurable sets are Lebesgue measurable by duality of the condition and countable union of Lebesgue measurable sets are also Lebesgue measurable. Therefore, every Borel set is Lebesgue measurable. Clearly, G_δ and F_σ are Borel sets and are Lebesgue measurable. $\square$

8. Lebesgue Measurability is translation invariant.

Proof. Let E be a Lebesgue measurable set. Let A be any subset of $\mathbb{R}$ and $y \in \mathbb{R}$. Then,

$$\begin{aligned} m^*(A) &= m^*(A - y) \\ &= m^*((A - y) \cap E) + m^*((A - y) \cap E^c) \\ &= m^*(A \cap (E + y)) + m^*(A \cap (E + y)^c) \end{aligned}$$

Thus, $E+y$ is Lebesgue measurable. And Lebesgue measurability is traslation invariant. $\square$

Exercise

11. Let $\mathcal{A}$ be the σ-algebra containing all intervals of the form (a, ∞). Every interval has one of the four forms,

$$(a, b] = (a, \infty) \cap (b, \infty)^c \tag{10.17}$$

$$(a, b) = (a, \infty) \cap \left[\bigcap_{k=1}^{\infty} \left(b - \frac{1}{k}, \infty\right)\right]^c \tag{10.18}$$

$$[a, b] = \left[\bigcap_{k=1}^{\infty} \left(a - \frac{1}{k}, \infty\right)\right] \cap (b, \infty)^c \tag{10.19}$$

$$[a, b) = \left[\bigcap_{k=1}^{\infty} \left(a - \frac{1}{k}, \infty\right)\right] \cap \left[\bigcap_{k=1}^{\infty} \left(b - \frac{1}{k}, \infty\right)\right]^c \tag{10.20}$$

12. **Borel sets** is the σ-algebra containing all open intervals. The equations from previous equations are sufficient. However, we have a simpler form for closed intervals, $[a, b]$.

$$[a, b] = [(-\infty, a) \cup (b, \infty)]^c \tag{10.21}$$

Clearly, every interval is a Borel set.

13. - $C \in F_\sigma \implies C = \bigcup_{k=1}^{\infty} C_k \implies \bigcup_{k=1}^{\infty} (C_k + y) = C + y \in F_\sigma$
    - $O \in G_\delta \implies O = \bigcap_{k=1}^{\infty} O_k \implies \bigcup_{k=1}^{\infty} (O_k + y) = O + y \in G_\delta$
    - $m^*(E) = 0 \implies 0 = \inf\{\sum_{k=1}^{\infty} l(I_k)\} = \inf\{\sum_{k=1}^{\infty} l(I_k + y)\} \implies m^*(E + y) = 0$

14. A subset E has positive Lebesgue measure if and only if it has a bounded subset of positive Lebesgue measure.

$$m^*(E) > 0 \iff \exists \text{ bounded subset } F \subset E, \text{ such that } m^*(F) > 0$$

Sufficient Part : By monotonicity, $F \subset E \implies m^*(F) \leq m^*(E)$. And $m^*(F) > 0 \implies 0 < m^*(F) \leq m^*(E)$.
Necessary Part :

15.

10.2.4 Outer and Inner Approximation

Definitions 10.2.4 (Excision Property)**.** Let A be a Lebesgue measurable set of finite measure and $A \subset B$. Then,

$$m^*(B \sim A) = m^*(B) - m^*(A) \tag{10.22}$$

Proof.

$$\begin{aligned} m^*(B) =& m^*(B \cap A) + m^*(B \cap A^c) \\ =& m^*(A) + m^*(B \sim A) \\ \implies m^*(B \sim A) =& m^*(B) - m^*(A) \end{aligned}$$

□

Theorem 10.2.1 (approximation)**.** *The following conditions are equivalent to Lebesgue measurability of E*

1. *$\forall \varepsilon > 0$, there is an open set $\mathcal{O}$ containing E for which $m^*(\mathcal{O} \sim E) < \varepsilon$.*

2. *There is a G_δ set G containining E such that $m^*(G \sim E) = 0$.*

3. *$\forall \varepsilon > 0$, there is a closed set F contained in E such that $m^*(E \sim F) < 0$.*

4. *There is an F_σ set F contained in E sucht hat $m^*(E \sim F) = 0$.*

Note : Equivalent conditions 1 & 2 are about outer approximation of a Lebesgue measurable set and 3 & 4 are about inner approximation of a Lebesgue measurable set.

Proof. **Measurability $\implies$ Open set - Outer approximation**

Suppose that E is Lebesgue measurable. And let $\varepsilon > 0$.

Case 1 : Suppose $m^*(E) < \infty$. By definition of Lebesgue outer measure, there exists an open cover $\{I_k\}_{k=1}^{\infty}$ such that $\sum_{k=1}^{\infty} l(I_k) < m^*(E) + \varepsilon$.

Define $\mathcal{O} = \bigcup_{k=1}^{\infty} I_k$. Then $\mathcal{O}$ is an open set containing E and thus Lebesgue measurable. Also, $m^*(\mathcal{O}) \leq \sum_{k=1}^{\infty} l(I_k) < m^*(E) + \varepsilon$. Therefore, by excision property we have $m^*(\mathcal{O} \sim E) = m^*(\mathcal{O}) - m^*(E) < \varepsilon$.

Case 2 : Suppose $m^*(E) = \infty$. Without loss of generality, E may be written as countable union Lebesgue measurable sets $\{E_k\}_{k=1}^{\infty}$ of finite measure.

By case 1, for every k, there exists $\mathcal{O}_k$ for each E_k of finite measure such that $m^*(\mathcal{O}_k \sim E_k) < \frac{\varepsilon}{2^k}$. Define $\mathcal{O} = \bigcup_{k=1}^{\infty} \mathcal{O}_k$. Then $\mathcal{O}$ is open, contains E and

$$
\begin{aligned}
m^*(\mathcal{O} \sim E) =& m^*\left(\bigcup_{k=1}^{\infty} \mathcal{O}_k \sim E\right) \\
\leq& m^*\left(\bigcup_{k=1}^{\infty} \mathcal{O}_k \sim E_k\right) \\
\leq& \sum_{k=1}^{\infty} m^*(\mathcal{O}_k \sim E_k) = \varepsilon \sum_{k=1}^{\infty} \frac{1}{2^k} = \varepsilon
\end{aligned}
$$

Open, Outer approximation $\implies$ G_δ, Outer approximation

Let E be a subset of real numbers such that Lebesgue measure of E has an open set inner approximation. That is, for every $\varepsilon > 0$, there exists an open set $\mathcal{O}$ such that $m^*(\mathcal{O} \sim E) < \varepsilon$.

Let $\mathcal{O}_k$ be open sets such that $m^*(\mathcal{O}_k \sim E) < \frac{1}{k}$. Define $G = \bigcap_{k=1}^{\infty} \mathcal{O}_k$. Then, G is a G_δ set containing E. And $G \sim E \subset \mathcal{O}_k \sim E$. Thus, by monotonicity, $m^*(G \sim E) \leq m^*(\mathcal{O}_k \sim E) < \frac{1}{k}$. Thus, we have a G_δ set G containing E such that $m^*(G \sim E) = 0$.

G_δ-Outer approximation $\implies$ Measurability

We have, $m^*(G \sim E) = 0$. Since every set of Lebesgue measure zero is Lebesgue measurable, $G \sim E$ is Lebesgue measurable. And its complement $(G \sim E)^c$ is also Lebesgue measurable. Also we have, G is a G_δ set, thus a Borel set and hence Lebesgue measurable.

Clearly,we have $E = G \cap (G \sim E)^c$. And therefore, E is Lebesgue measurable.

Open, Outer approximation $\iff$ Closed, Inner approximation

By duality of Lebesgue measurability, E is Lebesgue measurable if and only if E^c is Lebesgue measurable. And by de Morgan's Law, we have E^c has an open set - outer approximation $\mathcal{O}$ if and only if E has a closed set - inner approximation $\mathcal{O}^c$.

$$m^*(\mathcal{O} \sim E^c) < \varepsilon \iff m^*(E \sim \mathcal{O}^c) < \varepsilon$$

G_δ-Outer approximation $\iff$ F_σ-Inner approximation

Again, by duality of Lebesgue measurability and de Morgan's Law, we have E^c has a G_δ-outer approximation G if and only if E has an F_σ-inner approximation F.

$$m^*(G \sim E^c) = 0 \iff m^*(E \sim F) = 0$$

□

Theorem 10.2.2. *Let E be a Lebesgue measurable set of finite measure. Then for any $\varepsilon > 0$, there exists a finite collection of disjoint open sets $\{I_k\}_{k=1}^n$ such that $\mathcal{O} = \bigcup_{k=1}^{n} I_k$ and $m^*(\mathcal{O} \sim E) + m^*(E \sim \mathcal{O}) < \varepsilon$.*

Proof. Since E is Lebesgue measurable, by outer approximation theorem we have an open set U such that $E \subset U$ and

$$m^*(U \sim E) < \frac{\varepsilon}{2}$$

$$U = (U \cap E) \cup (U \cap E^c)$$

Since E is Lebesgue measurable, $m^*(E) < k$

$$\begin{aligned} m^*(U) &= m^*(U \cap E) + m^*(U \cap E^c) \\ &= m^*(E) + m^*(U \sim E) \end{aligned}$$

Since E has finite measure

$$\begin{aligned} m^*(U) &= m^*(E) + m^*(U \sim E) \\ &< k + \frac{\varepsilon}{2} < \infty \end{aligned}$$

That is, U is of finite measure.

Since U is open, U is countable†[3] union of a disjoint collection of open intervals, say $\{I_k\}_{k=1}^{\infty}$. Clearly,

$$\sum_{k=1}^{n} l(I_k) \leq \sum_{k=1}^{\infty} l(I_k) \leq m^*(U) < \infty$$

[3]By definition, **Open sets** are countable union of disjoint, open intervals

By characterisation of series convergence, there exists an integer n such that,

$$\sum_{k=n+1}^{\infty} l(I_k) < \frac{\varepsilon}{2}$$

Define $\mathcal{O} = \bigcup_{k=1}^{n} I_k$. Since $\mathcal{O} \sim E \subset U \sim E$, by monotonicity we have

$$m^*(\mathcal{O} \sim E) \leq m^*(U \sim E) < \frac{\varepsilon}{2} \tag{10.23}$$

Since $E \subset U$, we have $E \sim \mathcal{O} \subset U \sim \mathcal{O} = \bigcup_{k=1}^{n} I_k$. And clearly,

$$U \sim \mathcal{O} = \bigcup_{k=1}^{\infty} I_k \sim \bigcup_{k=1}^{n} I_k \subseteq \bigcup_{k=n+1}^{\infty} I_k$$

$$\text{Thus, } m^*(E \sim \mathcal{O}) \leq m^*(U \sim \mathcal{O}) \leq \sum_{k=n+1}^{\infty} l(I_k) < \frac{\varepsilon}{2} \tag{10.24}$$

Therefore,

$$m^*(E \sim \mathcal{O}) + m^*(\mathcal{O} \sim E) < \varepsilon$$

$\square$

Exercise

17. Let $\varepsilon > 0$ and E is Lebesgue measurable. Then there exists open set $\mathcal{O}$ and closed set F such that $F \subset E \subset \mathcal{O}$, $m^*(E \sim F) < \frac{\varepsilon}{2}$ and $m^*(\mathcal{O} \sim E) < \frac{\varepsilon}{2}$. Clearly, $\mathcal{O} \sim E$ and $E \sim F$ are disjoint and $\mathcal{O} \sim F = (\mathcal{O} \sim E) \cup (E \sim F)$. Thus by monotonicity of Lebesgue outer measure, we have $m^*(\mathcal{O} \sim F) \leq m^*(\mathcal{O} \sim E) + m^*(E \sim F) < \varepsilon$.

18. Suppose E has finite outer measure. †[4]
G_δ **set :** Let $\varepsilon > 0$. Then by the definition of Lebesgue outer measure, there exists a cover $\{I_k\}_{k=1}^{\infty}$ of E such that $\sum_{k=1}^{\infty} l(I_k) < m^*(E) - \frac{\varepsilon}{2}$. Define $G = \bigcup_{k=1}^{\infty} I_k$. Then G is a G_δ set and $m^*(G) \leq \sum_{k=1}^{\infty} l(I_k) < m^*(E) - \frac{\varepsilon}{2}$.

    F_σ **set :**

19. Let E be a set of finite outer measure. Suppose E is not Lebesgue measurable. And $\mathcal{O}$ be an open set containing E. Then $\mathcal{O} = (\mathcal{O} \sim E) \cup E$. By monotonicity, $m^*(\mathcal{O}) \leq m^*(\mathcal{O} \sim E) + m^*(E)$ †[5]. Since $m^*(E)$ is finite, we have $m^*(\mathcal{O}) - m^*(E) \leq m^*(\mathcal{O} \sim E)$.

[4] E has finite outer measure does not imply E is bounded or Lebesgue measurable.

[5] I am not able to change $\leq$ into $<$ as non-measurability doesn't mean that for this particular $\mathcal{O}$ the sum of Lebesgue outer measures should be greater. There may be a better proof.

20. Let E be a set of finite outer measure. Suppose E is Lebesgue measurable. Let (a,b) be any open, bounded interval. Then by the definiton of Lebesgue measurability, $b-a = m^*(a,b) = m^*((a,b)\cap E)+m^*((a,b)\cap E^c)$.

21. A subset E is Lebesgue measurable if there exists a G_δ set G containing E such that $m^*(G \sim E) = 0$. Suppose E_1 and E_2 are Lebesgue measurable sets. Then, we have G_δ sets G_1 and G_2. And two countable family of open intervals $\{\mathcal{O}_{1,k}\}_{k=1}^{\infty}$ and $\{\mathcal{O}_{2,k}\}_{k=1}^{\infty}$ such that $\cap\mathcal{O}_{1,k} = G_1$ and $\cap\mathcal{O}_{2,k} = G_2$. Let $\mathcal{O} = \{\mathcal{O}_k\}_{k=1}^{\infty}$ be collection of open intervals in $\mathcal{O}_1$ and $\mathcal{O}_2$. Define $G = \cap_{k=1}^{\infty}\mathcal{O}_k$. Then, $E = E_1 \cup E_2 \subset G_1 \cup G_2 = G$. Since $G \sim E = (G_1 \sim E_1) \cup (G_2 \cup E_2)$, by monotonicity of Lebesgue outer measure we have $m^*(G \sim E) \leq m^*(G_1 \sim E_1) + m^*(G_2 \sim E_2) = 0$.

22. Let m^{**} be a non-negative set function defined by $m^{**}(A) = \inf\{m^*(\mathcal{O}) : A \subset \mathcal{O}, \mathcal{O} \text{ is open}\}$. Suppose E is Lebesgue measurable. Then, by open set - outer approximation theorem we have $m^*(E) \leq m^{**}(E) < m^*(E)+\varepsilon$ for any $\varepsilon > 0$. Thus, $m^{**}(E) = m^*(E)$.

    In other words, for Lebesgue measurable sets m^* and m^{**} are the same.

23. Let m^{***} be a non-negative set function defined by $m^{***}(E) = \sup\{m^*(F) : F \subset E, F \text{ is closed }\}$. Let E be Lebesgue measurable set. Then, by closed set - inner approximation theorem we have $m^*(E) \geq m^{***}(E) > m^*(E) - \varepsilon$.

    In other words, for Lebesgue measurable sets m^* and m^{***} are the same.

10.2.5 Further Properties

The Lebesgue measure has the following properties.

1. Every Borel set is Lebesgue measurable.
2. Lebesgue Measure of an interval is its length.
3. Lebesuge Measure is translation invariant.
4. Lebesuge Measure is countably additive.
5. There exists non-measurable(Lebesgue) sets. eg. $C_E \subset E$
6. There exists uncountable set of zero measure. eg. Cantor set

Countable Subadditivity

Theorem 10.2.3. *The set function Lebesgue measure defined on σ algebra of Lebesgue measurable sets 1. assigns length to any interval, 2. is translation invariant, and 3. is countably additive.*

Proof. **Length of interval**
Let E be an interval, then E belongs to the σ algebra of Lebesgue measurable sets as Borel sets are Lebesgue measurable. Also, we have $m^*(E) = m(E)$ for any Lebesgue measurable set E. And Lebesgue outer measure m^* of an interval

is its length. Therefore, Lebesgue measure of any interval is its length.

Translation invariant
Let E be a Lebesgue measurable set. We have, $E + y$ is also Lebesgue measurable. Since $E + y$ is Lebesgue measurable and Lebesgue outer measure is translation invariant, we have $m^*(E) = m^*(E + y) = m(E + y)$. Clearly, $m(E) = m(E + y)$.

Countably additive [6]
Let $\{E_k\}_{k=1}^{\infty}$ be a countable family of disjoint, Lebesgue measurable sets. Lebesgue outer measure is countably subadditive and countable union of Lebesgue measurable sets is also Lebesgue measurable. Thus we have,

$$m\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty} m(E_k)$$

Since Lebesgue measure is finitely additive we have,

$$\begin{aligned} m\left(\bigcup_{k=1}^{\infty} E_k\right) \geq m\left(\bigcup_{k=1}^{n} E_k\right) &= \sum_{k=1}^{n} m(E_k) \\ \lim_{n\to\infty} m\left(\bigcup_{k=1}^{\infty} E_k\right) \geq \lim_{n\to\infty} m\left(\bigcup_{k=1}^{n} E_k\right) &= \lim_{n\to\infty} \sum_{k=1}^{n} m(E_k) \\ \implies m\left(\bigcup_{k=1}^{\infty} E_k\right) &\geq \sum_{k=1}^{\infty} m(E_k) \end{aligned}$$

Therefore, Lebesgue measure is countably additive.

$$m\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} m(E_k)$$

□

Continuity of Lebesgue measure

Theorem 10.2.4 (continuity)**.** *Let m be Lebesgue measure.*

1. *Suppose $\{A_k\}_{k=1}^{\infty}$ be an ascending [7] collection of Lebesgue measurable sets. Then,*

$$m\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{k\to\infty} m(A_k) \tag{10.25}$$

2. *Suppose $\{B_k\}_{k=1}^{\infty}$ be an descending [8] collection of Lebesgue measurable sets and $m(B_1) < \infty$. Then,*

$$m\left(\bigcap_{k=1}^{\infty} B_k\right) = \lim_{k\to\infty} m(B_k) \tag{10.26}$$

[6] A real-valued set function m is countably additive if $m(\cup_{k=1}^{\infty} E_k) = \sum_{k=1}^{\infty} m(E_k)$ for any disjoint family of sets $\{E_k\}$.
[7] $\{A_k\}$ is ascending if $A_1 \subset A_2 \subset \dots$
[8] $\{A_k\}$ is ascending if $B_1 \supset B_2 \supset \dots$

Proof. **Ascending Collection**
Let $\{A_k\}_{k=1}^{\infty}$ be an ascending collection of Lebesgue measurable sets. Define $A_0 = \phi$.

Case 1 : $\exists k' \in \mathbb{N},\ m(A_k) = \infty$
Suppose the collection has a Lebesgue measurable set $A_{k'}$ of infinite measure. Then for $\forall k \geq k',\ m(A_k) = \infty$. Clearly,

$$\lim_{k\to\infty} m(A_k) = \infty = m(A_k) \leq m\left(\bigcup_{k=1}^{\infty} A_k\right)$$

Case 2 : $\forall k \in \mathbb{N},\ m(A_k) < \infty$
Suppose that every Lebesgue measurable set in the collection is of finite measure. Consider the ascending collection of disjoint Lebesgue measurable sets, $\{C_k\}_{k=1}^{\infty}$ given by $C_k = A_k \sim A_{k-1}$. Clearly, $\cup_{k=1}^{\infty} A_k = \cup_{k=1}^{\infty} C_k$. By countable additiviy of Lebesgue measure, we have

$$m\left(\bigcup_{k=1}^{\infty} A_k\right) = m\left(\bigcup_{k=1}^{\infty} C_k\right) = \sum_{k=1}^{\infty} m(A_k \sim A_{k-1})$$

Also we have,

$$\begin{aligned}\sum_{k=1}^{\infty} m(A_k \sim A_{k-1}) &= \sum_{k=1}^{\infty} m(A_k) - m(A_{k-1}) \\ &= \lim_{n\to\infty} \sum_{k=1}^{n} m(A_k) - m(A_{k-1}) \\ &= \lim_{n\to\infty} m(A_n) - m(A_0)\end{aligned}$$

Therefore, $m\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{n\to\infty} A_n$.

Descending Collection Let $\{B_k\}$ be a descending collection of Lebesgue measurable sets and $m(B_1) < \infty$. Consider the asceding collection of Lebesgue measurable sets, $\{D_k\}_{k=1}^{\infty}$ given by $D_k = B_1 \sim B_k$. By the continuity of Lebesgue measure for ascending collection of sets, we have

$$m\left(\bigcup_{k=1}^{\infty} D_k\right) = \lim_{n\to\infty} m(D_n) = m(B_1) - \lim_{n\to\infty} m(B_n)$$

By de Morgan's law, we have $B_1 - \cap_k B_k = \cup_k (B_1 - B_k)$. Since B_1 is of finite measure, by excision property we have,

$$m\left(\bigcap_{k=1}^{\infty} B_k\right) = m\left(B_1 - \bigcup_{k=1}^{\infty} D_k\right) = \lim_{n\to\infty} m(B_n)$$

□

Borel-Cantelli Lemma

Definitions 10.2.5 (ae)**.** A property of real numbers is true except for a set of zero measure, then it is true **a**lmost **e**verywhere.

Lemma 10.2.5 (Borel-Cantelli)**.** *Let $\{E_k\}_{k=1}^{\infty}$ be a countable colection of Lebesgue measurable sets for which $\sum_{k=1}^{\infty} m(E_k) < \infty$. Then, almost all $x \in \mathbb{R}$ belongs to at most finitely many of the E_k's.*

Proof. We have $\sum_{k=1}^{\infty} m(E_k) < \infty$. Then, by convergence of the series

$$\lim_{n\to\infty} \sum_{k=n}^{\infty} m(E_k) = 0$$

Also, $\left\{ \bigcup_{k=n}^{\infty} E_k \right\}_{n=1}^{\infty}$ is a descending collection of Lebesgue measurable sets. By continuity of Lebesgue measure for descending collection of Lebesgue measurable sets, we have

$$m\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} E_k\right) = \lim_{n\to\infty} m\left(\bigcup_{k=n}^{\infty} E_k\right) = 0$$

Clearly, $\lim_{n\to\infty} \bigcup_{k=n}^{\infty} E_k$ is a set of zero measure. Suppose $x \in \mathbb{R}$ belongs to countably many E_k's. Then, for any $m \in \mathbb{N}$, there exists $k > m$ such that $x \in E_k$. Clearly, $x \in \lim_{n\to\infty} \bigcup_{k=n}^{\infty} E_k$. That is, x belongs to a set of zero measure. Therefore by contrapositivity, if x does not belong to a set of measuare zero, then x belongs to at most finitely many E_k's. In other words, almost every x in $\mathbb{R}$ belongs to at most finitely many E_k's. □

Exercise

24. $m(E_1 \cup E_2) + m(E_1 \cap E_2) = m(E_1) + m(E_2)$

25. $m(B_1) < \infty$ is necessary for continuity property of measure for descending collection of measurable sets.

26. $$m^*\left(A \cap \bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} m^*(A \cap E_k)$$

27. Let m' be set function on a σ-algebra and m' is countably additive.

    (a) m' is finitely additive, monotone, countably monotone, and has excision property

    (b) m' has continuity properties

28. continuity + finite additivity $\implies$ countable additivity

10.2.6 Non-measurable sets

Every measurable set of positive measure contains a non-measurable set.

Definitions 10.2.6 (Rational Equivalence)**.** Let E be any subset of $\mathbb{R}$. The relation $xRy \iff x - y \in \mathbb{Q}$ is an equivalence†[9] relation on $\mathbb{R}$.

Definitions 10.2.7 (Choice set, C_E)**.** Let E be any subset of $\mathbb{R}$ and R be an equivalence relation on E. By axiom of choice, there exists a choice set $C_E \subset E$ containing an exactly an element from each equivalence class.

Definitions 10.2.8 (translate)**.** Let E be a subset of $\mathbb{R}$. Let $\lambda \in \mathbb{R}$. Then $\lambda + E = \{\lambda + x : x \in E\}$ is a translate of E.

With the help of following lemma, we prove that for any measurable set E of positive measure, the subset C_E is non-measurable.

Lemma 10.2.6. *Let E be a bounded, measurable set. Suppose there exists a bounded, countably infinite set Λ for which the collection of translates of E under Λ, $\{\lambda + E\}_{\lambda\in\Lambda}$ are disjoint. Then $m(E) = 0$.*

In other words, if a bounded measurable set has countably many disjoint translates, then it is of measure zero. That is, there doesn't exists a bounded set of positive measure which has countably many disjoint translates.

Proof. The Lebesgue measure is translation invariant. Thus, the translates, $\lambda + E$ are measurable and $m(\lambda + E) = m(E)$.

We have, E and Λ are bounded, $\bigcup\limits_{\lambda\in\Lambda} \lambda+E$ is bounded and is of finite measure. The Lebesgue measure is countably additive. And since translates are disjoint,

$$m\left[\bigcup_{\lambda\in\Lambda} \lambda + E\right] = \sum_{\lambda\in\Lambda} m(\lambda + E) < \infty$$

Clearly, $m(E) = 0$. Suppose $m(E) = \varepsilon$. Then $\sum\limits_{\lambda\in\Lambda} m(\lambda + E) = \sum\limits_{\lambda\in\Lambda} \varepsilon = \infty$ since Λ has countably infinite elements. □

Theorem 10.2.7 (Vitali)**.** *Any set E of real numbers with positive outer measure contains a subset that fails to be measurable.*

More importantly, every measurable set of positive measure contains as non-measurable set.

Proof. **Case 1 : E is bounded and non-measurable**
Suppose E is not measurable. Then $E \subset E$ is a non-measurable set and the result is trivial.

Case 2 : E is bounded and measurable
Suppose E is a bounded, measurable subset of positive measure. Let C_E be a

[9] $x - x = 0 \in \mathbb{Q}$, $(y - x) = -(x - y) \in \mathbb{Q}$ and $x - z = (x - y) + (y - z) \in \mathbb{Q}$

choice set of E under rational equivalence. Let Λ be any bounded, countably infinite set of rational numbers. Clearly, the translates of C_E under Λ are disjoint.

Suppose $x \in (\lambda_1 + C_E) \cap (\lambda_2 + C_E)$. Then, $x = \lambda_1 + y = \lambda_2 + z$. Clearly, $y = z$ since $y - z \in \mathbb{Q}$ and we chose precisely one element from each equivalence class. Again, $y = z \implies \lambda_1 = \lambda_2$. In other words, the intersecting the translates are identical. That is, two distinct translates will be disjoint.

Suppose C_E is measurable. Since C_E and Λ are bounded and C_E has countably many disjoint translates under Λ, by lemma $m(C_E) = 0$. However, $E \subset \bigcup\limits_{\lambda \in [-2b, 2b] \cap \mathbb{Q}} \lambda + C_E$ for sufficiently large†[10] $b \in \mathbb{R}$.

$$m(E) \leq m\left(\bigcup_{\lambda \in [-2b,2b] \cap \mathbb{Q}} \lambda + C_E\right) = \sum_{\lambda \in [-2b,2b] \cap \mathbb{Q}} m(\lambda + C_E) = 0$$

which is a contradiction since E has positive measure.

Case 3 : E is unbounded and measurable
Suppose E is an unbounded subset of positive outer measure, then by the definition of Lebesuge outer measure, E has a bounded subset of positive outer measure. And by case 1 & 2, this set has a non-measurable subset. $\square$

Theorem 10.2.8. *There are disjont subsets A, B of real numbers for which*

$$m^*(A \cup B) < m^*(A) + m^*(B) \tag{10.27}$$

Proof. Suppose that for every disjoint pair of subsets $A, B \subset \mathbb{R}$, $m^*(A \cup B) = m^*(A) + m^*(B)$. Then, by the definition of Lebesgue measurability, every subset of real numbers is measurable. By, Vital's theorem there does exist non-measurable subsets of real numbers which is a contradiction. Therefore, there does exists disjoint subsets A, B such that $m^*(A \cup B) \neq m^*(A) + m^*(B)$. By subadditivity of Lebesgue outer measure, we have $m^*(A \cup B) \leq m^*(A) + m^*(B)$. Thererfore, $m^*(A \cup B) < m^*(A) + m^*(B)$. $\square$

Exercise

29. (a)
    (b) Rational Equivalence on $\mathbb{Q}$ gives singleton choice set as difference two rational numbers is always rational. $\frac{a}{b} - \frac{c}{d} = \frac{ad-bc}{bd}$. Thus, $\{0\}$ is a choice set.
    (c) Difference two numbers being irrational is not an equivalence relation as it violates transitivity. $x - y, y - z \notin \mathbb{Q} \not\Longrightarrow x - z \notin \mathbb{Q}$. For example, $\sqrt{2} - \sqrt{3}, \sqrt{3} - \sqrt{2} \notin \mathbb{Q}$. However, $\sqrt{2} - \sqrt{2} = 0 \in \mathbb{Q}$.
30.
31.
32.
33.

[10]Since E is bounded there exists $b \in \mathbb{R}$ such that $E \subset [-b, b]$

10.2.7 Cantor Set and Cantor-Lebesgue Function

Cantor set, C

We know that every countable set is of measure zero. However, subsets of zero measure are not necessarily countable. Cantor set is an uncountable set of zero measure.

Definitions 10.2.9 (Cantor set)**.** Consider unit interval, $I = [0, 1]$. Let $C_0 = I$. Let $\{I_k\}$ be the collection of subintervals in C_k. Construct C_{k+1} recursively by removing subintervals of length $\frac{l(I_k)}{3}$ from the middle of each I_k in C_k. Cantor set C is given by,

$$C = \bigcap_{k=1}^{\infty} C_k \tag{10.28}$$

Note that C_k are descending collection of 2^k disjoint, closed intervals, each of length $\frac{1}{3^k}$. Thus, effective length of C_k is $(\frac{2}{3})^k$.

Theorem 10.2.9. *Cantor set C is a closed, uncountable set of measure zero.*

Proof. **Step 1 : C is measurable**
By construction, Cantor set is an intersection of (countably many) closed subsets of real numbers. Since intersections of closed sets are closed, Cantor set is also closed. Also every closed subset is a Borel set and therefore measurable.

Step 2 : $m(C) = 0$
From the construction of C, we have $m(C_k) = \left(\frac{2}{3}\right)^k$. Clearly, $\{C_k\}_{k=1}^{\infty}$ is a descending collection of measurable sets and $m(C_1) < \infty$. Thus by continuity of Lebesgue measure,

$$m(C) = m\left(\bigcap_{k=1}^{\infty} C_k\right) = \lim_{k\to\infty} m(C_k) = \lim_{k\to\infty}\left(\frac{2}{3}\right)^k = 0$$

Step 3 : C is uncountable
Suppose C is countable. Then elements of C can be enumerated. That is, $C = \{x_k\}_{k=1}^{\infty}$. We have, $x_1 \in C \implies x_1 \in C_1$. There are two disjoint intervals in C_1. Clearly, C_1 has an interval F_1 which doesn't contain x_1. Similarly, there exists a closed interval, F_2 in C_2 such that $x_2 \notin F_2$ and $F_2 \subset F_1$. Continuing like this, we get a descending collection of closed intervals $\{F_k\}_{k=1}^{\infty}$ sucht that $x_k \notin F_k$.

By nested set theorem, intersection of descending collection of closed and bounded intervals is non-empty. Thus, $\bigcap_{k=1}^{\infty} F_k \neq \phi$. Let $x \in \bigcap_{k=1}^{\infty} F_k$. Clearly $x \neq c_k$ for any k as $c_k \notin F_k \implies c_k \notin \bigcap_{k=1}^{\infty} F_k$. However, $x \in C$ which is a contradiction to the assumption that $C = \{c_k : k = 1, 2, \dots\}$. Therefore, elements of C can't be enumerated. In other words, C is uncountable. $\square$

Cantor-Lebesgue Function, φ

Definitions 10.2.10 (increasing)**.** A real-valued function f is increasing if $f(x) \geq f(y)$ for every $x > y$.

Definitions 10.2.11 (strictly increasing)**.** A real-valued function f is strictly increasing if $f(x) > f(y)$ for every $x > y$.

Definitions 10.2.12 (Cantor-Lebesgue function φ)**.** Let C be the Cantor set. Define open set, $\mathcal{O} = I \sim C$.

$$\mathcal{O} = I \sim C = I \sim \left(\bigcap_{k=1}^{\infty} C_k \right) = \bigcup_{k=1}^{\infty} (I \sim C_k) = \bigcup_{k=1}^{\infty} \mathcal{O}_k$$

We know that $\mathcal{O}_k$ has 2^{k-1} open intervals $I_{k,1}, I_{k,2}, \ldots, I_{k,2^{k-1}}$. Define Cantor-Lebesgue function φ on $\mathcal{O}$ by $\varphi(x) = \frac{m}{2^k}$ for every $x \in I_{k,m}$. We may extend φ to unit interval such that $\varphi(0) = 0$ and $\varphi(x) = \sup \{\varphi(t) : t \in [0, x] \cap \mathcal{O}\}$.

Clearly, by the construction ϕ is an increasing real-valued function.

Theorem 10.2.10 (Properties of φ)**.** *Cantor-Lebesgue function $\varphi : I \to I$ is a surjection. And φ is differentiable in $\mathcal{O}$.*

Proof. We know from the definition of φ on $\mathcal{O}$, it is increasing on $\mathcal{O}$. And for extending φ from $\mathcal{O}$ to $[0, 1]$ we are considering the supremum(least upper bound) of all the previous values of φ on $\mathcal{O}$. Thus, function φ is increasing on the unit interval, $[0, 1]$.

For continuity of φ, it is enough to prove that φ doesn't have any jump discontinuities†[11] as it is an increasing function. Suppose $x \in C$ and $x \neq 0$. Then for sufficiently large k, x lies between two consecutive intervals of $\mathcal{O}_k$. Let a_k and b_k be the upper and lower bounds of intervals to the left and right of x in $\mathcal{O}_k$. Then $x \in (a_k, b_k)$.

We know that, $\varphi(b_k) - \varphi(a_k) = \frac{1}{2^k}$. And $\varphi(a_k) < \varphi(x) < \varphi(b_k)$, by the construction of φ. As $k \to \infty$, the jump $\lim_{k\to\infty} \varphi(b_k) - \varphi(a_k) = \lim_{k\to\infty} \dfrac{1}{2^k} = 0$. Thus, φ doesn't have any jump discontinuities. Therefore, φ is continuous on $[0, 1]$.

Clearly, φ is constant on each interval of $\mathcal{O}$. Therefore, its derivative exists and equals to zero everywhere on $\mathcal{O}$.

We have, $\mathcal{O} = [0, 1] \sim C$. Thus, $m(\mathcal{O}) = 1$, since $m(C) = 0$ and $m([0, 1]) = 1$. Also, φ is a an increasing, continuous function from $[0, 1]$ into $[0, 1]$ where $\varphi(0) = 0$ and $\varphi(1) = 1$. Therefore, Cantor-Lebesgue function, φ is onto unit interval $[0, 1]$. □

[11] If the function φ has unbounded variation(jump) at some points of $[0, 1]$. That is $\exists \varepsilon > 0$ such that $\forall \delta > 0, \exists x \in [0, 1]$ where $f(x + \varepsilon/2) - f(x - \varepsilon/2) > \delta$. —need revisit—

Measurable, non-Borel set

We already saw that there exists non-measurable subsets. Using the properties of the function ψ, we assert that there exist a measurable set $\psi^{-1}(W)$ which is not a Borel set.

Theorem 10.2.11. *The function $\psi : [0,1] \to [0,2]$ defined by $\psi(x) = \varphi(x) + x$*

1. *is a strictly increasing†*[12] *function which maps $[0,1]$ onto $[0,2]$ and*
2. *ψ maps Cantor set, C onto a set of positive measure and*
3. *ψ maps a measurable†*[13] *subset of C onto a non-measurable set.*

Proof. By definition, $\psi(x) = \varphi(x) + id(x)$. We know that the sum of continuous functions is continuous. Thus, ψ is continuous. Again, the sum of an increasing function and strictly increasing function is always strictly increasing. Thus, ψ is a strictly increasing function.

We know that ψ is only translating open intervals in $\mathcal{O}$. That is, $\psi(I_{k,m}) = I_{k,m} + \frac{m}{2^k}$. In other words, ψ translates every interval I_k in $\mathcal{O}$ into an interval of same length since φ is constant in each subinterval of $\mathcal{O}$. Thus,

$$m(\psi(\mathcal{O})) = \sum_{k=1}^{\infty} l(\psi(I_k)) = \sum_{k=1}^{\infty} l(I_k) = m(\mathcal{O}) = 1$$

We know that $[0,2] = \psi(\mathcal{O}) \cup \psi(C)$. Since $m(\psi(\mathcal{O})) = 1$, $m(\psi(C)) = 2 - 1 = 1$.

Clearly, $\psi(C)$ is a subset of positive measure. Thus, by Vitali's theorem $\psi(C)$ has a subset W which is not measurable. However, $\psi^{-1}(W)$ is a subset of C which is of measure zero. Therefore, C has a measurable subset which ψ maps onto a non-measurable set. □

Theorem 10.2.12. *Cantor set has a measurable subset which is not a Borel set.*

Proof. By Vitali's theorem, $\psi(C)$ has a non-measurable subset, say W. Clearly W is not a Borel set, since every Borel set is measurable.

Now $\psi : \psi^{-1}(W) \to W$ is a bijection. And $W \subset \psi(C) \implies \psi^{-1}(W) \subset C$. Also we know that, every subset of zero measure sets are measurable. Thus, $\psi^{-1}(W)$ is measurable subset of Cantor set, C.

Suppose $\psi^{-1}(W)$ is a Borel set. Then, W must be a Borel set, since continuous functions maps Borel sets onto Borel sets. This leads to a contradiction since W is not a Borel set. Therefore, Cantor set C has a measurable, non-Borel subset $\psi^{-1}(W)$. □

[12] Strictly increasing functions are one-to-one. Thus, ψ is a bijection.
[13] Cantor set is a subset of zero measure. Thus every subset of Cantor set is measurable.

10.3 Lebesgue Measurable Functions

10.3.1 Sum, Products, and Compositions

Theorem 10.3.1. *Let function f have a measurable domain. Then the following statements are equivalent :*

1. $\forall c \in \mathbb{R},\ \{x \in E : f(x) > c\}$ *is measurable*
2. $\forall c \in \mathbb{R},\ \{x \in E : f(x) \geq c\}$ *is measurable*
3. $\forall c \in \mathbb{R},\ \{x \in E : f(x) < c\}$ *is measurable*
4. $\forall c \in \mathbb{R},\ \{x \in E : f(x) \leq c\}$ *is measurable*

And each statement above imply that $\{x \in E : f(x) = c\}$ is measurable.

Proof. **Step 1 :** $1 \iff 4$ **and** $2 \iff 3$
The sets considered in statements 1 and 4 are complementary. Thus, if one set is measurable, then the other set is also measurable. Similarly, the sets in 2 and 3 are complementary. Thus, one statement implies the other.

Step 2 : $1 \implies 2$
Suppose $\{x \in E : f(x) > c\}$ is measurable for any real number c. Clearly, the sets $\{x \in E : f(x) > c - \frac{1}{k}\}$ is measurable for every natural number k. And countable intersection of measurable sets is measurable. Thus,

$$\bigcap_{k=1}^{\infty} \left\{x \in E : f(x) > c - \frac{1}{k}\right\} = \{x \in E : f(x) \geq c\} \text{ is measurable}$$

Step 3 : $2 \implies 1$ Suppose $\{x \in E : f(x) \geq c\}$ is measurable for any real number c. Clearly, the sets $\{x \in E : f(x) \geq c + \frac{1}{k}\}$ is measurable for every natural number k. Again, countable union of measurable sets is measurable. Thus,

$$\bigcup_{k=1}^{\infty} \left\{x \in E : f(x) \geq c + \frac{1}{k}\right\} = \{x \in E : f(x) > c\} \text{ is measurable}$$

Step 4 : $\forall c \in \mathbb{R},\ \{x \in E : f(x) = c\}$ **is measurable**
Suppose one of the statements is true. Then other three statements are also true. The sets $\{x \in E : f(x) \geq c\}$ and $\{x \in E : f(x) \leq c\}$ are both measurable. Thus, their intersection $\{x \in E : f(x) = c\}$ is also measurable. □

Definitions 10.3.1 (measurable function)**.** A function f is measurable if

1. the domain of f is measurable and
2. one of the following statements is true
   (a) $\forall c \in \mathbb{R},\ \{x \in E : f(x) > c\}$ is measurable
   (b) $\forall c \in \mathbb{R},\ \{x \in E : f(x) \geq c\}$ is measurable
   (c) $\forall c \in \mathbb{R},\ \{x \in E : f(x) < c\}$ is measurable
   (d) $\forall c \in \mathbb{R},\ \{x \in E : f(x) \leq c\}$ is measurable

Note : If f is a measurable function, then its domain is measurable and $f^{-1}(c)$ is measurable for any real number c. (why ?)

Study of measurable functions

1. **Inverse images of open sets measurable**
   Suppose function f is defined on a measurable set E. Then f is a measurable function $\iff \forall$ open set $\mathcal{O}$, $f^{-1}(\mathcal{O})$ is measurable.

   Proof. Given the domain of f is measurable.
   Step 1 : $f^{-1}(\mathcal{O})$ is measurable $\implies$ f is measurable
   Suppose inverse image of every open set is measurable. Then, $f^{-1}(c,\infty) = \{x \in E : f(x) > c\}$ is measurable. Then, by the definition, f is a measurable function.

   Step 2 : f measurable $\implies$ $f^{-1}(\mathcal{O})$ is measurable
   Suppose function f is measurable. Let $\mathcal{O}$ be an open set. Then it can be expressed as a countable union of open, bounded intervals, I_k. That is, $\mathcal{O} = \bigcup_{k=1}^{\infty} I_k$. We know that, $\forall k \in \mathbb{N}$, $I_k = (a_k, b_k) = (-\infty, b_k) \cap (a_k, \infty)$. By the definition of measurablility $\{x \in E : f(x) < b_k\} = f^{-1}(-\infty, b_k)$ and $\{x \in E : f(x) > a_k\} = f^{-1}(a_k,\infty)$ are measurable. And $f^{-1}(a_k, b_k) = f^{-1}(-\infty, b_k) \cap f^{-1}(a_k,\infty)$ is measurable. Again, inverse image of the open set $f^{-1}(\mathcal{O})$ which is a countable union of measurable sets $\{f^{-1}(I_k)\}_{k=1}^{\infty}$ is also measurable. Therefore, inverse image of every open set is measurable. □

2. **Continuous, real-valued functions are measurable**
   Suppose the function f has a measurable domain. And f is a real-valued, continuous function. Then f is measurable.

   Proof. Let E be a measurable set. And f be a continuous function on E. Let $\mathcal{O}$ be an open set, then by open set characterisation of continuous function we have $f^{-1}(\mathcal{O}) = E \cap U$ where U is an open set. Clearly, $f^{-1}(\mathcal{O})$ is a union of two measurable sets and thus measurable. Therefore, f is measurable since $\forall$ open set $\mathcal{O}$, $f^{-1}(\mathcal{O})$ is measurable. □

3. **Monotone function on an interval is measurable**

   Proof. Without loss of generality, suppose that f is an increasing function defined on an interval I. Since, every interval is measurable, f is defined on a measurable set.

   Consider $\{g_n\}_{n=1}^{\infty}$ defined by $g_n(x) = f(x) + \frac{x}{n}$. Since g_n are strictly increasing functions, $g_n^{-1}(c,\infty) = I \cap U$ where U is an interval. Thus, $g_n^{-1}(c,\infty) = \{x \in I : g_n(x) > c\}$ is always measurable. Therefore, $\{g_n\}$ is a family of measurable functions.

   Now, from the construction of g_n, we have $\lim_{n\to\infty} g_n = f$. We have, $\{g_n\}$ is a sequence of measurable functions converging pointwise to the limit function f (a.e.) on the interval I. Therefore, f is measurable.†[14] □

[14]Limit of measurable functions under pointwise convergence(a.e.) is measurable.
We will prove this result in the upcoming subsection .

4. **f measurable and $f = g$(a.e) $\implies$ g measurable**

   Proof. Suppose f, g are functions defined on a measurable set E. Suppose f is measurable. Let $A \subset E$ such that $A = \{x \in \mathbb{R} : f(x) \neq g(x)\}$. We have $f = g$ (a.e.). Thus, $m(A) = 0$. And we have,

   $$\begin{aligned}\{x \in E : g(x) > c\} =&\{x \in A : g(x) > c\} \cup \{x \in E \sim A : g(x) > c\}\\ =&\{x \in A : g(x) > c\} \cup \{x \in E \sim A : f(x) > c\}\\ =&\{x \in A : g(x) > c\} \cup [\{x \in E : f(x) > c\} \cap (E \sim A)]\end{aligned}$$

   Clearly, $\{x \in A : g(x) > c\}$ is a subset of a set of measurse zero and thus measurable. Also, $E \sim A$ measurable since both E and A are measurable. And since f is measurable, $\{x \in E : f(x) > c\}$ is measurable. Thus, $\{x \in E : g(x) > c\}$ is measurable for any $c \in \mathbb{R}$. Therefore, g is a measurable function. □

5. **f measurable $\iff$ $f|_D, D \sim E$ measurable ($\forall D$ measurable)**
   Suppose function f is an extended real-valued function on a measurable set E. For measurable subset D of E, f is measurable on E if and only if the resctriction of f to D and $D \sim E$ are measurable.

   Proof. **Part 1 : f measurable $\implies$ $f|_D$ measurable**
   Suppose f is a measurable function defined on E and D is a measurable subset of E. Since, f is measurable, E is measurable. Clearly $E \sim D$ is measurable.

   $$\begin{aligned}\{x \in D : f|_D(x) > c\} =&\{x \in D : f(x) > c\}\\ =&\{x \in E : f(x) > c\} \cap D\end{aligned}$$

   Since f is measurable, $\{x \in E : f(x) > c\}$ is measurable for any $c \in \mathbb{R}$. And intersection of measurable sets are measurable. Thus, $\{x \in D : f|_D(x) > c\}$ is measurable for any $c \in \mathbb{R}$. Therefore, $f|_D$ is a measurable function. □

Definitions 10.3.2 (characteristic function)**.** Let A be a subset of $\mathbb{R}$. The characteristic function χ_A of A is given by

$$\chi_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{10.29}$$

Definitions 10.3.3. Let $\{f_1, f_2, \ldots, f_n\}$ be a finite family of measurable functions on the same domain E. Then $\max\{f_1, f_2, \ldots, f_n\}$ on E is given by

$$\max\{f_1, f_2, \ldots, f_n\}(x) = \max\{f_1(x), f_2(x), \ldots, f_n(x)\},\ \forall x \in E \tag{10.30}$$

And function $\min\{f_1, f_2, \ldots, f_n\}$ on E is given by

$$\min\{f_1, f_2, \ldots, f_n\}(x) = \min\{f_1(x), f_2(x), \ldots, f_n(x)\},\ \forall x \in E \tag{10.31}$$

Properties of Measurable Functions

Suppose f, g are measurable functions on E and are finite a.e. on E.

1. **Linearity :** $\alpha f + \beta g$ is measurable $\forall \alpha, \beta$

   Proof. **Step 1 : f measurable $\implies \alpha f$ measurable**
   Suppose $\alpha = 0$. Then $\alpha f = 0$. And 0 function is trivially measurable.
   Suppose $\alpha \neq 0$. Then $\{x \in E : f(x) > c/\alpha\} = \{x \in E : \alpha f(x) > c\}$ is measurable for any $c \in \mathbb{R}$. Therefore αf is measurable for any $\alpha \in \mathbb{R}$.

   In other words, if f is measurable, then αf is measurable and if g is measurable, then βg is measurable for any $\alpha, \beta \in \mathbb{R}$. Therefore, it is enough to prove that f, g measurable $\implies f + g$ measurable.
   Step 2 : f, g are measurable $\implies f + g$ is measurable

   $$f(x) + g(x) < c \iff f(x) < c - g(x)$$

   Since $\mathbb{Q}$ is dense, there exists a rational number $q \in \mathbb{Q}$ between any two distinct real numbers

   $$\begin{aligned} f(x) + g(x) < c &\iff \exists q \in \mathbb{Q},\ f(x) < q < c - g(x) \\ &\iff f(x) < q \text{ AND } g(x) < c - q \end{aligned}$$

   $$\begin{aligned} \{x \in E : f + g(x) < c\} =& \{x \in E : f(x) + g(x) < c\} \\ &= \bigcup_{q \in \mathbb{Q}} [\{x \in E : f(x) < c\} \cap \{x \in E : g(x) < c - q\}] \end{aligned}$$

   Since, f, g are measurable, each set in the union is measurable. Thus, $\{x \in E : f + g(x) < c\}$ is measurable for every $c \in \mathbb{R}$, since rational numbers are countable, and countable union of measurable sets is measurable. Therefore, $f + g$ is measurable. □

2. **Product :** fg is measurable

   Proof. Suppose f, g are measurable functions which are finite (a.e.) on a measurable set E. And we have,

   $$fg = \frac{1}{2}\left[(f+g)^2 - f^2 - g^2\right] \tag{10.32}$$

   Step 1 : f is measurable $\implies f^2$ is measurable
   Suppose $c \geq 0$. If $c < 0$, then $\{x \in E : f^2(x) > c\} = E$ is measurable.

   $$\{x \in E : f^2(x) > c\} = \{x \in E : f(x) > \sqrt{c}\} \cup \{x \in E : f(x) < -\sqrt{c}\}$$

   is a union of two measurable sets and is measurable.

   Step 2 : fg is measurable
   We have f, g are measurable. Thus, $f + g$ is measurable, since linear combination of measurable sets is measurable. And, $f^2, g^2, (f + g)^2$ are measurable by Step 1. We know that fg is a linear combination of these measurable sets and therefore fg is measurable. □

3. **Composition of measurable functions is not necessarily measurable**
We know that, Cantor set C has subset $A = \psi^{-1}(W)$ such that ψ maps A onto a non-measurable set W.†[15] Then A is a measurable subset of open interval $(0, 1)$ and $\psi(A)$ is non-measurable.

Let χ_A be the characteristic function of A. Then $\chi_A \circ \psi^{-1}$ is not measurable since $\{x \in (0, 1) : \chi_A \circ \psi^{-1}(x) \geq 1\} = \psi(A)$ is not measurable. But, both the functions χ_A and ψ^{-1} are measurable functions since $\chi_A^{-1}(y)$ is either ϕ, A or A^c and ψ^{-1} is a continuous real-valued function defined on $(0, 1)$.

4. **If f is continuous and g is measurable, then $f \circ g$ is measurable**

Proof. Suppose f be a continuous function and g be a measurable function both defined on a common set E which is measurable. Since f is continuous, we know that f is measurable.

Let $\mathcal{O}$ be an open set. Then $(f \circ g)^{-1}(\mathcal{O}) = g^{-1}(f^{-1}(\mathcal{O}))$. By characterisation of continuity, $U = f^{-1}(\mathcal{O})$ is an open set since $\mathcal{O}$ is open. And by characterisation of measurability of g, we have $g^{-1}(U)$ is an open set since g is measurable and U is open. Now we have $(f \circ g)^{-1}(\mathcal{O})$ is an open set for any open set $\mathcal{O}$. Therefore by the characterisation of measurability, $f \circ g$ is measurable. □

5.
For a finite family of measurable functions $\{f_1, f_2, \ldots, f_n\}$ with common domain E, both the functions $\max\{f_1, f_2, \ldots, f_n\}$ and $\min\{f_1, f_2, \ldots, f_n\}$ are measurable.

Proof. Let $\{f_1, f_2, \ldots, f_n\}$ be a finite family of measurable functions defined on a common measurable set E. We have,

$$\{x \in E : max\{f_1, f_2, \ldots, f_n\}(x) > c\} = \bigcup_{k=1}^{n} \{x \in E : f_k(x) > c\}$$

is a finite union of measurable sets since each f_k is measurable. Therefore, $max\{f_1, f_2, \ldots, f_n\}$ is measurable. Similary, we have,

$$\{x \in E : min\{f_1, f_2, \ldots, f_n\}(x) < c\} = \bigcup_{k=1}^{n} \{x \in E : f_k(x) < c\}$$

Therefore, $min\{f_1, f_2, \ldots, f_n\}$ is also measurable.

□

[15] From the proof of Vitali's theorem and associated lemma, $\psi(C)$ has such a subset W which is a choice set of $\psi(C)$ under rational equaivalence.

Note : Let f be a measurable function on a measurable set E.
Then $-f$ is measurable since

$$\{x \in E : -f(x) > c\} = \{x \in E : f(x) < -c\}$$

And $|f|, f^+, f^-$ are also measurable since

$$|f| = max\{f, -f\},\ f^+ = max\{f, 0\},\ f^- = max\{-f, 0\}$$

These function f^+, f^- are quite important as we may write $f = f^+ - f^-$ where both the functions f^+ and f^- are non-negative. And Lebesgue integral for non-negative functions are sufficient to integrate any function f.

$$\int_E f = \int_E f^+ - \int_E f^-$$

10.3.2 Sequential Pointwise Limits and Simple Approximation

Definitions 10.3.4 (pointwise convergence)**.** A sequence of functions $\{f_n\}$ on a common domain E convergence pointwise to f if

$$\forall x \in E,\ \lim_{n\to\infty} f_n(x) = f(x)$$

Definitions 10.3.5 (pointwise convergence(a.e.))**.** Let E_0 be a set of zero measure. A sequence of functions $\{f_n\}$ on a common domain E convergence pointwise a.e. on $E_0 \subset E$ to f if

$$\forall x \in E \sim E_0,\ \lim_{n\to\infty} f_n(x) = f(x)$$

Definitions 10.3.6 (uniform convergence)**.** A sequence of functions $\{f_n\}$ on a common domain E convergence uniformly to f if

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N},\ \text{such that } \forall n \geq N,\ |f_n - f| < \varepsilon$$

Theorem 10.3.2 (Pointwise Convergence(a.e.) preserves measurability)**.** *Let $\{f_n\}$ be a sequence measurable functions on E. If $\{f_n\}$ converges to f pointwise a.e. on E, then f is measurable.*

Proof. Let $\{f_n\}$ be a sequence of measurable functions converging to f pointwise a.e. on E. Then $\{f_n\}_{n=1}^{\infty}$ converges pointwise on $E \sim E_0$ where E_0 is a set of measure zero. Without loss of generality, suppose that $\{f_n\}$ converges pointwise everywhere on E.†[16]

Let $c \in \mathbb{R}$. We claim that,

$$f(x) < c \iff \exists n, k \text{ such that } f_j(x) < c - \frac{1}{n},\ \forall j \geq k \tag{10.33}$$

Suppose that for every natural numbers n, k, there exists $j \geq k$ for which $f_j(x) \geq c - \frac{1}{n}$. Since $f(x)$ is the limit function, $f(x) \not< c$ which is a contradiction. Thus, we may write,

$$\{x \in E : f(x) < c\} = \bigcup_{n,k=1}^{\infty} \left[\bigcap_{j=k}^{\infty} \left\{ x \in E : f_j(x) < c - \frac{1}{n} \right\} \right] \tag{10.34}$$

[16] Consider $E' = E \sim E_0$

Countable union of countable intersections of measurable sets is also measurable. Thus $\{x \in E : f(x) < c\}$ is a measurable set for any $c \in \mathbb{R}$. Therefore, f is measurable. $\square$

Definitions 10.3.7 (simple function)**.** A real-valued function φ on a measurable set E is simple if it is measurable and assumes at most finitely many values.

Note : $xRy \iff \varphi(x) = \varphi(y)$ is an equivalence relation which partitions E into k disjoint subsets E.

Note : Suppose φ is a simple function. Then each equivalent class of E under above equivalence is also measurable by the definition of measurable function.(Why ?)

Definitions 10.3.8 (canonical representation)**.** The canonical representation of a simple function φ on E is given by,

$$\varphi = \sum_{k=1}^{n} c_k \cdot \chi_{E_k} \text{ where } E_k = \{x \in E : \varphi(x) = c_k\} \tag{10.35}$$

Definitions 10.3.9 (bounded)**.** Let f be a function on E. The function f is bounded if there exists $m \in \mathbb{R},\ m \geq 0$ such that $|f(x)| < m$ for every $x \in E$.

Lemma 10.3.3 (simple approximation)**.** *Let f be a measurable, real-valued function which is bounded on E. Then for any $\varepsilon > 0$, there exist simple functions φ_ε and ψ_ε approximating f (from below and above) such that $0 \leq \psi_\varepsilon - \varphi_\varepsilon < \varepsilon$ on E.*

Proof. Let f be a bounded, measurable function on E. Then there exists open, bounded interval (c, d) such that $f(E) \subset (c, d)$. Let $P = \{y_0, y_1, \ldots, y_n\}$ be a partitions of open interval (c, d) such that $y_0 = c < y_1 < \cdots < y_{n-1} < y_n = d$ and $y_k - y_{k-1} < \varepsilon$ for every k. Define $I_k = [y_{k-1}, y_k)$ and $E_k = f^{-1}(I_k)$. Then E_k are measurable since I_k are intervals and f is measurable.(Why ?)

Define $\varphi_\varepsilon = \sum_{k=1}^{n} y_{k-1} \cdot \chi_{E_k}$ and $\psi_\varepsilon = \sum_{k=1}^{n} y_k \cdot \chi_{E_k}$. Then for any $x \in E$, we have $\varphi_\varepsilon(x) = y_{k-1} \leq f(x) \leq y_k = \psi_\varepsilon(x)$ Clearly, $\varphi_\varepsilon \leq \psi_\varepsilon$ by the construction. And for each k, we have $\psi_\varepsilon(y) - \varphi_\varepsilon(y) = y_k - y_{k-1} < \varepsilon$ for any $y \in [y_{k-1}, y_k)$. Therefore, $0 \leq \psi_\varepsilon - \varphi_\varepsilon < \varepsilon$. $\square$

Theorem 10.3.4 (simple approximation)**.** *An extended real-valued function f on a measurable set E is measurable if and only if there exists a sequence of simple functions $\{\varphi_n\}$ converging to f pointwise on E and $|\varphi_n| \leq |f|$ on E for every $n \in \mathbb{N}$.*
And if f is non-negative, there exists an increasing sequence of functions $\{\varphi_n\}$ with the same property.

Proof. **Part 1**
Suppose there exists a sequence of simple functions $\{\varphi_n\}$ converging pointwise to f on E. Since simple functions are measurable and pointwise limit of measurable functions are measurable, f is measurable.

Part 2

Suppose f is measurable. Without loss of generality, suppose that $f \geq 0$ on E. Otherwise, $f = f^+ - f^-$ where $f^+, f^- \geq 0$. And f is measurable since both f^+ and f^- are measurable.

Step 1 : Construction of simple functions on E_n

Let $n \in \mathbb{N}$. Define $E_n = \{x \in E : f(x) \leq n\}$. Since f is measurable, E_n is a measurable subset of E. And we know that $f|_{E_n}$ is also a non-negative, bounded, measurable function.

Then by simple approximation lemma, for $\varepsilon = \frac{1}{n}$ there exists φ_n and ψ_n such that $0 \leq \varphi_n \leq f \leq \psi_n$ and $0 \leq \psi_n - \varphi_n < \frac{1}{n}$ on E_n.

Step 2 : Extending simple function to E

Extend φ_n to E by setting $\varphi_n(x) = n$ if $f(x) > n$. Again, φ_n is a simple function on E. And $0 \leq \varphi_n \leq f$ on E.

Step 3 : Sequence of simple functions converging to f

We claim that the sequence of simple functions $\{\varphi_n\}$ converges to f pointwise on E. Suppose $x \in E$. Suppose $f(x)$ is finite. Let $m \in \mathbb{N}$. Then $0 \leq f(x) - \varphi_n(x) \leq \frac{1}{n}$ for all $n \geq m$. Therefore, $\lim\limits_{n\to\infty} \varphi_n(x) = f(x)$. Suppose $f(x) = \infty$. Then $\varphi_n(x) = n$, $\forall n$. Therefore, $\lim\limits_{n\to\infty} \varphi_n(x) = f(x)$.

Replace each φ_n with $max\{\varphi_1, \varphi_2, \ldots, \varphi_n\}$†[17], we have an increasing sequence of simple functions $\varphi_n\}$ converging to f pointwise on E. □

Exercise

12.

13.

14.

15.

16.

17.

18.

19.

20.

21.

22.

23.

[17] We replace each simple function φ_k with the maximum of the finite subsequence from φ_1 upto φ_k. That is, $\varphi_1' = max\{\varphi_1\} = \varphi_1$, and $\varphi_2' = max\{\varphi_1, \varphi_2\}$, and $\varphi_3' = max\{\varphi_1, \varphi_2, \varphi_3\}$, ...

24.

10.3.3 Littlewood, Egoroff, Lusin

Out of syllabus - But important for our understanding of the subject.

Littlewoods' three principles

1. Every measurable set is **nearly** a finite union of intervals.
2. Every measurable function is **nearly** continuous.
3. Every convergent sequence of measurable functions is **nearly** uniform.

Theorem 10.3.5 (Egoroff)**.** *Suppose a sequence $\{f_n\}$ of measurable functions converges to f pointwise on a set E of finite measure. Then for any $\varepsilon > 0$, there exists a closed set $F \subset E$ such that $m(E \sim F) < \varepsilon$ and $\{f_n\}$ converges to f uniformly on F.*

Theorem 10.3.6. *Let f be a simple function on E. Then for any $\varepsilon > 0$, there exists a continuous function g on $\mathbb{R}$ and a closed set $F \subset E$ such that $f = g$ on F and $m(E \sim F) < \varepsilon$.*

Theorem 10.3.7 (Lusin)**.** *Let f be a real-valued measurable function on E. Then for any $\varepsilon > 0$, there exists a continuous function g on $\mathbb{R}$ and closed set $F \subset E$ such that $f = g$ on F and $m(E \sim F) < \varepsilon$.*

10.4 Lebesgue Integration

10.4.1 Riemann integral

Definitions 10.4.1 (partition)**.** A set $P = \{a, x_1, x_2, \dots, x_{n-1}, b\}$ is a partition of an interval $[a, b]$ into n subintervals if $a < x_1 < x_2 < \cdots < x_{n-1} < b$.

Definitions 10.4.2 (lower/upper Riemann integral)**.** Let f be a bounded, real-valued function on a closed interval $[a, b]$ and P be a partition of $[a, b]$. Let M_i, m_i be the supremum and infimum of f in each subinterval of the partition. Then upper and lower Riemann integrals are the infimum of upper Darboux sums and supremum of lower Darboux sums.

Riemann upper integral,

$$(R)\overline{\int_a^b} f = \inf_P \{U(f, P)\} = \inf \left\{ \sum_{i=1}^{n} M_i (x_i - x_{i-1}) \right\} \tag{10.36}$$

Riemann lower integral,

$$(R)\underline{\int_a^b} f = \sup_P \{L(f, P)\} = \sup \left\{ \sum_{i=1}^{n} m_i (x_i - x_{i-1}) \right\} \tag{10.37}$$

Definitions 10.4.3 (Riemann integrable)**.** A function f is Riemann integrable over $[a, b]$ if lower and upper Riemann integrals are equal. And Riemann integral of f over $[a, b]$ is given by,

$$(R)\int_a^b f = (R)\overline{\int_a^b} f = (R)\underline{\int_a^b} f \tag{10.38}$$

Definitions 10.4.4 (step function)**.** A function ψ is a step function if it assumes constant values in each subinterval of some partition of its domain.

$$\psi(x) = \sum_{i=1}^{n} c_i \cdot \chi_{(x_i, x_{i-1})}$$

$$\text{Note : } (R)\int_a^b \psi = \sum_{i=1}^{n} c_i(x_i - x_{i-1})$$

Definitions 10.4.5 (Dirchlet's function)**.** Dirchlet's function is given by,

$$f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$$

Note : Dirchlet's function is not Riemann integrable(why ?). But it is Lebesgue integrable since $\mathbb{Q}$ is Lebesgue measurable.

10.4.2 Lebesgue integral of a bounded, measurable function over a set of finite measure

Definitions 10.4.6. Let ψ be a simple function on a finite measure set E. Then,

$$\int_E \psi = \sum_{i=1}^{n} a_i \cdot m(E_i) \text{ where } \psi = \sum_{i=1}^{n} a_i \cdot \chi_{E_i} \tag{10.39}$$

Note : Step functions are simple functions. And the Riemann integral and Lebesgue integral of step functions are the same.

Note : Dirchlet's function is a simple function. And thus Lebesgue integrable.

Lemma 10.4.1. *Let $\{E_k\}_{k=1}^n$ be a disjoint collection of measurable subsets of a set E of finite measure. If $\varphi = \sum_{k=1}^n a_k \cdot \chi_{E_k}$, then $\int_E \varphi = \sum_{k=1}^n a_k \cdot m(E_k)$*

Proof. Suppose $\varphi = \sum_{k=1}^n a_k \cdot \chi_{E_k}$ where E_k are of finite measure. We consider distinct values of a_i, $\{\lambda_1, \lambda_2, \dots, \lambda_m\}$. Then, the function has the canonical representation, $\varphi = \sum_{r=1}^{m} \lambda_r \cdot \chi_{A_r}$ where $A_r = \bigcup_{\substack{k \\ a_k = \lambda_r}} E_k$.

Then, we have $\varphi^{-1}(\lambda_k) = A_k$ is a union of finite measurable set and thus measurable. Thus, φ is a measurable function which assumes at most m different values. And

$$\begin{aligned}\int_E \varphi &= \sum_{r=1}^{m} \lambda_r \cdot m(A_r)\\ &= \sum_{r=1}^{m} \lambda_r \cdot m\left(\bigcup_{\substack{k\\ a_k=\lambda_r}} E_k\right)\\ &= \sum_{r=1}^{m} \sum_{\substack{k\\ a_k=\lambda_r}} a_k \cdot m(E_k)\\ &= \sum_{k=1}^{n} a_k \cdot m(E_k)\end{aligned}$$

□

Theorem 10.4.2 (linearity + monotonicity of integral of simple functions)**.** *Let φ and ψ be simple functions defined a set E of finite measure. Then for any $\alpha, \beta \in \mathbb{R}$,*

$$\int_E (\alpha\varphi + \beta\psi) = \alpha \int_E \varphi + \beta \int_E \psi \tag{10.40}$$

Moreover, if $\varphi \leq \psi$ on E, then

$$\int_E \varphi \leq \int_E \psi \tag{10.41}$$

Proof. Suppose φ, ψ are simple functions and $\alpha, \beta \in \mathbb{R}$. Since φ, ψ are simple, there are n disjoint subset of E in which both φ and ψ are constant†[18]. Then,

$$\begin{aligned}\int_E (\alpha\varphi + \beta\psi) &= \int_E \alpha\left(\sum_{i=1}^{n} a_i \cdot \chi_{E_i}\right) + \beta\left(\sum_{i=1}^{n} b_i \cdot \chi_{E_i}\right)\\ &= \int_E (\alpha a_i + \beta b_i) \cdot \chi_{E_i}\\ &= \sum_{i=1}^{n} (\alpha a_i + \beta b_i) \cdot m(E_i)\\ &= \alpha \sum_{i=1}^{n} a_i \cdot m(E_i) + \beta \sum_{i=1}^{n} b_i \cdot m(E_i)\\ &= \alpha \int_E \varphi + \beta \int_E \psi\end{aligned}$$

Suppose $\varphi \leq \psi$. Then $\eta = \psi - \varphi$ is also a simple function. And $0 \leq \eta$. Then by linearity of Lebesgue integral of simple functions,

$$\begin{aligned}\int_E \eta &= \int_E \psi - \varphi\\ &= \int_E \psi - \int_E \varphi\end{aligned}$$

We have, η is a non-negative, measurable function. And thus, $\int_E \eta \geq 0$. Therefore, $\int_E \varphi \leq \int_E \psi$. □

[18]We don't assume that the values assume by ψ are distinct for distinct subsets.

Lebesgue integral for bounded functions

Now we define Lebesgue integral of bounded functions, the same way we have constructed Riemann integral. We use simple approximations and the definition of Lebesgue integral for simple functions for this construction.

Definitions 10.4.7 (upper/lower Lebesgue integral)**.** Let f be a bounded, real-valued function on a set E of finite measure. Any simple approximation lemma, we have two families of simple functions $\{\varphi_\varepsilon\}$ and $\{\psi_\varepsilon\}$ such that $\forall \varepsilon > 0,\ \varphi \leq f \leq \psi$ and $0 \leq \varphi - \psi < \varepsilon$.Then
Upper Lebesgue integral,

$$\overline{\int_E} f = \inf_{\psi_\varepsilon} \left\{ \int_E \psi_\varepsilon \right\} \tag{10.42}$$

Lower Lebesgue integral,

$$\underline{\int_E} f = \sup_{\varphi_\varepsilon} \left\{ \int_E \varphi_\varepsilon \right\} \tag{10.43}$$

Definitions 10.4.8 (Lebesgue integrability of bounded functions)**.** Let f be a bouned, real-valued function defined on a set E of finite measure. Function f is Lebesgue integrable on E if both lower and upper Lebesgue integrals of f over E are equal.

And the Lebesgue integral of a bounded, real-valued function f over E is given by,

$$\int_E f = \sup_{\varphi} \int_E \varphi = \inf_{\psi} \int_E \psi \tag{10.44}$$

Now we can prove that Lebesgue integral is a generalisation of Riemann integral for bounded functions.

Theorem 10.4.3. *Let f be a bounded, real-valued function defined on a close, bounded interval $[a, b]$. If f is Riemann integrable over $[a, b]$, then it is Lebesgue integrable over $[a, b]$. And both Riemann integral and Lebesgue integral of f over $[a, b]$ are equal.*

Proof. Suppose f is Riemann integrable over $[a, b]$. Then,

$$\sup \left\{ (R) \int_a^b \varphi : \varphi \text{ step },\varphi \leq f \right\} = \inf \int \left\{ \int_a^b \psi : \psi \text{ step}, f \leq \psi \right\} \tag{10.45}$$

However, every step function is simple and we have

$$\sup \left\{ \int_a^b \varphi : \varphi \text{ simple },\varphi \leq f \right\} = \inf \int \left\{ \int_a^b \psi : \psi \text{ simple}, f \leq \psi \right\} \tag{10.46}$$

Therefore, every bounded, real-valued function on closed,bounded interval is Lebesgue integrable function if it is Riemann integrable. $\square$

Theorem 10.4.4. *Every bounded, measurable function f defined on a set E of finite measure is Lebesgue integrable over E.*

Proof. Let E be a set of finite measure. And f be a bounded, measurable function on E. Let $n \in \mathbb{N}$. By simple approximation lemma, for $\varepsilon = \frac{1}{n}$ we have simple functions φ_n, ψ_n such that $\varphi_n \leq f \leq \psi_n$ and $\psi_n - \varphi_n < \frac{1}{n}$ on E.

$$0 \leq \psi_n - \varphi_n \leq \frac{1}{n} \implies 0 \leq \int_E \psi_n - \varphi_n = \int_E \psi_n - \int_E \varphi_n \leq \frac{1}{n} \int_E 1$$

We know that, $\inf\{r_1, r_2, \dots\} \leq r_k$ and $-\sup\{s_1, s_2, \dots\} \leq -s_k$.

$$\begin{aligned}
0 &\leq \inf\left\{\int_E \psi : \psi \text{ simple}, f \leq \psi\right\} - \sup\left\{\int_E \varphi : \varphi \text{ simple}, \varphi \leq f\right\} \\
&\leq \int_E \psi_n - \int_E \varphi_n \leq \frac{1}{n} m(E)
\end{aligned}$$

This inequality is true for any $n \in \mathbb{N}$. Since $m(E)$ is finite, $\frac{1}{n}m(E) \to 0$ as $n \to \infty$. Thus, upper and lower Lebesgue integrals of f are equal. Therefore, f is Lebesgue integrable. □

Theorem 10.4.5 (linearity + monotonicity of integral of bounded functions)**.** *Let f and g be bounded, measurable functions defined a set E of finite measure. Then for any $\alpha, \beta \in \mathbb{R}$,*

$$\int_E (\alpha f + \beta g) = \alpha \int_E f + \beta \int_E g \tag{10.47}$$

Moreover, if $f \leq g$ on E, then

$$\int_E f \leq \int_E g \tag{10.48}$$

Proof. Linear combination of bounded,measurable functions is measurable and bounded. Thus, if f, g are bounded, measurable functions, then $\alpha f + \beta g$ is also a bounded, measurable function. And $\alpha f + \beta g$ is integrable as every bounded, measurable function on E is integrable over E.

Step 1 : $\displaystyle\int_E \alpha f = \alpha \int_E f$

$$\begin{aligned}
\int_E \alpha f = &\inf_{\psi \geq \alpha f} \left\{\int_E \psi : \psi \text{ simple}\right\} \\
=&\alpha \inf_{\psi/\alpha \geq f} \left\{\int_E \psi/\alpha : \psi/\alpha \text{ simple}\right\} \\
=&\alpha \inf_{\psi' \geq f} \left\{\int_E \psi' : \psi' \text{ simple}\right\} \\
=&\alpha \int_E f
\end{aligned}$$

Step 2 : $\displaystyle\int_E f + g = \int_E f + \int_E g$

Since $f \leq \psi_f$ and $g \leq \psi_g$, we have $f + g \leq \psi_f + \psi_g$ and $\psi_f + \psi_g$ is a simple function. And

$$\int_E f + g = \inf \left\{ \int_E \psi : \psi \text{ simple, } f + g \leq \psi \right\} \leq \int_E \psi_f + \psi_g$$

Since integral of simple functions have linearity

$$\int_E f + g \leq \int_E \psi_f + \psi_g = \int_E \psi_f + \int_E \psi_g$$

This inequality is true for simple functions dominating f, g. Thus, it is true for the infimum of a family of such simple functions.

$$\int_E f + g \leq \inf \left\{ \int_E \psi_f : \psi_f \text{ simple, } f \leq \psi_f \right\} + \inf \left\{ \int_E \psi_g : \psi_g \text{ simple, } g \leq \psi_g \right\}$$

Therefore,

$$\int_E f + g \leq \int_E f + \int_E g \tag{10.49}$$

Similarly,

$$\begin{aligned} \int_E f + g &= \sup \left\{ \int_E \varphi : \varphi \text{ simple, } \varphi \leq f + g \right\} \\ &\geq \int_E \varphi_f + \varphi_g = \int_E \varphi_f + \int_E \varphi_g \\ &\geq \sup \left\{ \int_E \varphi_f : \varphi_f \text{ simple, } \varphi_f \leq f \right\} + \sup \left\{ \int_E \varphi_g : \varphi_g \text{ simple, } \varphi_g \leq g \right\} \end{aligned}$$

Therefore,

$$\int_E f + g \geq \int_E f + \int_E g \tag{10.50}$$

Thus, for any bounded, measurable function f and g, $\int_E f + g = \int_E f + \int_E g$.
Therefore, $\int_E \alpha f + \beta g = \int_E \alpha f + \int_E \beta g = \alpha \int_E f + \beta \int_E g$. □

Theorem 10.4.6 (additivity over the domain of integration)**.** *Let f be a bounded, measurable function on a set E of finite measure. Suppose A, B are disjoint, measurable subsets of E. Then*

$$\int_{A \cup B} f = \int_A f + \int_B f \tag{10.51}$$

Proof. We have $\chi_{A \cup B} = \chi_A + \chi_B$. And,

$$f \cdot \chi_{A \cup B} = f \cdot (\chi_A + \chi_B) = f \cdot \chi_A + f \cdot \chi_B \tag{10.52}$$

And if $E_1 \subset E$, then by the definition of Lebesgue integral

$$\int_E f \cdot \chi_{E_1} = \inf \left\{ \int_E \psi : \psi \text{ simple, } f . \chi_{E_1} \leq \psi \right\}$$

The collection of simple function ψ such that $f \cdot \chi_{E_1} \leq \psi$ on E is same as the collection of simple functions ψ such that $f \leq \psi$ on E_1.

$$= \inf\left\{\int_{E_1} \psi : \psi \text{ simple, } , f \leq \psi\right\} = \int_{E_1} f$$

Since integral has linearity, we have

$$\int_{A\cup B} f = \int_E f\cdot\chi_{A\cup B} = \int_E (f\cdot\chi_A + f\cdot\chi_B) = \int_E f\cdot\chi_A + \int_E f\cdot\chi_B = \int_A f + \int_B f$$

□

Corollary 10.4.6.1. *Let f be a bounded, measurable function on a set of finite measure E. Then,*

$$\left|\int_E f\right| \leq \int_E |f| \tag{10.53}$$

Proof. We have, $-|f| \leq f \leq |f|$ on E. Therefore by linearlity and monotonicity,

$$-\int_E |f| = \int_E -|f| \leq \int_E f \leq \int_E |f|$$

Therefore,

$$\left|\int_E f\right| \leq \int_E |f|$$

□

Theorem 10.4.7 (passage of limit under the integral sign)**.** *Let $\{f_n\}$ be a sequence of bounded, measurable functions on a set E of finite measure. If $\{f_n\}$ converges to f uniformly on E, then the sequence of the integrals $\left\{\int_E f_n\right\}$ converges to the integral of the limit function $\int_E f$.*

$$\lim_{n\to\infty} \int_E f_n = \int_E f \tag{10.54}$$

Proof. Suppose a sequence $\{f_n\}$ of bounded, measurable functions converges uniformly to f on a set of finite measure E. Uniform limit of bounded functions is bounded. And limit function of measurable functions is measurable. Therefore, f is also a bounded, measurable function.

Let $\varepsilon > 0$. Choose an N such that $|f - f_n| \leq \varepsilon/m(E)$, $\forall n \geq N$.(such an N exists since the convergence is uniform)
By linearlity and monotonicity,

$$\left|\int_E f - \int_E f_n\right| = \left|\int_E f - f_n\right| \leq \int_E |f - f_n| \leq \frac{\varepsilon}{m(E)} \cdot m(E) = \varepsilon$$

Clearly, $\left\{\int_E f_n\right\}$ is sequence of functions converging pointwise to $\int_E f$. Therefore we have, $\lim_{n\to\infty} \int_E f_n = \int_E f$. □

Definitions 10.4.9 (passage of limit under integral sign)**.** Suppose $\{f_n\}$ is a sequence of bounded, measurable functions converging to f pointwise (a.e.) on E. If

$$\lim_{n\to\infty}\int_E f_n = \int_E \lim_{n\to\infty} f_n = \int_E f \tag{10.55}$$

then this equality is the passage†[19] of the limit under the integral sign.

Note : Pointwise convergence is not sufficient for passage of limit under integral sign. In other words, there exists a sequence $\{f_n\}$ of bounded, measurable functions such that

$$\lim_{n\to\infty}\int_0^1 f_n \neq \int_0^1 f$$

where $\{f_n\}$ converges to limit function f pointwise on $[0,1]$.

Proof. Consider a family of functions defined by $f_n(\frac{1}{n}) = n$ and f_n increases from $f_n(0) = 0$ to $f_n(\frac{1}{n}) = n$ linearly and decreases from $f_n(\frac{1}{n}) = n$ to $f_n(\frac{2}{n}) = 0$ linearly. And vanishes for any $x \geq \frac{2}{n}$ on $[0,1]$.

$$f_n(x) = \begin{cases} 0 & x \in [\frac{2}{n}, 1] \\ nx(2-n) & x = [\frac{1}{n}, \frac{2}{n}) \\ n^2 x & x \in [0, \frac{1}{n}) \end{cases}$$

We know that these functions are bounded by their construction as $0 \leq f_n \leq \frac{1}{n}$ for every $n \in \mathbb{N}$. Each f_n is continuous, and thus measurable. Clearly, $\{f_n\}$ is a sequence of bounded, measurable functions defined on $[0,1]$.

We have, $\lim\limits_{n\to\infty} f_n = 0$. And, $\lim\limits_{n\to\infty}\int_0^1 f_n = 1 \neq 0 = \int_0^1 0.$ □

Note : However, the bounded convergence theorem gives us an additional constraint (uniformly pointwise bounded) for the passage of limit under integral sign.

Theorem 10.4.8 (bounded convergence)**.** *Let $\{f_n\}$ be a sequence of measurable, functions on a set of finite measure E. Suppose $\{f_n\}$ is uniformly pointwise bounded on E. If $\{f_n\}$ converges to f pointwise on E, then* $\lim\limits_{n\to\infty}\int_E f_n = \int_E f$.

Proof. Suppose sequence $\{f_n\}$ of bounded, measurable functions converges to f pointwise on E. Then the limit function f is also measurable. And since $\{f_n\}$ is uniformly pointwise bounded on E, f is bounded on E.

Let A be a measurable subset of E and $n \in \mathbb{N}$. By linearlity and monotonicity,

$$\int_E f_n - \int_E f = \int_E f_n - f = \int_A f_n - f + \int_{E\sim A} f_n + \int_{E\sim A} -f$$

By monotonicity of integral,

$$\left|\int_E f_n - \int_E f\right| \leq \int_A |f_n - f| + 2M \cdot m(E \sim A) \tag{10.56}$$

[19]'Passge of limit under integral' is used by I.P. Natanson, in his "Theory of functions of real variable"

Let $\varepsilon > 0$. Since E a set of finte measure and f is a real-valued function. By Egoroff's theorem, there exists a measurable subset $A \subset E$ on which $\{f_n\}$ converges to f uniformly and $m(E \sim A) < \varepsilon/4M$. By uniform convergence on A, there exists $N \in \mathbb{N}$ such that

$$|f_n - f| < \frac{\varepsilon}{2 \cdot m(E)} \text{ on } A, \ \forall n \geq N$$

Thus, by monotonicity of integration we have,

$$\left| \int_E f_n - \int_E f \right| \leq \frac{\varepsilon}{2 \cdot m(E)} m(A) + 2M \cdot m(E \sim A) < \varepsilon$$

Therefore, $\lim_{n\to\infty} \int_E f_n = \int_E f$. □

Exercise

1.
2.
3.
4.
5.
6.
7.
8.
9.
10. Let f be a bounded, measurable function on E. Let A be measurable subset of E. Then,

$$\begin{aligned} \int_E f \cdot \chi_A &= \sup \left\{ \int_E \psi : \psi \text{ simple, }, f \cdot \chi_A \leq \psi \right\} \\ &= \sup \left\{ \int_A \psi : \psi \text{ simple, } f \leq \psi \right\} = \int_A f \end{aligned}$$

11.
12.
13.
14.
15.
16.

10.4.3 Lebesgue integral of a measurable non-negative function

Definitions 10.4.10 (vanishing)**.** A function f on E vanishes outside $A \subset E$ if $f(x) = 0$ for every $x \in E \sim A$.

Definitions 10.4.11 (support)**.** Let f be a function on E. Then support of f is the set $\{x \in E : f(x) \neq 0\}$ of all non-vanishing points of f in E.

Warning : The defintion of support is slightly different for measure spaces and topological spaces.†[20]

Definitions 10.4.12 (finite support)**.** A function f on E has finite support if support of f is of finite measure.

Definitions 10.4.13. Let f be a non-negative measurable function on E. Then the integral of f over E is the supremum of the integrals of bounded, measurable functions of finite support h such that $0 \leq h \leq f$ in E.

$$\int_E f = \sup\left\{\int_E h : h \text{ bounded, measurable, finite support },\ 0 \leq h \leq f\right\} \tag{10.57}$$

Theorem 10.4.9 (Chebychev's inequality)**.** *Let f be a non-negative, measurable function on E. Then for any $\lambda > 0$, we have*

$$m\{x \in E : f(x) \geq \lambda\} \ \leq \ \frac{1}{\lambda} \cdot \int_E f \tag{10.58}$$

In other words $\lambda \cdot m(E_\lambda) \leq \displaystyle\int_E f$ where $E_\lambda = \{x \in E : f(x) \geq \lambda\}$.

Proof. Suppose f is a non-negative, measurable function on E. Let $\lambda > 0$ be a real number. Define $E_\lambda = \{x \in E : f(x) \geq \lambda\}$

Case 1 : $m(E_\lambda) = \infty$
Let $n \in \mathbb{N}$. Define $E_{\lambda,n} = E_\lambda \cap [-n, n]$ and $\psi_n = \lambda \cdot \chi_{E_{\lambda,n}}$. Clearly, ψ_n is a bounded, measurable function of finite support.

$$\lambda \cdot m(E_{\lambda,n}) = \int_E \psi_n \text{ and } 0 \leq \psi_n \leq f \text{ on } E \text{ for every } n$$

By continuity of measure we have,

$$\infty = \lambda \cdot m(E_\lambda) = \lambda \lim_{n\to\infty} m(E_{\lambda,n}) = \lim_{n\to\infty} \int_E \psi_n \leq \int_E f$$

Therefore,

$$\lambda \cdot m(E_\lambda) = \int_E f$$

[20] "Support of a function f on a topological space is the **closure** of the set of all non-vanishing points of f." We use topological version for differential forms in multivariate real analysis and above definition for Lebesgue integration in integral transforms. Luckily, it is not apparent in our syllabus.

Case 2 : $m(E_\lambda) < \infty$
Define $h = \lambda \cdot \chi_{E_\lambda}$. Clearly, h is a bounded, measurable function of finite support and $0 \le h \le f$ on E. By definition of integral,

$$\lambda \cdot m(E_\lambda) = \int_E h \le \int_E f$$

Therefore,

$$m(E_\lambda) = \frac{1}{\lambda}\int_E h \le \frac{1}{\lambda}\int_E f$$

□

Theorem 10.4.10. *Let f be a non-negative measurable function on E. Then,*

$$\int_E f = 0 \iff f = 0 \text{ (a.e.) on } E \tag{10.59}$$

Proof. **Part 1**
Suppose $\displaystyle\int_E f = 0$. Then by Chebychev's inequality,

$$\forall n \in \mathbb{N},\ m\left(\left\{x \in E : f(x) \ge \frac{1}{n}\right\}\right) \le n \int_E f = 0$$

We have,

$$\{x \in E : f(x) > 0\} = \bigcup_{n\in\mathbb{N}} \left\{x \in E : f(x) \ge \frac{1}{n}\right\} = 0$$

Part 2
Suppose $f = 0$ (a.e.) on E. Let φ be a simple function and h be a bounded, measurable function of finite support such that $0 \le \varphi \le h \le f$.

We have, $f = 0$ (a.e.) on E, thus $\varphi = 0$ (a.e.) on E. Then $\varphi = 0$ (a.e.) on E, and we have $\int_E \varphi = 0$. Since $\int_E \varphi = 0$ for every simple function $\varphi \le h$, we have $\int_E h = 0$. Again, $\int_E h = 0$ for every bounded, measurable function $h \le f$ with finite support. Therefore, $\int_E f = 0$. □

Theorem 10.4.11 (linearity + monotonicity of integral of non-negative functions)**.** *Let f, g be non-negative, measurable functions on E. Then for any real numbers $\alpha, \beta > 0$f*[21] *, we have*

$$\int_E (\alpha f + \beta g) = \alpha \int_E f + \beta \int_E f \tag{10.60}$$

And we have,

$$\text{If } f \le g \text{ on } E, \text{ then } \int_E f \le \int_E g \tag{10.61}$$

Proof. **Step 1 :** $\displaystyle\int_E \alpha f = \alpha \int_E f$
Suppose $\alpha > 0$. For any bounded, measurable function h with finite support,

[21] We know that $\alpha f + \beta g$ is non-negative only when α and β are non-negative.

dominated by f, we have αh is also a bounded, measurable function with finite support such that $0 \leq \alpha h \leq \alpha f$ on E.

$$0 \leq h \leq f \iff 0 \leq \alpha h \leq \alpha f$$

Thus by the definition of Lebesgue integral for non-negative, measurable functions,

$$0 \leq \int_E \alpha h = \alpha \int_E h \leq \alpha \int_E f$$

Therefore, $\int_E \alpha f = \alpha \int_E f, \ \forall \alpha > 0$.

Step 2 : $\int_E f + g = \int_E f + \int_E g$

Suppose $0 \leq h \leq f$ and $0 \leq k \leq g$ on E. Then, $h + k$ is also a bounded, measurable function with finite support such that $0 \leq h + k \leq f + g$ on E.

$$0 \leq h + k \leq f + g \implies \int_E h + \int_E k = \int_E h + k \leq \int_E f + g \tag{10.62}$$

Let l be the supremum of all the bounded, measurable functions with finite support such that $0 \leq l \leq f + g$ on E. Then, it is enough to prove that,

$$\int_E f + g = \int_E l \leq \int_E f + \int_E g$$

Define $h = min\{f, l\}$ and $k = l - h$ on E. Let $x \in E$.
If $l(x) \leq f(x)$, then $h(x) = l(x) \leq f(x)$ and $k(x) = h(x) - l(x) = 0 \leq g(x)$.
If $l(x) > f(x)$, then $l(x) < g(x)$ since $l \leq f + g$ on E.
Thus, $h(x) = f(x) < l(x) < g(x)$ and $k(x) = h(x) - l(x) < 0 < f(x)$. Clearly, h, k are bounded, measurable functions with finite support.

We have, $0 \leq h \leq f$, $0 \leq k \leq g$, and $l = h + k$ on E. Then,

$$\int_E l = \int_E h + k = \int_E h + \int_E k \leq \int_E f + \int_E g \tag{10.63}$$

Therefore, by Step 1

$$\int_E \alpha f + \beta g = \int_E \alpha f + \int_E \beta g = \alpha \int_E f + \beta \int_E g$$

Suppose $f \leq g$. Let h be any bounded, measurable function with finite support such that $0 \leq h \leq f$. Then $h \leq g$. Thus, for any bounded, measurable function h with finite support and $h \leq f$, we have $\int_E h \leq \int_E g$. Therefore,

$$\int_E f = \sup \left\{ \int_E h : \text{ bounded,measurable,finite support }, h \leq f \right\} \leq \int_E g$$

$\square$

Theorem 10.4.12 (additivity over domain of integration of non-negative functions)**.** *Let f be a non-negative, measurable function on E. If A, B are disjoint subsets of E, then*

$$\int_{A\cup B} f = \int_A f + \int_B f \tag{10.64}$$

In particular, if $E_0 \subset E$ is of measure zero, then

$$\int_E f = \int_{E\sim E_0} f \tag{10.65}$$

Proof. Let f be a non-negative, measurable function defined on a measurable set E and A, B are two disjoint measurable subsets of E.

We have $\displaystyle\int_E f\cdot\chi_A = \int_A f$ and $f\cdot\chi_{A\cup B} = f\cdot(\chi_A+\chi_B) = f\cdot\chi_A + f\cdot\chi_B$. And if f is a bounded, measurable function with finite support, then $f\cdot\chi_A$ is also a bounded, measurable function with finite support for any measurable subset $A\subset E$. Therefore, by linearity of the Lebesgue integral of non-negative functions,

$$\int_{A\cup B} f = \int_E f\cdot\chi_{A\cup B} = \int_E f\cdot\chi_A + f\cdot\chi_B = \int_E f\cdot\chi_A + \int_E f\cdot\chi_B = \int_A f + \int_B f$$

Suppose E_0 is a subset of E of measure zero. Then, we have the **excision formula** by the additivity of integral over domain

$$\int_{(E\sim E_0)\cup E_0} f = \int_{E\sim E_0} f + \int_{E_0} f = \int_{E\sim E_0} f \tag{10.66}$$

since $m(E_0) = 0$ and $\displaystyle\int_{E_0} f = 0$. □

Fatou's Lemma and Passage of limit under Integral sign

Fatou's Lemma is a important tool for establishing passage of limit under integral sign.

Lemma 10.4.13 (Fatou)**.** *Let $\{f_n\}$ be a sequence of non-negative, measurable functions on E. If $\{f_n\}$ converges to f pointwise (a.e.) on E, then*

$$\int_E f \le \liminf \int_E f_n$$

Proof. Let $\{f_n\}$ be a sequence of functions converging to f pointwise a.e. on E. Let h be a bounded, measurable function with finite support and $0 \le h \le f$. Then $0 \le h \le M$, h has support, $E_0 = \{x \in E : h(x) \neq 0\}$ and $m(E_0) < \infty$.

Step 1 : Construction of h_n
Define $h_n = \min\{h, f_n\}$ on E. And $0 \le h_n \le M$. Then, h_n are measurable since each h_n is a bounded function defined on a set of finite measure. And h_n has finite support $E_n \subset E_0$.

Step 2 : Bounded Convergence Theorem
For each $x \in E$, $\{h_n(x)\}$ converges to $h(x)$ since $\{f_n(x)\}$ converges to $f(x)$. Since h_n are uniformly bounded, by Bounded Convgerence thereom we have

$$\lim_{n\to\infty} \int_E h_n = \lim_{n\to\infty} \int_{E_0} h_n = \int_{E_0} h$$

However $h_n \le f_n$. Therefore, $\displaystyle\int_E h_n \le \int_E f_n \le \liminf \int_E f_n$. ☐

Theorem 10.4.14 (monotone convergence)**.** *Let $\{f_n\}$ be an increasing sequence of non-negative, measurable functions on E. If $\{f_n\}$ converges to f pointwise (a.e.) on E, then $\displaystyle\lim_{n\to\infty} \int_E f_n = \int_E f$.*

Proof. Let $\{f_n\}$ be an increasing sequence of non-negative, measurable functions on E. Then, by Fatou's lemma

$$\int_E f \le \liminf \int_E f_n \tag{10.67}$$

However, $\forall n \in \mathbb{N}$, $f_n \le f$. Otherwise, $\{f_n\}$ won't converge to f. Thus, by monotonicity of Lebesgue integral for non-negative functions,

$$\forall n \in \mathbb{N}, \int_E f_n \le \int_E f \implies \limsup \int_E f_n \le \int_E f \tag{10.68}$$

Thus, limit inferior and limit superior are equal by sandwitch lemma.
Therefore, the sequence $\left\{\displaystyle\int_E f_n\right\}$ converges and $\displaystyle\int_E f = \lim_{n\to\infty} \int_E f_n$. ☐

Corollary 10.4.14.1. *Let $\{u_n\}$ be a sequence of non-negative, measurable functions on E. If $f = \displaystyle\sum_{n=1}^{\infty} u_n$ pointwise (a.e.) on E, then $\displaystyle\int_E f = \sum_{n=1}^{\infty} \int_E u_n$.*

Proof. Let $\{u_n\}$ be a sequence of non-negative, measurable functions on E and $f = \sum_{n=1}^{\infty} u_n$. Define $f_n = \sum_{k=1}^{n} u_k$. Then $\{f_n\}$ is an increasing sequence of non-negative, measurable functions. And by linearity and monotone convergence theorem,

$$\lim_{n\to\infty} \sum_{k=1}^{n} \int_E u_k = \lim_{n\to\infty} \int_E \sum_{k=1}^{n} u_k = \lim_{n\to\infty} \int_E f_n = \int_E f$$

☐

Definitions 10.4.14 (integrable function)**.** A non-negative, measurable function f on a measurable set E is integrable over E if $\displaystyle\int_E f < \infty$.

Theorem 10.4.15. *Let f be non-negative, measurable function which is integrable over E. Then f is finite a.e. on E.*

Proof. Suppose f is a non-negative measurable function defined on a measurable set E. And f is integrable over E. Then,

$$\int_E f = M < \infty$$

However by monotonicity and Chebychev's inequality,

$$m\{x \in E : f(x) = \infty\} \leq m\{x \in E : f(x) \geq n\} \leq \frac{1}{n}\int_E f = \frac{M}{n}, \forall n \in \mathbb{N}$$

As $n \to \infty$, we have $m\{x \in E : f(x) = \infty\} \to 0$.
Therefore, f is finite a.e. on E. □

Theorem 10.4.16 (Beppo Levi's Lemma)**.** *Let $\{f_n\}$ be an increasing sequence of non-negative, measurable functions on E. If the sequence of integrals $\left\{\int_E f_n\right\}$ are bounded, then $\{f_n\}$ converges pointwise on E to a measurable function f which is finite a.e. on E and*

$$\lim_{n\to\infty}\int_E f_n = \int_E f < \infty$$

Proof. Let $\{f_n\}$ be an increasing, sequence of non-negative, measurable functions on E. Since every monotone sequence of extended real numbers converges to an extended real number, $\{f_n\}$ converges to extended real valued function f on E. And by monotone convergence theorem, $\{\int_E f_n\}$ converges to $\int_E f$.

Suppose $\left\{\int_E f_n\right\}$ is a bounded sequence, then $\int_E f$ is also bounded. Thus, $\int_E f < \infty$. Therefore, f is finite a.e. on E, since f is integrable over E. □

Exercise

17.

18.

19.

20.

21.

22.

23.

24.

25.

26.

27.

10.4.4 General Lebesgue integral

Theorem 10.4.17. *Let f be a measurable function on E. Then f^+, f^- are integrable over E if and only if $|f|$ is integrable over E.*

Proof. Let f be a measurable function on E. Suppose f^+, f^- are non-negative, integrable functions over E. Then $|f| = f^+ + f^-$ is integrable over E by linearity.

Suppose $|f|$ is a non-negative, integrable function over E. Then, we have $0 \leq f^+ \leq |f|$ and $0 \leq f^- \leq |f|$. Thus, by monotonicity $\int_E f^+ \leq \int_E |f| < \infty$. Similarly, $\int_E f^- < \infty$. Therefore, f^+, f^- are integrable over E. □

Note : A measurable function f is integrable over E if $|f|$ is integrable over E.

Definitions 10.4.15 (general Lebesgue integral)**.** Let f be a measurable function such that $|f|$ is integrable over E. Then integral of f over E is

$$\int_E f = \int_E f^+ - \int_E f^- \tag{10.69}$$

Theorem 10.4.18. *Let f be integrable over E. Then f is finite a.e. on E and*

$$\int_E f = \int_{E \sim E_0} f \tag{10.70}$$

where E_0 is a subset of E and is of measure zero.

Proof. Suppose f is an integrable function. Then $f = f^+ - f^-$ where f^+, f^- are non-negative, integrable functions. And $|f| = f^+ + f^-$ is a non-negative integrable function on E, thus $|f|$ is finite a.e. on E. Therefore f is finite a.e. on E.

Suppose $E_0 \subset E$ such that $m(E_0) = 0$. Then by linearity and additivity over domain of the integral of non-negative measurable functions,

$$\begin{aligned}
\int_E f &= \int_E f^+ - f^- = \int_E f^+ + \int_E -f^- \\
&= \int_{E_0} f^+ + \int_{E \sim E_0} f^+ + \int_{E_0} -f^- + \int_{E \sim E_0} -f^- \\
&= \int_{E \sim E_0} f^+ + \int_{E \sim E_0} -f^- = \int_{E \sim E_0} f^+ - f^- = \int_{E \sim E_0} f
\end{aligned}$$

□

Theorem 10.4.19 (integral comparison test)**.** *Let f be a measurable function on E. Suppose there exists a non-negative function g such that g integrable over E and g dominates f on E. Then f is integrable over E and*

$$\left| \int_E f \right| \leq \int_E |f| \tag{10.71}$$

Proof. Let f be a measurable function. And g be a non-negative, measurable function such that g is integrable over E and $|f| \leq g$ on E. Clearly, by monotonicity of integral $|f|$ is integrable over E. Therefore, f is also integrable over E.

Also we have,

$$\left|\int_E f\right| = \left|\int_E f^+ - f^-\right| = \left|\int_E f^+ - \int_E f^-\right|$$

By triangular inequality,

$$\begin{aligned}
&\leq \left|\int_E f^+\right| + \left|\int_E f^-\right| = \int_E f^+ + \int_E f^- \text{ since } 0 \leq \int_E f^+ \\
&\leq \int_E f^+ + f^- = \int_E |f|
\end{aligned}$$

□

Linearity, Monotonicity, and Additivity over the domain of integration

Theorem 10.4.20 (linearlity+monotonicity of general integral)**.** *Let functions f, g be integrable over E. Then for any $\alpha, \beta \in \mathbb{R}$, function $\alpha f + \beta g$ is integrable over E and*

$$\int_E (\alpha f + \beta g) = \alpha \int_E f + \beta \int_E g \tag{10.72}$$

Moreover,

$$\textit{If } f \leq g \textit{ a.e. on } E, \textit{ then } \int_E f \leq \int_E g \tag{10.73}$$

Proof. **Step 1 :** $\int_E \alpha f = \alpha \int_E f$
If $\alpha > 0$, then $[\alpha f]^+ = \alpha f^+$ and $[\alpha f]^- = \alpha f^-$.
If $\alpha < 0$, then $[\alpha f]^+ = -\alpha f^-$ and $[\alpha f]^- = -\alpha f^+$.
Thus,

$$\begin{aligned}
\int_E \alpha f &= \int_E [\alpha f]^+ - \int_E [\alpha f]^- \\
&= \int_E \alpha f^+ - \int_E \alpha f^- = \alpha \int_E f^+ - \alpha \int_E f^- \\
&= \alpha \left(\int_E f^+ - \int_E f^-\right) = \alpha \int_E f
\end{aligned}$$

Step 2 : $\int_E f + g = \int_E f + \int_E g$
We have,

$$[f+g]^+ - [f+g]^- = f + g = (f^+ - f^-) + (g^+ - g^-)$$

Therefore,

$$[f+g]^+ + f^- + g^- = [f+g]^- + f^+ + g^+ \tag{10.74}$$

And,

$$\int_E [f+g]^+ + f^- + g^- = \int_E [f+g]^- + f^+ + g^+$$

By linearity of non-negative, measurable functions,

$$\int_E [f+g]^+ + \int_E f^- + \int_E g^- = \int_E [f+g]^- + \int_E f^+ + \int_E g^+$$

Since all the terms are finite, rearranging the terms we get,

$$\int_E [f+g]^+ - \int_E [f+g]^- = \int_E f^+ - \int_E f^- + \int_E g^+ - \int_E g^-$$

Therefore,

$$\int_E f+g = \int_E f + \int_E g$$

Let f, g be measurable functions such that $f \leq g$ on E. Then $h = g - f$ is a non-negative, measurable function. And

$$\int_E h = \int_E g - f = \int_E g - \int_E f \geq 0$$

Therefore,

$$\int_E f \leq \int_E g$$

$\square$

Theorem 10.4.21 (additivity over domain of integration)**.** *Let function f be integrable over E. Let A, B are disjoint, measurable subsets of E. Then,*

$$\int_{A\cup B} f = \int_A f + \int_B f \tag{10.75}$$

Proof. We have, $f \cdot \chi_{A\cup B} = f \cdot \chi_A + f \cdot \chi_B$.
Also by monotonicity, $|f \cdot \chi_A| \leq |f| \implies \displaystyle\int_E |f \cdot \chi_A| \leq \int_E |f| < \infty$.
Thus, $|f \cdot \chi_A|$ and $|f \cdot \chi_B|$ are integrable.

$$\int_{A\cup B} f = \int_E f \cdot (\chi_{A\cup B}) = \int_E f \cdot \chi_A + \int_E f \cdot \chi_B = \int_A f + \int_B f$$

$\square$

Sufficient Condition for Passage of limit under integral sign

Lebesgue dominated convergece theorems shows that **the existence of an integrable function or a sequence of functions (with integrable limit function) dominating every function in the sequence** is sufficient for the passage of limit under the integral sign.

Theorem 10.4.22 (Lebesgue dominated convergence)**.** *Let $\{f_n\}$ be a sequence of measurable functions on E. Suppose there exists a funtion g which is integrable over E and dominates every function in the sequence $\{f_n\}$ on E. If $\{f_n\}$ converges to f pointwise a.e. on E, then f is integrable on E and*

$$\lim_{n\to\infty} \int_E f_n = \int_E f \tag{10.76}$$

Proof. Suppose $\{f_n\}$ is a sequence of measurable functions on E. And g is a function which is integrable over E and $|f_n| \leq g$ on E. Suppose $\{f_n\}$ converges to f pointwise on E. Clearly, $|f| \leq g$ a.e. on E. And by monotonicity, f, f_n are integrable over E. We know that, since f, f_n are integrable over E, they are finite a.e. on E.

We have, $g - f$, $g - f_n$ are non-negative, measurable functions on E. And $\{g - f_n\}$ converges to $g - f$ pointwise a.e. on E. By Fatou's lemma,

$$\int_E (g - f) \leq \liminf \int_E (g - f_n)$$
$$\int_E g - \int_E f \leq \int_E g - \limsup \int_E f_n$$

Then,

$$\limsup \int_E f_n \leq \int_E f$$

Now consider, $\{g + f_n\}$ converging to $g + f$ pointwise a.e. on E. Again by Fatou's lemma,

$$\int_E (g + f) \leq \liminf \int_E (g + f_n)$$
$$\int_E g + \int_E f \leq \int_E g + \liminf \int_E f_n$$

Then,

$$\int_E f \leq \liminf \int_E f_n$$

Therefore,

$$\int_E f = \lim_{n\to\infty} \int_E f_n$$

□

Theorem 10.4.23 (General Lebesgue dominated convergence)**.** *Let $\{f_n\}$ be a sequence of measurable functions on E. And $\{f_n\}$ converges pointwise to f a.e. on E. Suppose there exists a sequence $\{g_n\}$ of non-negative, measurable functions on E that converges pointwise to g a.e. on E and g dominates every function in the sequence $\{f_n\}$.*

$$\text{If } \lim_{n\to\infty} \int_E g_n = \int_E g < \infty, \text{ then } \lim_{n\to\infty} \int_E f_n = \int_E f \tag{10.77}$$

Proof. Suppose $\{f_n\}$ is a sequence of measurable functions on E converging to f pointwise a.e. on E. Also suppose $\{g_n\}$ is another sequence of measurable functions on E convering to g pointwise a.e. on E such that $|f_n| \leq g_n$ and g is integrable on E. Clearly, $|f_n| \leq g_n \implies |f| \leq g$. And by monotonicity, f_n, f are integrable over E.

Consider, $\{g_n - f_n\}$ converging to $g - f$ pointwise a.e. on E. By Fatou's lemma,

$$\int_E (g - f) \leq \liminf \int_E (g_n - f_n)$$
$$\int_E g - \int_E f \leq \liminf \int_E g_n - \limsup \int_E f_n \dagger^{22}$$

Then,

$$\limsup \int_E f_n \leq \int_E f$$

Now consider, $\{g_n + f_n\}$ converging to $g + f$ pointwise a.e. on E. Again, by Fatou's lemma,

$$\int_E (g + f) \leq \liminf \int_E (g_n + f_n)$$
$$\int_E g + \int_E f \leq \liminf \int_E g_n + \liminf \int_E f_n \dagger^{23}$$

Then,

$$\int_E f \leq \liminf \int_E f_n$$

Therefore,

$$\int_E f = \lim_{n\to\infty} \int_E f_n$$

□

Exercise

28.

29.

30.

31.

32.

33.

34.

35.

36.

10.17 General Measure Spaces

Let X be any non-empty set. Let μ be any non-negative set function defined on a family of subsets of X. Then there exists a σ-algrebra $\mathcal{M}$ of measurable functions with respect to the outer measure μ^* induced by μ.

10.17.1 Measures and Measurable Sets

Definitions 10.17.1 (measure space)**.** Let X be non-empty set and $\mathcal{M}$ be a σ-algebra of subset of X. Then $(X, \mathcal{M})$ is a measure space. And $E \subset X$ is a measurable subset of X if $E \in \mathcal{M}$.

Definitions 10.17.2 (measure)**.** A set function $\mu : \mathcal{M} \to [0, \infty]$ is a measure if $\mu(\phi) = 0$ and μ is countably additive†[24].

A few examples of measure space

1. $(\mathbb{R}, \mathcal{L}, m)$
   The set of real numbers $\mathbb{R}$ together with the σ-algebra $\mathcal{L}$ of Lebesgue measurable functions and the Lebesuge measure m is a measure space.

2. $(\mathbb{R}, \mathcal{B}, m)$
   The set of real numbers together with the σ-algebra $\mathcal{B}$ of Borel sets and Lebesgue measure m is a measure space.

3. $(X, 2^X, \eta)$
   A non-empty set X together with its powerset 2^X and counting measure, $\eta : \mathcal{M} \to [0, \infty)$ given by $\eta(A) = |A|$ is a measure space.

4. $(X, \mathcal{M}, \delta_{x_0})$
   A non-empty set X together with a σ-algebra $\mathcal{M}$ of subset of X and Dirac measure concentrated about $x_0 \in X$, $\delta_{x_0} : \mathcal{M} \to \{0, 1\}$ given by

$$\delta_{x_0}(A) = \begin{cases} 1 & x_0 \in A \\ 0 & x_0 \notin A \end{cases}$$

   is a measure space.

5. $(X, \mathcal{C}, \mu)$
   A non-empty set X together with the family of all countable subsets and their complements, $\mathcal{C}$ and measure $\mu : \mathcal{C} \to \{0, 1\}$ given by

$$\mu(A) = \begin{cases} 0 & A \text{ is countable} \\ 1 & A^c \text{ is countable} \end{cases}$$

   is a measure space.

Note : Lebesgue measure offers an addition property, that is measure of an interval is its length.

[24] We have dropped an axiom - "length of the interval is its measure."

Measure Subspace

Let $(X, \mathcal{M}, \mu)$ be a measure space. Let $X_0 \in \mathcal{M}$. Then $(X_0, \mathcal{M}_0, \mu_0)$ is also a measure space where $\mathcal{M}_0$ is $\mathcal{M}$ respectricted to X_0 and μ_0 is μ restricted to X_0.

For example, let $I = [0, 1]$. Clearly, $I \in \mathcal{L}$ since I is a Borel set, and measurable. Then $(I, \mathcal{L}_{|_I}, m_{|_I})$ is a measure space on the closed interval I obtained from the Lebesgue measure Space $(\mathbb{R}, \mathcal{L}, m)$.

Properties : Additivity, Monotonicity, Excision

1. Finite Additivity
   For any finite, disjoint collection of measurable sets $\{E_k\}_{k=1}^n$, we have
   $$\mu\left(\bigcup_{k=1}^n E_k\right) = \sum_{k=1}^n \mu(E_k)$$
   Proof. Countably additive implies finitely additive. □

2. Monotonicity
   Suppose A, B are measurable subsets and $A \subset B$, then $\mu(A) \leq \mu(B)$.

   Proof. $\mu(B) = \mu(A) + \mu(B \sim A) \implies \mu(B) \geq \mu(A)$, since $\mu(B \sim A) \geq 0$. □

3. Excision
   Suppose A, B are measurable subsets, $A \subset B$ and $\mu(A) < \infty$, then $\mu(B \sim A) = \mu(B) - \mu(A)$. And if $\mu(A) = 0$, then $\mu(B \sim A) = \mu(B)$.

   Proof. $\mu(B) = \mu(A) + \mu(B \sim A) \implies \mu(B \sim A) = \mu(B) - \mu(A)$. □

4. Countable Monotonicity
   Suppose a countable collection of measurable sets $\{E_k\}_{k=1}^\infty$ covers another measurable sets E. Then, $\mu(E) \leq \displaystyle\sum_{k=1}^\infty \mu(E_k)$.

   Proof. Suppose $\{E_k\}_{k=1}^\infty$ is a countable collection of measurable sets covering E. Then, we can construct $\{F_k\}_{k=1}^\infty$ is a countable collection of disjoint, measurable sets where $F_k = E_k \sim \displaystyle\bigcup_{n=1}^{k-1} E_n$.

   Clearly, $\displaystyle\bigcup_{k=1}^\infty E_k = \bigcup_{k=1}^\infty F_k$ and $\forall k,\ \mu(F_k) \leq \mu(E_k)$ since $F_k \subset E_k$.

   Therefore $\mu(E) \leq \mu\left(\displaystyle\bigcup_{k=1}^\infty E_k\right) = \mu\left(\bigcup_{k=1}^\infty F_k\right) = \sum_{k=1}^\infty \mu(F_k) \leq \sum_{k=1}^\infty \mu(E_k)$.
   □

Properties : Continuity of Measure

1. If $\{A_k\}_{k=1}^{\infty}$ is an ascending sequence of measurable sets, then

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \lim_{k\to\infty} \mu(A_k)$$

2. If $\{B_k\}_{k=1}^{\infty}$ is a descending sequence of measurable sets and $\mu(B_1) < \infty$, then

$$\mu\left(\bigcap_{k=1}^{\infty} B_k\right) = \lim_{k\to\infty} \mu(B_k)$$

Proof. Refer proof on page 212 □

Properties : Borel-Cantelli Lemma

Let $(X, \mathcal{M}, \mu)$ be a measure space. And $\{E_k\}_{k=1}^{\infty}$ a countable collection of measurable sets such tha t $\sum_{k=1}^{\infty} E_k < \infty$. Then almost all $x \in X$ belongs to at most a finite number of E_k's.

Proof. Refer proof on page 213 □

σ-finite Measure

Definitions 10.17.3 (σ-finite)**.** Let $(X, \mathcal{M}, \mu)$ be a measure space. Measure μ is finite if $\mu(X) < \infty$. Measure μ is σ-finite if X is a countble union of measurable sets of finite measure. A measurable subset E is of finite measure if $\mu(E) < \infty$. And E is σ-finite if E is a countable union of measurable sets of finite measure.

For example, Lebesgue measure is σ-finite since $\mathbb{R} = \bigcup_{n\in\mathbb{Z}}[n, n+1]$. But, counting measure on $[0, 1]$ is not σ-finite since $[0, 1]$ is uncountably infinite and countable union of finite subsets is countable.

Complete Measure Space

Definitions 10.17.4 (complete measure space)**.** Let $(X, \mathcal{M}, \mu)$ be a measure space. Then the measure space is complete if every subset of sets of measure zero are measurable.

Note : $(\mathbb{R}, \mathcal{L}, m)$ is complete, however, $(\mathbb{R}, \mathcal{B}, m)$ is not complete since cantor set C of measure zero has a non-borel subset $\psi^{-1}(W)$ where W is a choice set of $\psi(C)$ under rational equivalence.

Theorem 10.17.1 (completion)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let $\mathcal{M}_0 = \{A \cup B : A \subset C \in \mathcal{M}, \mu(C) = 0, B \in \mathcal{M}\}$. And $\mu_0(E) = \mu(B)$ where $E = A \cup B$. Then, $(X, \mathcal{M}_0, \mu_0)$ is a completion of $(X, \mathcal{M}, \mu)$.*

Proof. Let $(X, \mathcal{M}, \mu)$ be a measure space.

Step 1 : $\mathcal{M}_0$ is a σ-algebra containing $\mathcal{M}$
Define a family of subsets of X, $\mathcal{M}_0$ in such a way that a subset $E \in \mathcal{M}_0$ if it can be expressed as union of a subset C such that $\mu(C) = 0$ and a set $B \in \mathcal{M}$. Clearly, $E \in \mathcal{M} \implies E = E \cup \phi \in \mathcal{M}_0$. Thus, $\mathcal{M} \subset \mathcal{M}_0$.

Suppose $E = A \cup B$, then

$$X \sim E = X \sim (A \cup B) = (X \sim D) \cup (X \sim (C \cup B))$$

where $D = C \sim (A \cup B)$. Clearly $D \subset C$ and $C \cup B$ is a union of measurable sets and therefore measurable. Therefore, for any set $E \in \mathcal{M}_0$, we have $X \sim E \in \mathcal{M}_0$.

Suppose $\{E_k\}$ be a collection of sets in $\mathcal{M}_0$. Let $E_k = A_k \cup B_k$ where A_k's are subsets of some set C_k such that $\mu(C_k) = 0$. Clearly,$\mu(C) = \mu(\cup_{k=1}^{\infty} C_k) = 0$. And $A = \cup_{k=1}^{\infty} A_k$ is a subset of C. Also $B = \cup_{k=1} \infty B_k \in \mathcal{M}$ since $\mathcal{M}$ is a σ-algebra. Therefore,

$$E = \bigcup_{k=1}^{\infty} E_k = \bigcup_{k=1}^{\infty} (A_k \cup B_k) = \bigcup_{k=1}^{\infty} A_k \cup \bigcup_{k=1}^{\infty} B_k = A \cup B \in \mathcal{M}_0$$

Therefore, by de Morgan's law, $\mathcal{M}_0$ is a σ-algebra.

Step 2 : μ_0 is well-defined
Let $E \in \mathcal{M}_0$. We have $E = A \cup B$ where A is a subset C such that $\mu(C) = 0$, $B \in \mathcal{M}$ and $\mu_0(E) = \mu(B)$. Suppose there exists A', B' such that $E = A' \cup B'$ satifying the conditions. Then $\mu(B) = \mu(B')$. Suppose $\mu(B) < \mu(B')$, without loss of generality. Then $\mu(B' \sim B) > 0 \implies 0 < \mu(A \sim A') \leq \mu(C) = 0$ which is a contradiction. Therefore, $\mu_0(E)$ is well-defined irrespective of the representation $E = A \cup B$.

Step 3 : $(X, \mathcal{M}_0, \mu_0)$ is complete
Suppose A is a subset of some set C such that $\mu(C) = 0$. Then there exists $E \in \mathcal{M}_0$ such that $E = A \cup \phi$ where $\phi \in \mathcal{M}$. Therefore, $(X, \mathcal{M}_0, \mu_0)$ is complete. □

Exercise

Definitions 10.17.5 (semi-finite)**.** Measure μ is semi-finite if every subset of infinite measure contains measurable subsets of arbitrarily large finite measure.

Definitions 10.17.6 (locally measurable)**.** A subset E is locally measurable if $B \cap E$ is measurable for any subset B of finite measure.

Definitions 10.17.7 (saturated)**.** A measure space is saturated if every locally measurable subset is measurable.

1.

2.

3.

4.

5.

6.

7. (a) If μ, ν are measures on $(X, \mathcal{M})$, then $\lambda = \mu + \nu$ is also a measuare on $(X, \mathcal{M})$.
   (b) Suppose ν, μ are measures on $(X, \mathcal{M})$. If $\nu \leq \mu$, then there exists measure λ such that $\nu + \lambda = \mu$.
   (c) Suppose $\nu \leq \mu$ and μ is σ-finite, then λ is unique.
   (d) Suppose $\nu \leq \mu$. Then there exists a minimal λ such that $\nu + \lambda = \mu$.

8. (a) Every σ-finite measure is semi-finite.
   (b) Any measure $\mu = \mu_1 + \mu_2$ where measure μ_1 is semifinite and measure μ_2 assumes either 0 or ∞.

9. The completion $(X, \mathcal{M}_0, \mu_0)$ of $(X, \mathcal{M}, \mu)$ is minimal with respect to ?

10. (a) Every σ-finite measure is saturated.
    (b) Locally measurable sets is a σ-algebra.
    (c) $(X, 2^X, \underline{\mu})$ is a saturation of $(X, \mathcal{M}, \mu)$ where $\overline{\mu}(E) = \infty$ if $E \notin \mathcal{M}$.
    (d) $(X, \mathcal{C}, \overline{\mu})$ is a saturation of $(X, \mathcal{M}, \mu)$ where $\mathcal{C}$ is the σ-algebra of locally measurable sets and $\overline{\mu}(E) = \sup\{\mu(B) : B \subset E,\ B \in \mathcal{M}\}$

11. If μ, η are measures on $(X, \mathcal{M})$, then $\max\{\mu, \eta\}$ is also a measure on $\mathcal{M}$.

10.17.2 Signed measures

Definitions 10.17.8 (signed measure)**.** A set function $\nu : \mathcal{M} \to [-\infty, \infty]$ is a signed measure on $X, \mathcal{M}$ if

1. ν assumes at most one of the infinities
2. $\nu(\phi) = 0$ and
3. If ν is countably additive for a disjoint collection of measurable sets $\{E_k\}_{k=1}^{\infty}$ and $\nu(\cup_{k=1}^{\infty} E_k) < \infty$, then the series $\sum_{k=1}^{\infty} \nu(E_k)$ is absolutely convergent.

Note : Difference of two measures, one of which is finite is always a signed measure. And we later prove that evey signed measure is of this form.

Definitions 10.17.9. Let ν be a signed measure.

positive set A set is positive with respect to ν if it is measurable and every measurable subset $E \subset A$ has positive measure.

negative set A set is negative with respect to ν if it is measurable and every measurable subset of it has negative measure.

null (measure) set A set is null with respect to ν if it is measurable and every measurable subset of it has measure zero.

Note : Signed measure restricted to any positive set is a measure.

Theorem 10.17.2. *Let ν be a signed measure on $(X, \mathcal{M})$. Then every measurable subset of a positive set is positive and union of countable collection of positive sets is positive.*

Proof. Let $\{A_k\}_{k=1}^{\infty}$ be countable collection of postive sets. Let $A = \displaystyle\bigcup_{k=1}^{\infty} A_k$ and E be a measurable subset of A. Consider the disjoint collection of countable subsets, $\{E_k\}_{k=1}^{\infty}$ where

$$E_k = (E \cap A_k) \sim \bigcup_{k=1}^{n-1} E_k$$

Clearly $\displaystyle\bigcup_{k=1}^{\infty} E_k = E$ and $\nu(E) = \displaystyle\sum_{k=1}^{\infty} \nu(E_k) \geq 0$. Since $E \subset A$ is arbitrary, every measurable subset of A is of non-negative measure. Therefore, union of positive sets is also a positive set. □

Lemma 10.17.3 (Hahn)**.** *Let ν be a signed measure on $(X, \mathcal{M})$. And E is a measurable set of finite, positive measure. Then there exists subset $A \subset E$ such that A is positive and is of positive measure.*

Proof. Suppose E is not a positive set, otherwise the proof is complete. Let m_1 be the smallest natural number†[25] such that there exists a measurable set with measure less than $-\frac{1}{m}$. Let $E_1 \subset E$ be a measurable subset with $\nu(E_1) < -\frac{1}{m_1}$.

Let n be a natural number and $m_1, m_2, \ldots, m_n$ are natural numbers where m_k's are the smallest natural numbers such that there exists a measurable subset E_k,

$$E_k \subset E \sim \bigcup_{j=1}^{k-1} E_j \text{ and } \nu(E_k) < -\frac{1}{m_k}$$

If this sequence terminates, then $E \sim \displaystyle\bigcup_{j=1}^{n} E_j$ is a positive set with positive measrue. Suppose the sequence does not terminate. Define $A = E \sim \displaystyle\bigcup_{j=1}^{\infty} E_j$. Then A is a positive set by construction. Suppose A has a measurable subset B, then $\nu(B) \geq \displaystyle\lim_{k\to\infty} -\frac{1}{m_k} = 0$. And measure of A is positive, since we have $E = A \cup \displaystyle\bigcup_{j=1}^{\infty} E_j$ and $0 < \nu(E) = \nu(A) + \displaystyle\sum_{j=1}^{\infty} \nu(E_j)$ where $\nu(E_j)$'s are all negative. Therefore, every measurable set E with positive measure has a postive subset A with positive measure. □

[25] Since E has measurable subsets of negative measure and negative measures are bounded above by 0, there exists a least upper bound $-M < -\frac{1}{m}$ for a natural number m.

Theorem 10.17.4 (Hahn Decomposition). *Let ν be a signed measure on $(X, \mathcal{M})$. Then there exists a positive set A and a negative set B for which $X = A \cup B$ and $A \cap B = \phi$. And we write $[A, B] = X$.*

Proof. Let $(X, \mathcal{M}$ be measure space with signed measure ν. Let $\mathcal{P}$ be the collection of positive subset of X. Let $\lambda = \sup\{\nu(E) : E \in \mathcal{P}\}$. And $\lambda \geq 0$ since ϕ is trivially a positive set in $\mathcal{P}$ with measure zero.

Let $\{A_n\}$ be a countable collection of positive subsets of X such that $\lim\limits_{n\to\infty} \nu(A_k) = \lambda$. Define $A = \bigcup\limits_{n=1}^{\infty} A_n$. Then A is also a positive set and $\nu(A) \leq \sup\{\nu(E) : E \in \mathcal{P}\} = \lambda$ since $A \in \mathcal{P}$. Also for each k, $A \sim A_k \subset A$ and $\nu(A) = \nu(A \sim A_k) + \nu(A_k) \geq \nu(A_k)$. This inequality is true for every k, and for the limit as well. Thus, $\nu(A) \geq \lambda$.

Let $B = X \sim A$. Then B is a negative set by construction as $X \sim A$ doesn't have any measurable subset of positive measure. Suppose E be a subset of B with positive measure. Then $\nu(A \cap E) = \nu(A) + \nu(E) < \nu(A) = \lambda$ which is a contradiction. $\square$

Definitions 10.17.10 (singular measures). Two signed measures ν_1, ν_2 are singular if there exists disjoint subsets $A, B \subset X$ such that $X = A \cup B$ and $\nu_1(A) = \nu_2(B) = 0$. If ν_1, ν_2 are singular, then we write $\nu_1 \perp \nu_2$.

Theorem 10.17.5 (Jordan-Hahn Decomposition). *Let ν be a signed measure on $(X, \mathcal{M})$. Then there two mutually singular measures ν^+, ν^- on $(X, \mathcal{M})$ such that $\nu = \nu^+ - \nu^-$. Moreover, there is only one such pair of mutually singular measures.*

Proof. Let $(X, \mathcal{M})$ be a measure space with signed meausre ν. Then by Hahn decomposition theorem, X has a partition $X = A \cup B$, $A \cap B = \phi$ such that A, B are positive and negative sets[ν] respectively.

Define $\nu^+(E) = \nu(E \cap A)$ and $\nu^-(E) = -\nu(E \cap B)$. Then ν^+, ν^- are measure and $\nu = \nu^+ - \nu^-$. Also, $\nu^+(B) = \nu(B \cap A) = \nu(\phi) = 0$ and $\nu^-(A) = \nu(A \cap B) = \nu(\phi) = 0$. $\square$

Exercise

12.

13.

14.

15.

16.

17.

10.17.3 Caratheodory measure

Definitions 10.17.11 (outer measure)**.** A set function $\mu^* : 2^X \to [0, \infty]$ is an outer measure if $\mu^*(\phi) = 0$. and μ^* is countably monotone.†[26]

Definitions 10.17.12 (measurable set)**.** A set E is measurable if for any subset A of X, we have

$$\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^c)$$

Since μ^* is monotone, we can use simpler condition

$$\mu^*(A) \geq \mu^*(A \cap E) + \mu^*(A \cap E^c)$$

Theorem 10.17.6. *Union of finite collection of measurable sets is measurable.*

Proof. Refer proof on page 202 □

Theorem 10.17.7. *Let $A \subset X$ and $\{E_k\}_{k=1}^n$ be a finite, disjoint collection of measurable sets. Then*

$$\mu^*\left(A \cap \bigcup_{k=1}^{n} E_k\right) = \sum_{k=1}^{n} \mu^*(A \cap E_k)$$

In particular, the restriction of μ^* to the finite collection of measurable sets is finitely additive.

Proof. —yet to complete— □

Theorem 10.17.8. *Union of countable collection of measurable sets is measurable.*

Proof. Refer proof on page 203 □

Theorem 10.17.9. *Let μ^* be an outer measure on 2^X. Then the collection $\mathcal{M}$ of measurable subsets is a σ-algebra. If $\overline{\mu}$ is the restriction of μ^* to $\mathcal{M}$, then $(X, \mathcal{M}, \mu)$ is a complete measure space.*

Proof. —yet to complete— □

10.18 Integration over General measure Spaces

10.18.1 Measurable Functions

Theorem 10.18.1. *Let $(X, \mathcal{M})$ be a measure space. Let function f be defined on X. Then the following statements are equivalent*

1. $\forall c \in \mathbb{R},\ \{x \in X : f(x) < c\}$ *is measurable.*
2. $\forall c \in \mathbb{R},\ \{x \in X : f(x) \leq c\}$ *is measurable.*
3. $\forall c \in \mathbb{R},\ \{x \in X : f(x) > c\}$ *is measurable.*
4. $\forall c \in \mathbb{R},\ \{x \in X : f(x) \geq c\}$ *is measurable.*

[26] Further simplifying the axiom countable additivity into countable monotonicity.

Each of the above statements imply that $\forall c \in \mathbb{R}$, $\{x \in X : f(x) = c\}$ *is measurable.*

Proof. Refer proof on page 219 $\square$

Definitions 10.18.1 (measurable function)**.** Let $(X, \mathcal{M})$ be a measure space. A function f on X is measurable if one of the equivalent statements above and hence all are true.

Theorem 10.18.2. *Let* $(X, \mathcal{M})$ *be a measure space and function* f *be defined on* X*. Then* f *is measurable if and only if for every open set* $\mathcal{O}$, $f^{-1}(\mathcal{O})$ *is open.*

Proof. Refer proof on page 220 $\square$

Theorem 10.18.3. *Let* $(X, \mathcal{M})$ *be a complete measure space. And* X_0 *is a measurable subset of* X *for which* $\mu(X \sim X_0) = 0$*. Then* f *is measurable if and only if its restriction to* X_0 *is measurable. In particular, let two functions* $g = h$ *a.e. on* X*,* g *is measurable if and only if* h *is measurable.*

Proof. Refer proof on page 221 $\square$

Theorem 10.18.4. *Let* $(X, \mathcal{M})$ *be a measure space. And* f, g *are measurable functions on* X*.*

1. *Linearity :* $\alpha f + \beta g$ *is measurable* $\forall \alpha, \beta \in \mathbb{R}$.

2. *Products :* $f \cdot g$ *is measurable.*

3. *minimum/maximum :* $\max\{f, g\}, \min\{f, g\}$ *are measurable.*

Proof. $\square$

Theorem 10.18.5. *Let* $(X, \mathcal{M})$ *be a measure space. Let* f *be continuous function and* g *be measurable, then* $f \circ g$ *is measurable.*

Proof. $\square$

Theorem 10.18.6. *Let* $(X, \mathcal{M}, \mu)$ *be a measure space. Let* $\{f_n\}$ *be a sequence of measurable functions on* X *for which* $\{f_n\}$ *converges to* f *pointwise a.e. on* X*. If either* $(X, \mathcal{M}, \mu)$ *is complete or convergence is pointwise on* X*, then* f *is measurable.*

Proof. $\square$

Corollary 10.18.6.1. *Let* $(X, \mathcal{M})$ *be a measure space. Let* $\{f_n\}$ *be a sequence of measurable functions on* X*. Then* $\sup\{f_n\}$, $\inf\{f_n\}, \limsup\{f_n\}, \liminf\{f_n\}$ *are measurable.*

Proof. $\square$

10.18.2 Integration of non-negative measurable functions

Definitions 10.18.2. Let $(X, \mathcal{M}, \mu)$ be a measure space. And ψ be a non-negative simple function on X. Then

$$\int_X \psi \; d\mu = \sum_{k=1}^{n} c_k \cdot \mu(E_k) \text{ where } \psi = \sum_{k=1}^{n} c_k \cdot \chi_{E_k} \tag{10.78}$$

And for $E \subset X$, we have

$$\int_E \psi \; d\mu = \int_X \psi \cdot \chi_E \; d\mu = \sum_{k=1}^{n} c_k \cdot \mu(E_k \cap E) \text{ where } \psi = \sum_{k=1}^{n} c_k \cdot \chi_{E_k} \tag{10.79}$$

Theorem 10.18.7. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let ψ, φ be non-negative simple functions.*

1. *Linearity*
   Let α, β be positive real numbers.
$$\int_X (\alpha\psi + \beta\varphi) \; d\mu = \alpha \int_X \psi \; d\mu + \beta \int_X \varphi \; d\mu \tag{10.80}$$

2. *Additivity over domain of integration*
   Let A, B be disjoint measurable subsets of X
$$\int_{A\cup B} \psi \; d\mu = \int_A \psi \; d\mu + \int_B \psi \; d\mu \tag{10.81}$$

3. *Suppose $X_0 \subset X$ such that $\mu(X \sim X_0) = 0$, then*
$$\int_X \psi \; d\mu = \int_{X_0} \psi \; d\mu \tag{10.82}$$

4. *Monotonicity*
   Suppose $\psi \le \varphi$ a.e. on X, then
$$\int_X \psi \; d\mu \le \int_X \varphi \; d\mu \tag{10.83}$$

Proof. □

Definitions 10.18.3. Let $(X, \mathcal{M}, \mu)$ be a measure space. And f be a non-negative, extended real-valued, measurable function on X. Then

$$\int_X f \; d\mu = \sup\left\{\int_X \varphi \; d\mu : \varphi \text{ simple, } 0 \le \varphi \le f\right\} \tag{10.84}$$

And if $E \subset X$, then

$$\int_E f \; d\mu = \int_X f \cdot \chi_E \; d\mu = \sup\left\{\int_X \varphi \cdot \chi_E \; d\mu : \varphi \text{ simple, } 0 \le \varphi \le f \text{ on E}\right\} \tag{10.85}$$

Theorem 10.18.8. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let g, h be non-negative, measurable functions. Then*

1. *Linearity (Part 1)*
   Let α be a positive real number.

$$\int_X \alpha f \; d\mu = \alpha \int_X f \; d\mu \tag{10.86}$$

2. *Suppose $X_0 \subset X$ such that $\mu(X \sim X_0) = 0$, then*

$$\int_X f \; d\mu = \int_{X_0} f \; d\mu \tag{10.87}$$

3. *Monotonicity*
   Suppose $f \leq g$ a.e. on X, then

$$\int_X f \; d\mu \leq \int_X g \; d\mu \tag{10.88}$$

Proof. □

Theorem 10.18.9 (Chebychev's Inequality)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f be a non-negative measurable function on X. Let $\lambda \in \mathbb{R}$.*

$$\mu\{x \in X : f(x) \geq \lambda\} \leq \frac{1}{\lambda} \int_X f \; d\mu$$

Proof. Refer proof on page 236 □

Theorem 10.18.10. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f be a non-negative measurable function on X for which $\int_X f \; d\mu < \infty$. Then f is finite a.e. on X and $\{x \in X : f(x) \geq \frac{1}{n}\}$ is σ-finite.*

Proof. □

Lemma 10.18.11 (Fatou)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let $\{f_n\}$ be a sequence of non-negative, measurable functions converging to f pointwise a.e. on X. If f is measurable, then*

$$\int_X f \; d\mu \leq \liminf \int_X f_n \; d\mu$$

Proof. □

Theorem 10.18.12 (monotone convergence)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. And $\{f_n\}$ be an increasing sequence of non-negative measurable functions converging to f pointwise on X. Then*

$$\lim_{n \to \infty} \int_X f_n \; d\mu = \int_X f \; d\mu$$

Proof. □

Lemma 10.18.13 (Beppo-Levi)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let $\{f_n\}$ be an increasing sequence of non-negative measurable functions on X. If the sequnce of integrals $\left\{\int_X f_n \, d\mu\right\}$ is bounded, then $\{f_n\}$ converges to a function f pointwise on X which is measurable, finite a.e. on X and*

$$\lim_{n\to\infty} \int_X f_n \, d\mu = \int_X f \, d\mu < \infty$$

Proof. □

Theorem 10.18.14 (simple approximation)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f be a non-negative measurable function on X. Then there exists an increasing sequence $\{\psi_n\}$ of simple functions on X that converges pointwise on X and*

$$\lim_{n\to\infty} \int_X \psi_n \, d\mu = \int_X f \, d\mu$$

Proof. Refer proof on page 225 □

Theorem 10.18.15. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f, g abe non-negative, measurable functions on X. If $\alpha, \beta > 0$, then*

$$\int_X (\alpha f + \beta g) \, d\mu = \alpha \int_X f \, d\mu + \beta \int_X g \, d\mu$$

Proof. □

Theorem 10.18.16. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f be a non-negative, measurable function on X. Then there exists an increasing sequnce of simple functions $\{\psi_n\}$ that converges to f pointwise on X and*

$$\lim_{n\to\infty} \int_X \psi_n \, d\mu = \int_X f \, d\mu$$

Proof. □

Theorem 10.18.17. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f, g be non-negative measurable functions on X. If $\alpha, \beta > 0$, then*

$$\int_E (\alpha f + \beta g) \, d\mu = \alpha \int_X f \, d\mu + \beta \int_X g \, d\mu$$

Proof. □

Definitions 10.18.4 (integrable)**.** Let $(X, \mathcal{M}, \mu)$ be a measure space. A non-negative, measurable function f is integrable over X with respect to μ if $\int_X f \, d\mu < \infty$.

Exercise

17.

18.

19.

20.

21.

22.

23.

24.

25.

26.

10.18.3 Integration of general measurable functions

Definitions 10.18.5. Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f be a measurable function on X. Then f be an integrable function over X with respect to μ, if $|f|$ is integrable over X with respect to μ. And

$$\int_X f \; d\mu = \int_X f^+ \; d\mu - \int_X f^- \; d\mu$$

Theorem 10.18.18 (integral comparison test)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f be a measurable function on X. If g is integrable over X and g dominates f on X, ie, $|f| \leq g$. Then f is integrable over X and*

$$\left| \int_X f \; d\mu \right| \leq \int_X |f| \; d\mu \leq \int_X g \; d\mu$$

Proof. □

Theorem 10.18.19. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f, g be integrable over X with respect to μ.*

1. *Linearlity*

$$\int_X (\alpha f + \beta g) \; d\mu = \alpha \int_X f \; d\mu + \beta \int_X f \; d\mu$$

2. *Monotonicity*
   If $f \leq g$ a.e. on X, then

$$\int_X f \; d\mu \leq \int_X g \; d\mu$$

3. *Additivity over domain of integration*
   Let A, B be disjoint, measurable subsets of X. Then

$$\int_{A\cup B} f\ d\mu = \int_A f\ d\mu + \int_B f\ d\mu$$

Proof. □

Theorem 10.18.20 (countably additivity over the domain of integration)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f be a function integrable over X with respect to μ. Let $\{X_n\}_{n=1}^{\infty}$ be a disjoint, collection of measurable sets whose union is X. Then*

$$\int_X f\ d\mu = \sum_{n=1}^{\infty} \int_{X_n} f\ d\mu$$

Proof. □

Theorem 10.18.21 (continuity of integation)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f be a function integrable over X with respect to μ.*

1. *If $\{X_n\}_{n=1}^{\infty}$ is an ascending collection of measurable sets whose union is X. Then*

$$\int_X f\ d\mu = \lim_{n\to\infty} \int_{X_n} f\ d\mu$$

2. *If $\{X_n\}_{n=1}^{\infty}$ is a descending collection of measurable sets, then*

$$\int_{\bigcap_{n=1}^{\infty} X_n} f\ d\mu = \lim_{n\to\infty} \int_{X_n} f\ d\mu$$

Proof. □

Theorem 10.18.22. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let f be a measurable function on X. If f is bounded and vanishes outside a set of finite measure, then f is integrable over X.*

Proof. □

Corollary 10.18.22.1. *Let X be a compact, topological space. Let $\mathcal{M}$ be a σ-algebra containing the topology. If f is continuous, real-valued function on X and $(X, \mathcal{M}, \mu)$ is a finite measure space, then f is integrable over X with respect to μ.*

Proof. □

Theorem 10.18.23 (Lebesgue Dominated Convergence)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let $\{f_n\}$ be a sequence of measurable functions converging to a measurable function f pointwise a.e. on X. Let g be an integrable, non-negative function dominating the sequence $\{f_n\}$. Then f is integrable over X and,*

$$\lim_{n\to\infty} \int_X f_n\ d\mu = \int_X f\ d\mu$$

Proof. □

Vitali Convergence Theorem

Definitions 10.18.6 (uniformly integrable)**.** Let $(X, \mathcal{M}, \mu)$ be a measure space. Let $\{f_n\}$ be a sequence of functions integrable over X. The sequence $\{f_n\}$ is uniformly integrable over X if for any $\varepsilon > 0$, there exist $\delta > 0$ such that for any measurable subset $E \subset X$ of measure $\mu(E) < \delta$, $\displaystyle\int_E |f_n| \, d\mu < \varepsilon, \ \forall n \in \mathbb{N}$.

Definitions 10.18.7 (tight sequence)**.** Let $(X, \mathcal{M}, \mu)$ be a measure space. A sequence of function $\{f_n\}$ is tight over X if for any $\varepsilon > 0$, there exists a subset of finite measure X_0 such that $\displaystyle\int_{X \sim X_0} |f_n| \, d\mu < \varepsilon, \ \forall n \in \mathbb{N}$.

Theorem 10.18.24. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let function f be integrable over X. Then for any $\varepsilon > 0$, there exists $\delta > 0$ such that for any subset E with $\mu(E) < \delta$, $\displaystyle\int_E |f| \, d\mu < \varepsilon$.*

Furthermore, for any $\varepsilon > 0$, there exists a subset of finite measure $X_0 \subset X$ such that $\displaystyle\int_{X \sim X_0} |f| \, d\mu < \varepsilon$.

Proof. Without loss of generality, suppose $0 \leq f$ on X. Otherwise, consider $f = f^+ - f^-$ and apply the following proof for those non-negative functions.

Let $\varepsilon > 0$. Since f is integrable, $\int_X f \, d\mu$ is finite. By the definition of integral of a non-negative functions, there exists a simple function ψ on X such that $0 \leq \psi \leq f$ on X and

$$0 \leq \int_X f \, d\mu - \int_X \psi \, d\mu < \varepsilon/2$$

Since ψ assume only finite number of values, we can choose $M > 0$ such that $0 \leq \psi < M$ on X.

$$\int_E f \, d\mu = \int_E \psi \, d\mu + \int_E (f - \psi) \, d\mu \leq M \cdot \mu(E) + \varepsilon/2$$

Choose $\delta = \varepsilon/2M$, then $\mu(E) < \varepsilon/2M$ and $\int_E f \, d\mu < \varepsilon$.

Let $X_0 = \{x \in X : f(x) > 0\}$. Then $f = f - \psi$ on $X \sim X_0$. Thus,

$$\int_{X \sim X_0} |f| \, d\mu = \int_{X \sim X_0} f \, d\mu = \int_{X \sim X_0} (f - \psi) \, d\mu \leq \int_X (f - \psi) \, d\mu < \varepsilon$$

□

Theorem 10.18.25 (Vitali Convergence theorem)**.** *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let $\{f_n\}$ be a sequence of functions which are uniformly integrable and tight over X. Suppose $\{f_n\}$ converges to function f pointwise a.e. on X and f is integrable over X. Then, $\displaystyle\lim_{n\to\infty} \int_E f_n \, d\mu = \int_E f \, d\mu$.*

Proof. By integral test, we have

$$\begin{aligned}\left|\int_X (f_n - f)\ d\mu\right| &\leq \int_X |f_n - f|\ d\mu \\ &\leq \int_{X_1} |f_n - f|\ d\mu + \int_{X_0\sim X_1} |f_n - f|\ d\mu + \int_{X\sim X_0} |f_n - f|\ d\mu \\ &\leq \int_{X_1} |f_n - f|\ d\mu + \int_{X_0\sim X_1} (|f_n| + |f|)\ d\mu + \int_{X\sim X_0} (|f_n| + |f|)\ d\mu\end{aligned}$$

Let $\varepsilon > 0$. Then since $\{f_n\}$ is tight and f is integrable, there exists $X_0 \subset X$ such that X_0 is a measurable subset of finite measure and

$$\int_{X\sim X_0} |f_n|\ d\mu + \int_{X\sim X_0} |f|\ d\mu < \varepsilon/3 \tag{10.89}$$

Since f is integrable over X, f is finite a.e. on X and $\mu(X_0) < \infty$. Thus by Egoroff's theorem, X_0 has a subset X_1 with $\mu(X_0 \sim X_1) < \delta$ such that $\{f_n\}$ is uniformly convergent on X_1. By uniform integrability of $\{f_n\}$, there exits $\delta > 0$ such that for any measurable set E with $\mu(E) < \delta$.

$$\int_{X_0\sim X_1} (|f_n| + |f|)\ d\mu = \int_{X_0\sim X_1} |f_n|\ d\mu + \int_{X_0\sim X_1} |f|\ d\mu < \varepsilon/3 \tag{10.90}$$

By uniform convergence of $\{f_n\}$ into f on X_1, there exists N such that for every $n \geq N$.

$$\int_{X_1} |f_n - f|\ d\mu \leq \sup_{x\in X_1} |f_n(x) - f(x)| \cdot \mu(X - 1) < \varepsilon/3 \tag{10.91}$$

□

Corollary 10.18.25.1. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let $\{h_n\}$ be sequence on non-negative, integrable functions over X. Suppose $\{h_n\}$ converges to 0 for almost all $x \in X$. Then $\lim_{n\to\infty} \int_E h_n\ d\mu = 0$ if and only if $\{h_n\}$ is uniformly integrable and tight over X.*

Proof. Since $\{h_n\}$ is uniformly convergent to 0 almost everywhere on X and is tight, from theorem we have

$$\lim_{n\to\infty} \int_X h_n\ d\mu = 0$$

Suppose $\lim_{n\to\infty} \int_X h_n\ d\mu = 0$. Then by Fatou's lemma,

$$\int_X \lim_{n\to\infty} h_n\ d\mu < \liminf \int_X h_n(x)\ d\mu = 0$$

□

Exercises

27.

28.

29.

30.

31.

32.

33.

34.

35.

36.

37.

38.

39.

40.

41.

42.

43.

44.

45.

46.

47.

48.

10.18.4 Radon-Nikodym Theorem

Definitions 10.18.8 (absolutely continuous)**.** Let $(X, \mathcal{M})$ be a measure space. Let μ, ν be measures on X. The measure ν is absolutely continuous with respect to μ if $\nu(E) = 0$ whenever $\mu(E) = 0$ for any measure subset $E \in \mathcal{M}$. And we write $\nu \ll \mu$.

For example : The set function $\nu(E) = \displaystyle\int_E f \, d\mu$ is a measure on $(X, \mathcal{M})$ and is absolutely continuous with respect to μ.

Theorem 10.18.26. *Let $(X, \mathcal{M}, \mu)$ be a measure space. Let ν be a measure on $(X, \mathcal{M})$. Then ν is absolutely continuous with respect to μ if and only if for any $\varepsilon > 0$, there exists $\delta > 0$ such that for any measurable subset $E \in \mathcal{M}$ with $\mu(E) < \delta$, $\nu(E) < \varepsilon$.*

Proof. Suppose for $\varepsilon > 0$, there exists $\delta > 0$ such that $\nu(E) < \varepsilon$ if $\mu(E) < \delta$. Since measure has monotonicity, as $\varepsilon \to 0$, $\delta \to 0$. Therefore, ν is absolutely continuous with respect to μ.

Let $\varepsilon > 0$. Suppose that for any $\delta > 0$, there exists some measurable set E with $\mu(E) < \delta$, but $\nu(E) \geq \varepsilon$. Then, we can construct a sequence of measurable sets, $\{E_n\}$ such that $\mu(E_n) \leq \frac{1}{2^n}$ and $\nu(E_n) \geq \varepsilon$. Define a descencding sequence $\{A_n\}$ where $A_n = \bigcup_{k=n}^{\infty} E_k$. By the continuity of measure, $\lim_{n\to\infty} \mu(A_n) = \mu\left(\bigcap_{n=1}^{\infty} A_n\right) = 0$ which is a contradiction to absolute continuity since $\lim_{n\to\infty} \nu(A_n) \geq \varepsilon$. Therefore, if $\varepsilon > 0$, then there exists $\delta > 0$ such that $\mu(E) < \delta \implies \nu(E) < \varepsilon$. □

Theorem 10.18.27 (Radon-Nikodym)**.** *Let $(X, \mathcal{M}, \mu)$ be a σ-finite measure space. If there exists a σ-finite measure ν on $(X, \mathcal{M})$ which is absolutely continuous with respect to μ. Then there exists a non-negative function f such that f is measurable (with respect to $\mathcal{M}$) and*

$$\nu(E) = \int_E f \, d\mu, \ \forall E \in \mathcal{M} \tag{10.92}$$

And this function f is unique with respect to μ. That is, for any other non-negative, measurable[μ] function g defining ν, $f = g$ a.e.[μ] on X.

Proof. Suppose μ, ν are finite measures. If $\nu(E) = 0$ for any $E \in \mathcal{M}$, then the equation is true for $f = 0$ on X.

Suppose $\nu(E) \neq 0$ for some $E \in \mathcal{M}$. First, we prove that there exists a non-negative function f on X such that

$$\int_X f \, d\mu > 0 \text{ and } \int_E f \, d\mu \leq \nu(E), \ \forall E \in \mathcal{M}$$

Then we prove that the supremum of all such function is the unique function satisfying our requirements.

Step 1 : Existence of f
Let $\lambda > 0$. Then $\nu - \lambda\mu$ is difference of two finite measures and thus a signed measure. According to Hahn decomposition therem, there exists a Hahn decomposition $\{P_\lambda, N_\lambda\}$ of signed measure $\nu - \lambda\mu$. Thus, for each λ, we have $X = P_\lambda \cup N_\lambda$ and $P_\lambda \cap N_\lambda = \phi$. Also the sets P_λ, N_λ are positive and negative with respect to $\nu - \lambda\mu$ respectively.

We claim that there exists $\lambda > 0$ such that $\mu(P_\lambda) > 0$.
Suppose that for any $\lambda > 0$, $\mu(P_\lambda) = 0$. Then by monotonicity, $\mu(E) = 0$ for any

measurable subset $E \subset P_\lambda$. We have P_λ is a positive set with respect to $\nu - \lambda\mu$. Also ν is absolutely continuous with respect to μ. Thus, we have $\nu(E) = 0$. And since N_λ is negative with respect to $\nu - \lambda\mu$, we have $\nu(E) \le \lambda\mu(E)$. Clearly, every measurable subset E is of the form $E = E_P \cup E_N$ where $E_P \subset P_\lambda$ and $E_N \subset N_\lambda$. And $\nu(E) = \nu(E_P) + \nu(E_N) \le 0 + \lambda\mu(E_N) \le \lambda\mu(E)$. Thus, $\nu \le \lambda\mu$ for every measurable subset $E \in \mathcal{M}$ and for any real-number $\lambda > 0$. Clearly, $\mu(E) = 0 \implies \nu(E) = 0$ and $\mu(E) < \infty \implies \nu(E) = 0$ by archimedian property. This is a contradiction, since we assumed that $\nu(E) \ne 0$ for some $E \in \mathcal{M}$.

Let $\lambda_0 > 0$ such that $\mu(P_{\lambda_0}) > 0$. Define $f = \lambda_0 \cdot \chi_{P_{\lambda_0}}$. Then, we have $\displaystyle\int_X f\ d\mu = \lambda_0 \cdot \mu(P_{\lambda_0}) > 0$. And for any measurable subset $E \in \mathcal{M}$, we have

$$\int_E f\ d\mu = \lambda_0 \cdot \mu(P_{\lambda_0} \cap E) \le \nu(P_{\lambda_0} \cap E) \le \nu(E)$$

And by construction, f is a non-negative, measurable function on X.

Step 2 : Construction of f

Let $\mathcal{F}$ be the collection of all such non-negative, measurable functions on X. Define $M = \sup\limits_{f \in \mathcal{F}} \left\{ \displaystyle\int_X f\ d\mu \right\}$. Suppose $g, h \in \mathcal{F}$, then $\max\{g, h\} \in \mathcal{F}$. And,

$$\int_E \max\{g, h\}\ d\mu = \int_{E_1} h\ d\mu + \int_{E_2} g\ d\mu$$

where $E_1 = \{x \in X : h(x) > g(x)\}$ and $E_2 = \{x \in X : h(x) \le g(x)\}$.

Consider a sequence $\{f_n\}$ in $\mathcal{F}$ such that $\displaystyle\lim_{n\to\infty} \int_X f_n\ d\mu = M$. Without loss of generality, this is an increasing sequence. If the sequence is not increasing, then replace f_k with $\displaystyle\sum_{n=1}^{k} f_n$.

Define $f(x) = \displaystyle\lim_{n\to\infty} \int_X f_n\ d\mu$. By monotone convergence theorem, we have $\int_X f\ d\mu = M$. Define $\eta(E) = \nu(E) - \int_E f\ d\mu$. Since integral is coutably additive and $\nu(X) < \infty$, η is a signed measure. Also η is absolutely continuous with respect to ν by its construction.

We claim that $\eta = 0$ on $\mathcal{M}$.

Suppose $\eta \ne 0$. Then by performing Step 1, 2 for η instead of ν, we get a non-negative measurable function $\hat{f}$ such that

$$\int_X \hat{f}\ d\mu > 0 \text{ and } \int_E \hat{f}\ d\mu \le \eta(E) = \nu(E) - \int_E f\ d\mu$$

Clearly $f + \hat{f} \in \mathcal{F}$ and $\displaystyle\int_X (f + \hat{f})\ d\mu > \int_X f\ d\mu = M$ which is a contradiction.

Step 3 : Uniqueness of f

Suppose f_1, f_2 are two limit function on $\mathcal{F}$ satisfying the conditions. Then, for every measurable subset $E \in \mathcal{M}$, we have $s \int_E (f_1 - f_2)\ d\mu = 0$ and $f_1 = f_2$ a.e.$[\mu]$ on X. Therefore, the choice of f is unique upto measure μ.

Step 4 : Extending the proof to σ-finite measures
Suppose μ, ν are σ-finite measures. Then from the defintion of σ-finite measure, X has a countable partition, $X = \cup_{n=1}^{\infty} X_n$ such that $\mu(X_n) < \infty$ and $\nu(X_n) < \infty$ for every n. Now we may represent σ-finite measures as countable sum of finite measures, μ_n, ν_n on X_n.
That is, $\mu = \sum_{n=1}^{\infty} \mu_n \cdot \chi_{X_n}$ and $\nu = \sum_{n=1}^{\infty} \nu_n \cdot \chi_{X_n}$
Now $\nu - \lambda\mu = \sum_{n=1}^{\infty} (\nu_n - \lambda\mu_n) \cdot \chi_{X_n}$ is a σ-finite signed measure. From Step 1,2 and 3, we have a non-negative measurable function f such that integral of f is the supremum of integrals of all the non-negative functions dominated by ν. Then,

$$\int_X f\ d\mu = \sum_{n=1}^{\infty} \int_{X_n} f\ d\mu_n > 0$$

And

$$\int_E f\ d\mu = \sum_{n=1}^{\infty} \int_{E \cap X_n} f\ d\mu_n < \sum_{n=1}^{\infty} \nu_n(E \cap X_n) = \nu(E)$$

□

Theorem 10.18.28. *Let $(X, \mathcal{M}, \mu)$ be a σ-finite measure space. If there exists a finite signed measure ν which is absolutely continuous with respect to μ on X. Then there exist a function f which is integrable over X with respect to μ and*

$$\nu(E) = \int_E f\ d\mu,\ \forall E \in \mathcal{M} \tag{10.93}$$

Proof. □

Theorem 10.18.29 (Lebesgue Decomposition)**.** *Let $(X, \mathcal{M}, \mu)$ be a σ-finite measure space. Let ν be a σ-finite measure on $(X, \mathcal{M})$. Then there exists unique measures ν_0, ν_1 such that $\nu = \nu_0 + \nu_1$, ν_0 is singlular with respect to μ and ν_1 is absolutely continuous with respect to μ.*

Proof. □

Definitions 10.18.9 (Radon-Nikodym derivative)**.** Let $(X, \mathcal{M})$ be a measure space. Let μ, ν be two σ-finite measures on $(X, \mathcal{M})$ such that ν is absolutely continuous with respect to μ. By, Radon-Nikodym theorem, there exists a measurable function f which is unique upto μ such that $\nu = \int_E f\ d\mu$. This function f is the Radon-Nikodym derivative $f = \frac{d\nu}{d\mu}$.

Exercises

49.

50.

51.

52.

53.

54.

55.

56.

57.

58.

59.

60.

10.20 The Construction of a particular measure

10.20.1 Product measures

Definitions 10.20.1 (measurable rectangle)**.** Let $(X, \mathcal{A}, \mu)$, $(Y, \mathcal{B}, \nu)$ be two measure spaces. A rectangle is a subset of $X \times Y$ of the form $A \times B$ where $A \subset X$ and $B \subset Y$. A measurable rectangle is a subset of $X \times Y$ of the form $A \times B$ where $A \in \mathcal{A}$ and $B \in \mathcal{B}$.

Lemma 10.20.1. *Let $\{A_k \times B_k\}_{k=1}^{\infty}$ be a countable disjoint, collection of measurable rectangles such that $\bigcup_{k=1}^{\infty} (A_k \times B_k) = A \times B$ is also a meausurable rectangle. Then $\mu(A) \times \nu(B) = \sum_{k=1}^{\infty} \mu(A_k) \times \mu(B_k)$*

Proof. □

Theorem 10.20.2. *Let $(X, \mathcal{A}, \mu)$, $(Y, \mathcal{B}, \nu)$ be two measure spaces. Let $\mathcal{R}$ be the collection of all measurable renctangles in $X \times Y$. Let $\lambda : \mathcal{R} \to [0, \infty]$ given by $\lambda(A \times B) = \mu(A) \cdot \nu(B)$. Then $\langle \mathcal{R}, \lambda \rangle$ is a semiring and λ is a premeasure.*

Proof. □

Definitions 10.20.2. Let $(X, \mathcal{A}, \mu)$, $(Y, \mathcal{B}, \nu)$ be two measure spaces. Let $\mathcal{R}$ be the collection of measurable rectangles in $X \times Y$ and λ is the premeasure defined by $\lambda : \mathbb{R} \to [0, \infty]$, $\lambda(A \times B) = \mu(A) \cdot \nu(B)$. The product measure $\lambda = \mu \times \nu$ is the Caratheodory extension of $\lambda : \mathbb{R} \to [0, \infty]$ defined by the σ-algebra of $(\mu \times \nu)^*$-measurable subsets of $X \times Y$.

Definitions 10.20.3 (x-section)**.** Let E be a subset of $X \times Y$ and f be a function on E. Let $x \in X$. Then the x-section of f, $E_x = \{y \in Y : (x, y) \in E\}$.

Fubini's Theorem

Lemma 10.20.3. *Let $E \subset X \times Y$ be an $\mathcal{R}_{\sigma\delta}$ set for which $(\mu \times \nu)(E) < \infty$. Then for every $x \in X$, the x-section of E, E_x is a ν-measurable subset of Y, the function which maps each $x \in X$ into $\nu(E_x)$ is a μ-measurable function and*

$$(\mu \times \nu)(E) = \int_X \nu(E_x)\ d\mu(x) \tag{10.94}$$

Proof. □

Lemma 10.20.4. *Suppose measure ν is complete. Let $E \subset X \times Y$ be measurable with respect to $\mu \times \nu$. If $(\mu \times \nu)(E) = 0$, then for almost all $x \in X$, the x-section of E, E_x is ν-measurable and $\nu(E_x) = 0$. Therefore,*

$$(\mu \times \nu)(E) = \int_X \nu(E_x)\ d\mu(x) \tag{10.95}$$

Proof. □

Theorem 10.20.5. *Suppose measure ν is complete. et $E \subset X \times Y$ be measurable with respect to $\mu \times \nu$ and $(\mu \times \nu)(E) < \infty$. Then for almost all $x \in X$, the x-section of E, E_x is a ν-measurable subset of Y, the function which maps each $x \in X$ into $\nu(E_x)$ is a μ-measurable function and*

$$(\mu \times \nu)(E) = \int_X \nu(E_x)\ d\mu(x) \tag{10.96}$$

Proof. □

Theorem 10.20.6. *Suppose measure ν is complete. Let $\varphi : X \times Y \to \mathbb{R}$ be asimple functions that is integrable over $X \times Y$ with respect to $\mu \times \nu$. Then for almost all $x \in X$, the x-section of φ, $\varphi(x, \cdot)$ is integrable over Y with respect to ν and*

$$\int_{X \times Y} \varphi\ d(\mu \times \nu) = \int_X \left(\int_Y \varphi(x, y)\ d\nu(y) \right)\ d\mu(x) \tag{10.97}$$

Proof. □

Theorem 10.20.7 (Fubini)**.** *Let $(X, \mathcal{A}, \mu)$, $(Y, \mathcal{B}, \nu)$ be two measure spaces where ν is complete. Let f be integrable over $X \times Y$ with respect to the product measure $\mu \times \nu$. Then for almost all $x \in X$, the x-section of f is integrable over Y with respect to ν and*

$$\int_{X \times Y} f\ d(\mu \times \nu) = \int_X \left(\int_Y f(x, y)\ d\nu(y) \right) d\mu(x) \tag{10.98}$$

Proof. □

Tonelli's Theorem

Theorem 10.20.8 (Tonelli). *Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be two σ-finite measure spaces where ν is complete. Let f be a non-negative $(\mu \times \nu)$-measurable function on $X \times Y$. Then for almost all $x \in X$, the x-section of f, $f(x, \cdot)$ is ν-measurable and the function which maps almost every $x \in X$ into integral of $f(x, \cdot)$ over Y with respect to ν is μ-measurable. Moreover,*

$$\int_{X \times Y} f \; d(\mu \times \nu) = \int_X \left(\int_Y f(x, y) \; d\nu(y) \right) \; d\mu(x) \tag{10.99}$$

Proof. □

Corollary 10.20.8.1. *Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be two σ-finite, complete measure spaces. Let f be a non-negative $(\mu \times \nu)$-measurable function on $X \times Y$. Then*

1. *for almost all $x \in X$, the x-section of f, $f(x, \cdot)$ is ν-measurable function and the function defined almost everywhere on X by $x \to$ the integral of $f(x, \cdot)$ over Y with respect to ν is μ-measurable and*

2. *for almost all $y \in Y$, the y-section of f, $f(\cdot, y)$ is μ-measurable and the function defined almost everywhere on Y by $y \to$ the integral of $f(\cdot, y)$ over X with respect to μ is ν-measurable if*

$$\int_X \left(\int_Y f(x, y) \; d\nu(y) \right) \; d\mu(x) < \infty \tag{10.100}$$

then f is integrable over $X \times Y$ with respect to $\mu \times \nu$ and

$$\begin{aligned} \int_Y \left(\int_X f(x, y) \; d\mu(x) \right) \; d\nu(y) &= \int_{X \times Y} f d(\mu \times \nu) \qquad (10.101) \\ &= \int_X \left(\int_Y f(x, y) \; d\nu(y) \right) \; d\mu(x) \end{aligned}$$

Proof. □

Exercises

Definitions 10.20.4 (probability measure space). A probability measure space is a measure space $(X, \mathcal{M}, \mu)$ where $\mu(X) = 1$.

1.
2.
3.
4.
5.
6.
7.

8.

9.

10.

11.

12.

13.

14.

Semester III

Subject 11

ME010301 Advanced Complex Analysis

Subject 12

ME010302 Partial Differential Equations

Subject 13

ME010303 Multivariate Calculus & Integral Transforms

13.1 Integral Transforms

13.1.1 The Weierstrass Approximation Theorem

Every continuous, real valued function on a compact interval has a polynomial approximation. [Apostol, 1973, Theorem 11.17]

Theorem 13.1.1 (Weierstrass)**.** *Let f be a real-valued, continuous function on a compact interval $[a, b]$. Then for every $\epsilon > 0$, there is a polynomial p such that $|f(x) - p(x)| < \epsilon$ for every $x \in [a, b]$.*

Synopsis. Given a real-valued continuous function on compact interval $[a, b]$, we can construct a real-valued, continous function g on $\mathbb{R}$ which is periodic with period 2π. We have, if $f \in L(I)$ and f is bounded almost everywhere in I, then $f \in L^2(I)$. [Apostol, 1973, Theorem 10.52]. By Fejer's theorem ([Apostol, 1973, Theorem 11.15]), the fourier series generated by g ([Apostol, 1973, definition 11.3]) converges to the Cesaro sum ([Apostol, 1973, Definition 8.47]), which is g itself in this case. Thus for any $\epsilon > 0$, there is a finite sum of trignometric functions. The power series expansions of trignometric functions ([Apostol, 1973, definition 9.27]) being uniformly convergent, there exists a polynomial p_m which approximates g. And we can construct p (polynomial approximation of g) using p_m.

Proof. Define $g : \mathbb{R} \to \mathbb{R}$,

$$g(t) = \begin{cases} f(a + (b - a)t/\pi), \ t \in [0, \pi) \\ f(a + (2\pi - t)(b - a)/\pi), \ t \in [\pi, 2\pi] \\ g(t - 2n\pi), \ t > 2\pi, \ n \in \mathbb{N} \\ g(t + 2n\pi), \ t < 0, \ n \in \mathbb{N} \end{cases}$$

Thus g is a continuous, real-valued, periodic function with period 2π such that

$$f(x) = g\left(\frac{\pi(x-a)}{b-a}\right), \ x \in [a,b] \tag{13.1}$$

The fourier series generated by g is given by,

$$g(t) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kt + b_k \sin kt)$$

$$\text{where } a_k = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos kt \ dt, \ b_k = \frac{1}{\pi}\int_0^{2\pi} f(t)\sin kt \ dt$$

Let $\{s_n(t)\}$ be the sequence of partial sums of the fourier series generated by g. And $\{\sigma_n(t)\}$ be the sequence of averages of $s_n(t)$ given by,

$$\sigma_n(t) = \frac{1}{n}\sum_{k=1}^{n} s_k(t), \ \text{where } s_k(t) = \frac{a_0}{2} + \sum_{j=1}^{k}(a_j \cos jt + b_j \sin jt)$$

Function $f \in L(I)$ being real-valued continuous function on a compact interval, it is bounded and hence is Lebesgue square integrable. ie, $f \in L^2(I)$. Thus, $g \in L^2(I)$.

Since g is continous on $\mathbb{R}$, the function $s : \mathbb{R} \to \mathbb{R}$ defined by,

$$s(t) = \lim_{h\to 0^+} \frac{g(t+h) - g(t-h)}{2}$$

is well-defined on $\mathbb{R}$ and $s(t) = g(t), \ \forall t \in \mathbb{R}$.

Then by Fejer's Theorem, the sequence $\{\sigma_n(t)\}$ converges uniformly to $g(t)$ for every $t \in \mathbb{R}$. Thus, given $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $\forall t \in \mathbb{R}$, $|g(t) - \sigma_N(t)| < \frac{\epsilon}{2}$.

We have,

$$\sigma_N(t) = \sum_{k=0}^{N}(A_k \cos kt + B_k \sin kt), \ \text{where } A_k, B_k \in \mathbb{R} \tag{13.2}$$

By the power series expansion of the trignometric functions about origin,

$$\cos kt = \sum_{j=1}^{\infty}\left(\frac{\cos^{(j)} 0}{j!}(kt)^j\right) = \sum_{j=1}^{\infty} A'_j t^j \text{ where } A'_j \in \mathbb{R} \tag{13.3}$$

$$\sin kt = \sum_{j=1}^{\infty}\left(\frac{\sin^{(j)} 0}{j!}(kt)^j\right) = \sum_{j=1}^{\infty} B'_j t^j \text{ where } B'_j \in \mathbb{R} \tag{13.4}$$

Since the above power series expansions of trignometric functions are uniformly convergent, their finite linear combination $\{\sigma_N(t)\}$ is also uniformly convergent. ie, Given $\epsilon > 0$ there exists $m \in \mathbb{N}$ such that for every $t \in \mathbb{R}$

$$\left|\sum_{k=0}^{m} C_k t^k - \sigma_N(t)\right| < \frac{\epsilon}{2} \text{ where } C_k \in \mathbb{R}$$

Therefore, $|p_m(t)-g(t)| \leq |p_m(t)-\sigma_N(t)|+|\sigma_N(t)-g(t)| < \epsilon$ where $p_m(t) = \sum_{k=0}^{m} C_k t^k$. Define $p : [a,b] \to \mathbb{R}$ by,

$$p(x) = p_m\left(\frac{\pi(x-a)}{b-a}\right) \tag{13.5}$$

By equations 13.1 and 13.5, $|p(x)-f(x)| < \epsilon$ for every $x \in [a,b]$. □

13.1.2 Other Forms of Fourier Series

Let $f \in L([0,2\pi])$, then the fourier series generated by f is given by,

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$$

$$\text{where } a_n = \frac{1}{\pi}\int_0^{2\pi} f(t)\cos nt\ dt, \qquad b_n = \frac{1}{\pi}\int_0^{2\pi} f(t)\sin nt\ dt$$

By Euler's forumula $e^{inx} = \cos nx + i\sin nx$. We have, $\cos nx = \frac{(e^{inx}+e^{-inx})}{2}$ and $\sin nx = \frac{(e^{inx}-e^{-inx})}{2i}$

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(\alpha_n e^{inx} + \beta_n e^{-inx}\right)$$

$$\text{where } \alpha_n = \frac{(a_n - ib_n)}{2} \qquad \beta_n = \frac{(a_n + ib_n)}{2}$$

Therefore, by assigning $\alpha_0 = a_0/2$, $\alpha_{-n} = \beta_n$, we get the following exponential form of fourier series generated by f,

$$f(x) \sim \sum_{n=-\infty}^{\infty} \alpha_n e^{inx} \text{ where } \alpha_n = \frac{1}{2\pi}\int_0^{2\pi} f(t)\ e^{-int}\ dt$$

Note : If f is periodic with period 2π, then the interval of integration $[0,2\pi]$ can be replaced with any interval of length 2π. eg. $[-\pi,\pi]$

Periodic with period p

Let $f \in L([0,p])$ and f is periodic with period p. Then

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n \cos\frac{2\pi nx}{p} + b_n \sin\frac{2\pi nx}{p}\right)$$

$$\text{where } a_n = \frac{2}{p}\int_0^{p} f(t)\cos\frac{2\pi nt}{p}\ dt \qquad b_n = \frac{2}{p}\int_0^{p} f(t)\sin\frac{2\pi nt}{p}\ dt$$

Therefore, we have the exponential form of the above fourier series given by,

$$f(x) \sim \sum_{n=-\infty}^{\infty} \alpha_n e^{\frac{2\pi inx}{p}}, \text{ where } \alpha_n = \frac{1}{p}\int_0^{p} f(t)\ e^{\frac{-2\pi int}{p}}\ dt$$

13.1.3 Fourier Integral Theorem

Theorem 13.1.2 (Fourier Integral Theorem). *Let $f \in L(-\infty, \infty)$. Suppose $x \in \mathbb{R}$ and an interval $[x-\delta, x+\delta]$ about x such that either*

1. *f is of bounded variation on an interval $[x-\delta, x+\delta]$ about x or*
2. *both limits $f(x+)$ and $f(x-)$ exists and both Lebesgue intergrals*

$$\int_0^\delta \frac{f(x+t)-f(x+)}{t} dt \text{ and } \int_0^\delta \frac{f(x-t)-f(x-)}{t} dt$$

   exists.

Then,

$$\frac{f(x+)+f(x-)}{2} = \frac{1}{\pi} \int_0^\infty \int_{-\infty}^\infty f(u) \cos v(u-x)\ du\ dv,$$

the integral $\int_0^\infty$ being an improper Riemann integral.

Synopsis.

$$f(x+t)\frac{\sin \alpha t}{\pi t} dt \to f(u)\frac{\sin \alpha(u-x)}{\pi(u-x)} \to \frac{f(u)}{\pi} \int_0^\alpha \cos v(u-x) dv$$

By Riemann-Lebesgue lemma[Apostol, 1973, Theorem 11.6],

$$f \in L(I) \implies \lim_{\alpha \to +\infty} \int_I f(x) \sin \alpha t\ dt = 0$$

By Jordan's Theorem[Apostol, 1973, Theorem 10.8], if g is of bounded variation on $[0, \delta]$, then

$$\lim_{\alpha \to +\infty} \frac{2}{\pi} \int_0^\delta g(t) \frac{\sin \alpha t}{t} dt = g(0+)$$

By Dini's Theorem[Apostol, 1973, Theorem 10.9], if the limit $g(x+)$ exists and Lebesgue integral $\int_0^\delta \frac{g(t)+g(0+)}{t} dt$ exists for some $\delta > 0$, then

$$\lim_{\alpha \to +\infty} \frac{2}{\pi} \int_0^\delta g(t) \frac{\sin \alpha t}{t} dt = g(0+)$$

The order of Lebesgue integrals can be interchanged.[Apostol, 1973, Theorem 10.40]

Suppose $f \in L(X)$ and $g \in L(Y)$. Then

$$\int_X f(x) \left(\int_Y g(y) k(x,y) dy \right) dx = \int_Y g(y) \left(\int_X f(x) k(x,y) dx \right) dy$$

Proof. Consider $\int_{-\infty}^\infty f(x+t) \frac{\sin \alpha t}{\pi t} dt$. We prove that this integral is equal to the either sides.

$$\int_{-\infty}^\infty f(x+t) \frac{\sin \alpha t}{\pi t} dt = \int_{-\infty}^{-\delta} + \int_{-\delta}^0 + \int_0^{-\delta} + \int_\delta^\infty f(x+t) \frac{\sin \alpha t}{\pi t} dt$$

We have, function $\frac{f(x+t)}{\pi t}$ is bounded on $(-\infty, -\delta) \cup (\delta, \infty)$, hence $\frac{f(x+t)}{\pi t}$ is Lebesgue integrable on $(-\infty, -\delta) \cup (\delta, \infty)$.

By Riemann Lebesgue lemma,

$$\frac{f(x+t)}{\pi t} \in L(-\infty,-\delta) \implies \int_{-\infty}^{-\delta} f(x+t)\frac{\sin \alpha t}{\pi t}dt = 0,$$

$$\frac{f(x+t)}{\pi t} \in L(\delta,\infty) \implies \int_{\delta}^{\infty} f(x+t)\frac{\sin \alpha t}{\pi t}dt = 0$$

Case 1 Suppose f is of bounded variation on $[x-\delta, x+\delta]$, put $g(t) = f(x+t)$ then g is of bounded variation on $[-\delta, \delta]$. Thus g is of bounded variation on $[0, \delta]$. Then by Jordan's Theorem

$$\lim_{\alpha\to+\infty} \frac{2}{\pi}\int_0^{\delta} f(x+t)\frac{\sin \alpha t}{t}dt = \lim_{\alpha\to+\infty} \frac{2}{\pi}\int_0^{\delta} g(t)\frac{\sin \alpha t}{t}dt = g(0+) = f(x+)$$

Case 2 Suppose both the limits $f(x+)$ and $f(x-)$ exists and both Lebesgue integrals

$$\int_0^{\delta} \frac{f(x+t)-f(x+)}{t}dt \text{ and } \int_0^{\delta} \frac{f(x-t)-f(x-)}{t}dt$$

exists.

Thus, we have $f(x+)$ exists and the Lebesgue integral $\int_0^{\delta} \frac{f(x+t)-f(x+)}{t}dt$ exists. Put $g(t) = f(x+t)$, then $g(0+) = f(x+)$ exists and the Lebesgue integral $\int_0^{\delta} \frac{g(t)-g(0+)}{t}dt$ exists, then by Dini's Theorem,

$$\lim_{\alpha\to+\infty} \frac{2}{\pi}\int_0^{\delta} f(x+t)\frac{\sin \alpha t}{t}dt = \lim_{\alpha\to+\infty} \frac{2}{\pi}\int_0^{\delta} g(t)\frac{\sin \alpha t}{t}dt = g(0+) = f(x+)$$

Similarly, $f(x-)$ exists and the Lebesgue integral $\int_0^{\delta} \frac{f(x-t)-f(x-)}{t}dt$ exists. Put $g(t) = f(x-t)$, then $g(0+) = f(x-)$ exists and the Lebesgue integral $\int_0^{\delta} \frac{g(t)-g(0+)}{t}dt$ exists, then by Dini's Theorem,

$$\begin{aligned}\lim_{\alpha\to+\infty} \frac{2}{\pi}\int_{-\delta}^{0} f(x+t)\frac{\sin \alpha t}{t}dt &= \lim_{\alpha\to+\infty} \frac{2}{\pi}\int_0^{\delta} f(x-\tau)\frac{\sin \alpha \tau}{\tau}d\tau \\ &= \lim_{\alpha\to+\infty} \frac{2}{\pi}\int_0^{\delta} g(\tau)\frac{\sin \alpha \tau}{\tau}d\tau = g(0+) = f(x-)\end{aligned}$$

Then by either cases,

$$\begin{aligned}\lim_{\alpha\to+\infty} \int_{-\infty}^{\infty} f(x+t)\frac{\sin \alpha t}{\pi t}dt &= \lim_{\alpha\to+\infty} \int_{-\delta}^{0} + \int_0^{\delta} f(x+t)\frac{\sin \alpha t}{\pi t}dt \\ &= \frac{f(x+)+f(x-)}{2}\end{aligned}$$

We have, $\int_0^\alpha \cos v(u-x)dv = \frac{\sin v(u-x)}{u-x}$.

$$\begin{aligned}
\lim_{\alpha\to+\infty}\int_{-\infty}^{\infty} f(x)\frac{\sin\alpha t}{\pi t}dt &= \lim_{\alpha\to+\infty}\int_{-\infty}^{\infty} f(u)\frac{\sin\alpha(u-x)}{u-x}du, \text{ (put } u = x+t)\\
&= \lim_{\alpha\to+\infty}\int_{-\infty}^{\infty} f(u)\left(\int_0^\alpha \cos v(u-x)dv\right)du\\
&= \lim_{\alpha\to+\infty}\int_0^{\alpha}\left(\int_{-\infty}^{\infty} f(u)\cos v(u-x)du\right)dv,
\end{aligned}$$

since, the order of Lebesgue integrals can be reversed.

$$= \int_0^{\infty}\left(\int_{-\infty}^{\infty} f(u)\cos v(u-x)du\right)dv$$

where, $\displaystyle\int_0^\infty$ is not a Lebesgue integral, but an improper Riemann integral

Therefore,

$$\begin{aligned}
\int_0^{\infty}\left(\int_{-\infty}^{\infty} f(u)\cos v(u-x)du\right)dv &= \lim_{\alpha\to+\infty}\int_{-\infty}^{\infty} f(x)\frac{\sin\alpha t}{\pi t}dt\\
&= \frac{f(x+)+f(x-)}{2}
\end{aligned}$$

□

Remark. If a function f on $(-\infty,\infty)$ is non-periodic, then it may not have a fourier series represenation. In such cases, we have fourier intergral represen-taion.

13.1.4 Exponential form of Fourier Integral Theorem

Let $f \in L(-\infty,\infty)$. Suppose $x \in \mathbb{R}$ and an interval $[x-\delta, x+\delta]$ about x such that either

1. f is of bounded variation on an interval $[x-\delta, x+\delta]$ about x or
2. both limits $f(x+)$ and $f(x-)$ exists and both Lebesgue intergrals

$$\int_0^\delta \frac{f(x+t)-f(x+)}{t}dt \text{ and } \int_0^\delta \frac{f(x-t)-f(x-)}{t}dt$$

exists.

Then,

$$\frac{f(x+)+f(x-)}{2} = \lim_{\alpha\to\infty}\frac{1}{2\pi}\int_{-\alpha}^{\alpha}\left(\int_{-\infty}^{\infty} f(u)e^{iv(u-x)}\,du\right)dv$$

Proof. Let $F(v) = \int_{-\infty}^{\infty} f(u)\cos v(u-x)du$. Then $F(v) = F(-v)$ and

$$\begin{aligned}
\lim_{\alpha\to\infty}\frac{1}{2\pi}\int_{-\alpha}^{\alpha} F(v)dv &= \lim_{\alpha\to\infty}\frac{1}{\pi}\int_0^{\alpha}\int_{-\infty}^{\infty} f(u)\cos v(u-x)dudv\\
&= \frac{f(x+)+f(x-)}{2}
\end{aligned}$$

Let $G(v) = \int_{-\infty}^{\infty} f(u) \sin v(u-x) du$. Then $G(v) = -G(-v)$ and

$$\lim_{\alpha \to \infty} \frac{1}{2\pi} \int_{-\alpha}^{\alpha} G(v) dv = 0$$

Thus

$$\lim_{\alpha \to \infty} \frac{1}{2\pi} \int_{-\alpha}^{\alpha} F(v) + iG(v) dv = \frac{f(x+) + f(x-)}{2}$$

□

13.1.5 Integral Transforms

Definitions 13.1.1. Integral transform $g(y)$ of $f(x)$ is a Lebesgue integral or Improper Riemann integral of the form

$$g(y) = \int_{-\infty}^{\infty} K(x,y) f(x)\ dx$$

, where K is the kernal of the transform. We write $g = \mathscr{K}(f)$.

Remark. Integral transforms(operators) are linear operators. ie, $\mathscr{K}(af_1 + bf_2) = a\mathscr{K} f_1 + b\mathscr{K} f_2$

Remark. A few commonly used integral transforms,

1. Exponential Fourier Transform $\mathscr{F}$,
$$\mathscr{F} f = \int_{-\infty}^{\infty} e^{-ixy} f(x)\ dx$$

2. Fourier Cosine Transform $\mathscr{C}$,
$$\mathscr{C} f = \int_{0}^{\infty} \cos xy f(x)\ dx$$

3. Fourier Sine Transform $\mathscr{S}$,
$$\mathscr{S} f = \int_{0}^{\infty} \sin xy f(x)\ dx$$

4. Laplace Transform $\mathscr{L}$,
$$\mathscr{L} f = \int_{0}^{\infty} e^{-xy} f(x)\ dx$$

5. Mellin Transform $\mathscr{M}$,
$$\mathscr{M} f = \int_{0}^{\infty} x^{y-1} f(x)\ dx$$

Remark. Suppose $f(x) = 0,\ \forall x < 0$.

$$\int_{-\infty}^{\infty} e^{-ixy} f(x)\ dx = \int_{0}^{\infty} e^{-ixy} f(x)\ dx = \int_{0}^{\infty} \cos xy\ f(x)\ dx + i \int_{0}^{\infty} \sin xy\ f(x)\ dx$$

$$\mathscr{F} f = \mathscr{C} f + i \mathscr{S} f$$

Therefore Fourier Cosine $\mathscr{C}$ and Sine $\mathscr{S}$ transforms are special cases of fourier integral transform, $\mathscr{F}$ provided f vanishes on negative real axis.

Remark. Let $y = u + iv$, $f(x) = 0$, $\forall x < 0$.

$$\int_0^\infty e^{-xy} f(x) = \int_0^\infty e^{-xu} e^{-ixv} f(x)\ dx = \int_0^\infty e^{-ixv} \phi_u(x) dx$$

where $\phi_u(x) = e^{-xu} f(x)$.

$$\mathscr{L} f = \mathscr{F} \phi_u$$

Therefore Laplace transform, $\mathscr{L}$ is a special case of Fourier integral transform, $\mathscr{F}$.

Remark. Let $g(y) = \mathscr{F} f(x)$.

$$g(y) = \int_{-\infty}^\infty e^{-ixy} f(x)\ dx$$

Suppose f is continuous at x, then by fourier integral theorem,

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^\infty \left(\int_{-\infty}^\infty f(u) e^{iv(u-x)} du \right) dv \\ &= \int_{-\infty}^\infty e^{-ivx} \left(\frac{1}{2\pi} \int_{-\infty}^\infty e^{ivu} f(u)\ du \right) dv \\ &= \int_{-\infty}^\infty g(v) e^{-ivx} dv = \mathscr{F} g \text{ where } g(v) = \frac{1}{2\pi} \int_{-\infty}^\infty f(u) e^{ivu} du \end{aligned}$$

The above function $g(v)$ gives the **inverse fourier transformation** of f.

Let g be fourier transform of f, then f is uniquely determined by its fourier transform g by,

$$f(x) = \mathscr{F}^{-1} g(y) = \frac{1}{2\pi} \lim_{\alpha \to \infty} \int_{-\alpha}^\alpha g(y) e^{ixy} dy$$

6. Inverse Fourier Transform $\mathscr{F}^{-1}$,

$$\mathscr{F}^{-1} f = \int_{-\infty}^\infty \frac{e^{ixy}}{2\pi} f(x)\ dx$$

13.1.6 Convolutions

Definitions 13.1.2. Let $f, g \in L(-\infty, \infty)$. Let S be the set of all points x for which the Lebesgue integral

$$h(x) = \int_{-\infty}^\infty f(t) g(x-t) dt$$

exists. Then the function $h : S \to \mathbb{R}$ is a **convolution** of f and g. And $h = f * g$.

Remark. Convolution operator is commutative. ie, $h = f * g = g * f$ (hint : take $u = x - t$)

Remark. Suppose f, g vanishes on negative real axis, then

$$h(x) = \int_{-\infty}^\infty f(t)\ g(x-t)\ dt = \int_{-\infty}^0 + \int_0^x + \int_x^\infty f(t)\ g(x-t)\ dt = \int_0^x f(t)\ g(x-t)\ dt$$

Remark. Singularity of convolution is a point at which the convolution integral fails to exists.

Theorem 13.1.3. *Let $f, g \in L(\mathbb{R})$ and either f or g is bounded in $\mathbb{R}$. Then the convoluton integral*

$$h(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt$$

exists for every $x \in \mathbb{R}$ and the function h so defined is bouned in $\mathbb{R}$. In addition, if the bounded function is continuous on $\mathbb{R}$, then h is continuous and $h \in L(\mathbb{R})$.

Synopsis.

Proof. □

Remark. If f, g are both unbounded, the convolution integral may not exist.

$$\text{eg: } f(t) = \frac{1}{\sqrt{t}},\ g(t) = \frac{1}{\sqrt{1-t}}$$

Theorem 13.1.4. *Let $f, g \in L^2(\mathbb{R})$. Then the convolution integral $f * g$ exists for each $x \in \mathbb{R}$ and the function $h : \mathbb{R} \to \mathbb{R}$ defined by $h(x) = f * g(x)$ is bounded in $\mathbb{R}$.*

Synopsis.

Proof. □

13.1.7 The Convolution Theorem for Fourier Tranforms

Theorem 13.1.5. *Let $f, g \in L(\mathbb{R})$ and at least one of f or g is continuous and bounded on $\mathbb{R}$. Let $h = f * g$. Then for every real u,*

$$\int_{-\infty}^{\infty} h(x)e^{-ixu}dx = \left(\int_{-\infty}^{\infty} f(t)e^{-itu}dt\right)\left(\int_{-\infty}^{\infty} g(y)e^{-iyu}dy\right)$$

The integral on the left exists both as a Lebesgue integral and an improper Riemann integral.

Synopsis.

Proof. □

Remark (Application of Convolution Theorem)*.*

$$B(p,q) = \frac{\Gamma p \Gamma q}{\Gamma p + q},\ \text{where } B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx,\ \Gamma p = \int_0^{\infty} t^{p-1}e^{-t}dt$$

13.2 Multivariate Differential Calculus

In this chapter, we deal with real functions of several variables. Instead of **c**, we write $\bar{c} \in \mathbb{R}^n$, then $\bar{c} = (c_1, c_2, \ldots, c_n)$ where $c_j \in \mathbb{R}$ for every $j = 1, 2, \ldots, n$. Again, suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ and $f(\bar{x}) = \bar{y}$, then $\bar{y} = (y_1, y_2, \ldots, y_m)$ where each y_k is real. The unit co-ordinate vector, $\bar{u}_k$ is given by $u_{kj} = \delta_{j,k}$

13.2.1 Directional Derivative

Motivation : The existence of all partial derivatives of a multivariate real function f at a point $\bar{c}$ doesn't imply the continuity of f at $\bar{c}$. Thus, we need a suitable generalisation for the partial derivative which could characterise continuity. And directional derivative is such an attempt.

Definitions 13.2.1 (Directional Derivative)**.** Let $S \subset \mathbb{R}^n$ and $f : S \to \mathbb{R}^m$. Let $\bar{c}$ be an interior points of S and $\bar{u} \in \mathbb{R}^n$, then there exists an open ball $B(\bar{c}, r)$ in S. Also for some $\delta > 0$ the line segment $\alpha : [0, \delta] \to S$ given by $\alpha(t) = \bar{c} + t\bar{u}$ lie in $B(\bar{c}, r)$. Then the **directional derivative** of f at an interior point $\bar{c}$ in the direction $\bar{u}$ is given by

$$f'(\bar{c}, \bar{u}) = \lim_{h \to 0} \frac{f(\bar{c} + h\bar{u}) - f(\bar{c})}{h}$$

Remark. The direction derivative of f at an interior point $\bar{c}$ in the direction $\bar{u}$ exists only if the above limit exists.

Remark. Example, [Apostol, 1973, Exercise 12.2a]

Suppose $\bar{x}, \bar{a}, \bar{c}, \bar{u} \in \mathbb{R}^n$. Let $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(\bar{x}) = \bar{a} \cdot \bar{x}$. Then

$$f'(\bar{c}, \bar{u}) = \lim_{h \to 0} \frac{\bar{a} \cdot (\bar{c} + h\bar{u}) - \bar{a} \cdot \bar{c}}{h} = \bar{a} \cdot \bar{u}$$

Remark (Properties). Let $f : S \to \mathbb{R}^m$, where $S \subset \mathbb{R}^n$

1. $f'(\bar{c}, \bar{0}) = \bar{0}$
   Note : The zero vectors belongs to $\mathbb{R}^n, \mathbb{R}^m$ respectively.

2. $f'(\bar{c}, \bar{u}_k) = \frac{\partial f}{\partial u_k}(\bar{c}) = D_k f(\bar{c})$, the k^{th} partial derivative of f.

3. Let $f = (f_1, f_2, \cdots, f_m)$, such that $f(\bar{c}) = (f_1(\bar{c}), f_2(\bar{c}), \ldots, f_m(\bar{c}))$. Then,

   $$\exists f'(\bar{c}, \bar{u}) \iff \forall k, \exists f_k'(\bar{c}, \bar{u}) \text{ and } f'(\bar{c}, \bar{u}) = (f_1'(\bar{c}, \bar{u}), f_2'(\bar{c}, \bar{u}), \ldots, f_m'(\bar{c}, \bar{u}))$$

   ie, Directional derivative of f exists iff directional derivative of each component function f_k exists. And the components of the directional derivatives of f are the directional derivaties of the components of f.

   Thus $D_k f(\bar{c}) = (D_k f_1(\bar{c}), D_k f_2(\bar{c}), \ldots, D_k f_m(\bar{c}))$ holds.

4. Let $F(t) = f(\bar{c} + t\bar{u})$, then $F'(0) = f'(\bar{c}, \bar{u})$ and $F'(t) = f'(\bar{c} + t\bar{u}, \bar{u})$

5. Let $f(\bar{c}) = \bar{c} \cdot \bar{c} = \|\bar{c}\|^2$, and $F(t) = f(\bar{c} + t\bar{u})$, then $F'(t) = 2\bar{c} \cdot \bar{u} + 2t\|\bar{u}\|^2$ and $F'(0) = f'(\bar{c}, \bar{u}) = 2\bar{c} \cdot \bar{u}$

6. Let f be linear, then $f'(\bar{c}, \bar{u}) = f(\bar{u})$

7. Existence of all partial derivatives doesn't imply existence of all directional derivatives.

   $$f(x, y) = \begin{cases} x + y & \text{if } x = 0 \text{ or } y = 0 \\ 1 & \text{otherwise} \end{cases}$$

   For above f, directional derivatives exists only along the co-ordinates (ie, partial derivatives).

8. Existence of all directional derivatives doesn't imply continuity.

$$f(x,y) = \begin{cases} xy^2(x^2+y^4) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

Above f is discontinuous at $(0,0)$, however all directional derivatives exists and has finite value.

13.2.2 Total Derivative

We may define a total derivative $T_c(h) = hf'(c)$ in the case of real-functions of single variable as follows :-

$$\text{Let } E_c(h) = \begin{cases} \frac{f(c+h)-f(c)}{h} - f'(c), & h \neq 0 \\ 0, & h = 0 \end{cases}$$

Then, $f(c+h) = f(c) + hf'(c) + hE_c(h)$ and as $h \to 0$, $E_c(h) \to 0$. Also $T_c(h) = f'(c)h$ is a linear function of h. ie, $T_c(ah_1 + bh_2) = aT_c(h_1) + bT_c(h_2)$. Now, we will define a total derivative of multivariate function that has these two properties.

Definitions 13.2.2 (Total Derivative)**.** The function $f : \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable** at $\bar{c}$ if there exists a `linear` function $T_{\bar{c}} : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(\bar{c}+\bar{v}) = f(\bar{c}) + T_{\bar{c}}(\bar{v}) + \|\bar{v}\|E_{\bar{c}}(\bar{v})$ where $E_{\bar{c}}(\bar{v}) \to \bar{0}$ as $\bar{v} \to \bar{0}$.

Remark. The linear function $T_{\bar{c}}$ is the total derivative of f at $\bar{c}$, $T_{\bar{c}}(\bar{0}) = \bar{0}$ and the condition above gives the First Order Taylor's Formula for $f(\bar{c}+\bar{v}) - f(\bar{c})$.

Remark (Properties). Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $f'(\bar{c})(\bar{v}) = T_{\bar{c}}(\bar{v})$ be the total derivative of f at $\bar{c}$ evaluated at $\bar{v}$. Then,

1. $f'(\bar{c})(\bar{v}) = f'(\bar{c}, \bar{u})$
2. If f is differentiable at $\bar{c}$, then f is continuous at $\bar{c}$.
3. $f'(\bar{c})(\bar{v}) = v_1 D_1 f(\bar{c}) + v_2 D_2 f(\bar{c}) + \cdots + v_n D_n f(\bar{c})$

Note. The above f' is a function from $\mathbb{R}^n$ to the set of all linear functions $\mathscr{L} = \{h : \mathbb{R}^n \to \mathbb{R}^m\}$. $f'(\bar{c})$ is a linear function (in fact, total derivative $T_{\bar{c}}$) which maps $\bar{v}$ into the directional derivatives of f at $\bar{c}$ in the direction $\bar{v}$. This notation generalises f' for univariate f as well.(put $n = m = 1$)

In this subject, we use the following notations,

$D_k f(\bar{c})$ partial derivative

$f'(\bar{c}, \bar{v})$ directional derivative

$f'(\bar{c})(\bar{v})$ total derivative

$\nabla f(\bar{c})$ gradient vector

Theorem 13.2.1. *If f is differentiable at $\bar{c}$ with total derivative $T_{\bar{c}}$, then for every $\bar{u} \in \mathbb{R}^n$, $T_{\bar{c}}(\bar{u}) = f'(\bar{c}, \bar{u})$. (ie, $f'(\bar{c})(\bar{v}) = f'(\bar{c}, \bar{v})$)*

Proof. For $\bar{v} = \bar{0}$, we have $T_{\bar{c}}(\bar{0}) = 0 = f'(\bar{c}, \bar{0})$.

Suppose $\bar{v} \neq \bar{0}$, then put $\bar{v} = h\bar{u}$. Since f is differentiable at $\bar{c}$, f has total derivative at $\bar{c}$. That is, there exists a linear function $T_{\bar{c}}$ such that $f(\bar{c} + h\bar{u}) = f(\bar{c}) + T_{\bar{c}}(h\bar{u}) + \|h\bar{u}\|E_{\bar{c}}(h\bar{u})$ where $E_{\bar{c}}(h\bar{u}) \to \bar{0}$ as $h\bar{u} \to \bar{0}$.

$$\begin{aligned}
&\implies f(\bar{c} + h\bar{u}) = f(\bar{c}) + hT_{\bar{c}}(\bar{u}) + |h|\|\bar{u}\|E_{\bar{c}}(h\bar{u}),\ E_{\bar{c}}(h\bar{u}) \to \bar{0} \text{ as } h\bar{u} \to \bar{0} \\
&\implies \frac{f(\bar{c} + h\bar{u}) - f(\bar{c})}{h} = T_{\bar{c}}(\bar{u}) + \frac{|h|\|\bar{u}\|E_{\bar{c}}(h\bar{u})}{h},\ E_{\bar{c}}(h\bar{u}) \to \bar{0} \text{ as } h \to 0 \\
&\implies \lim_{h\to 0} \frac{f(\bar{c} + h\bar{u}) - f(\bar{c})}{h} = T_{\bar{c}}(\bar{u}) + \lim_{h\to 0} \frac{|h|\|\bar{u}\|E_{\bar{c}}(h\bar{u})}{h} \\
&\implies f'(\bar{c}, \bar{u}) = T_{\bar{c}}(\bar{u})
\end{aligned}$$

□

Note. $T_{\bar{c}}$ is linear, however $E_{\bar{c}}$ is not linear. Thus $E_{\bar{c}}(h\bar{u}) \neq hE_{\bar{c}}(\bar{u})$.

As $h \to 0$, $h\bar{u} \to \bar{0}$ and $E_{\bar{c}}(h\bar{u}) \to \bar{0}$. Since the order of the function $E_{\bar{c}}(h\bar{u})$ is much smaller than that of h, the limit on the right converges to 0.

Theorem 13.2.2. *If f is differentiable at $\bar{c}$, then f is continuous at $\bar{c}$.*

Proof. Let $\bar{v} \neq 0$, then

$$\begin{aligned}
\bar{v} &= v_1\bar{u}_1 + v_2\bar{u}_2 + \cdots + v_n\bar{u}_n, \\
\bar{v} \to \bar{0} &\implies \forall j,\ v_j \to 0 \\
T \text{ is linear } &\implies T_{\bar{c}}(\bar{v}) = v_1T_{\bar{c}}(\bar{u}_1) + v_2T_{\bar{c}}(\bar{u}_2) + \cdots + v_nT_{\bar{c}}(\bar{u}_n) \\
&\text{Thus, } T_{\bar{c}}(\bar{v}) \to \bar{0} \text{ as } \bar{v} \to 0
\end{aligned}$$

Since f differentiable at $\bar{c}$, there exists linear function $T_{\bar{c}}$ such that

$$\begin{aligned}
&f(\bar{c} + \bar{v}) = f(\bar{c}) + T_{\bar{c}}(\bar{v}) + \|v\|E_{\bar{c}}(\bar{v}) \\
&\implies \lim_{\bar{v}\to\bar{0}} f(\bar{c} + \bar{v}) = f(\bar{c}) + \lim_{\bar{v}\to\bar{0}} T_{\bar{c}}(\bar{v}) + \lim_{\bar{v}\to\bar{0}} \|v\|E_{\bar{c}}(\bar{v}) \\
&\implies \lim_{\bar{v}\to\bar{0}} f(\bar{c} + \bar{v}) = f(\bar{c})
\end{aligned}$$

□

Theorem 13.2.3. *Let $S \subset \mathbb{R}^n$ and $f : S \to \mathbb{R}^m$ be differentiable at an interior point $\bar{c}$ of S, where $S \subseteq \mathbb{R}^n$. If $\bar{v} = v_1\bar{u}_1 + v_2\bar{u}_2 + \cdots + v_n\bar{u}_n$, then*

$$f'(\bar{c})(\bar{v}) = \sum_{k=1}^{n} v_kD_kf(\bar{c})$$

In particular, if f is real-valued $(m = 1)$ we have, $f'(\bar{c})(\bar{v}) = \nabla f(\bar{c}).\bar{v}$

Proof. Suppose $f : S \to \mathbb{R}^m$ is differentiable at $\bar{c}$, then there exists a linear function $f'(\bar{c}) : S \to \mathbb{R}^m$ such that $f(\bar{c} + \bar{v}) = f(\bar{c}) + f'(\bar{c})(\bar{v}) + \|\bar{v}\|E_{\bar{c}}(\bar{c})$ where

$E_{\bar{c}} \to \bar{0}$ as $\bar{v} \to \bar{0}$.

$$\begin{aligned} f'(\bar{c})(\bar{v}) &= f'(\bar{c})\left(\sum_{k=1}^{n} v_k \bar{u}_k\right) \\ &= \sum_{k=1}^{n} v_k f'(\bar{c})(\bar{u}_k), \text{ since } f'(\bar{c}) \text{ is linear} \\ &= \sum_{k=1}^{n} v_k D_k f(\bar{c}), \text{ since } f'(\bar{c})(\bar{u}_k) = f'(\bar{c}, \bar{u}_k) = D_k f(\bar{c}) \end{aligned}$$

Let $m = 1$, then $f : S \to \mathbb{R}$

$$\begin{aligned} f'(\bar{c})(\bar{v}) &= \sum_{k=1}^{n} v_k D_k f(\bar{c}) = \nabla f(\bar{c}).\bar{v} \\ &\text{since } \nabla f(\bar{c}) = (D_1 f(\bar{c}),\ D_2 f(\bar{c}),\ \cdots,\ D_n f(\bar{c})) \end{aligned}$$

□

Remark. Let $f : S \to \mathbb{R}$, then $f(\bar{c} + \bar{v}) = f(\bar{c}) + \nabla f(\bar{c}).\bar{v} + o(\|\bar{v}\|)$ as $\bar{v} \to \bar{0}$.

Remark (Complex-valued Functions).

13.2.3 Matrix of Linear Function

Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be a linear function. Let $\{\bar{u}_1,\ \bar{u}_2,\ \cdots,\ \bar{u}_n\}$ be standard basis for $\mathbb{R}^n$ and $\{\bar{e}_1,\ \bar{e}_2,\ \cdots,\ \bar{e}_m\}$ be standard basis for $\mathbb{R}^m$. Let $\bar{v} \in \mathbb{R}^n$, then $\bar{v} = \sum_{k=1}^{n} v_k \bar{u}_k$ and $T(\bar{v}) = \sum_{k=1}^{n} v_k T(\bar{u}_k)$ and

$$\begin{aligned} T(\bar{v}) &= \begin{bmatrix} v_1 & v_2 & \vdots & v_n \end{bmatrix} \begin{bmatrix} T(\bar{u}_1) \\ T(\bar{u}_2) \\ \cdots \\ T(\bar{u}_n) \end{bmatrix} \\ &= \begin{bmatrix} v_1 & v_2 & \vdots & v_n \end{bmatrix} \begin{bmatrix} t_{11}\bar{e}_1 + t_{21}\bar{e}_2 + \cdots + t_{m1}\bar{e}_m \\ t_{12}\bar{e}_1 + t_{22}\bar{e}_2 + \cdots + t_{m2}\bar{e}_m \\ \cdots \\ t_{1n}\bar{e}_1 + t_{2n}\bar{e}_2 + \cdots + t_{mn}\bar{e}_m \end{bmatrix} \\ &= \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} t_{11} & t_{21} & \cdots & t_{m1} \\ t_{12} & t_{22} & \cdots & t_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1n} & t_{2n} & \cdots & t_{mn} \end{bmatrix} \begin{bmatrix} \bar{e}_1 \\ \bar{e}_2 \\ \cdots \\ \bar{e}_m \end{bmatrix} \end{aligned}$$

We may take the transpose,

$$T(\bar{v}) = \begin{bmatrix} \bar{e}_1 & \bar{e}_2 & \cdots & \bar{e}_m \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \cdots \\ v_n \end{bmatrix}$$

$$T(\bar{v}) = T\left(\sum_{k=1}^{n} v_k \bar{u}_k\right) = \sum_{k=1}^{n} v_k T(\bar{u}_k) = \sum_{k=1}^{n} v_n \sum_{j=1}^{m} t_{kj} \bar{e}_j$$

Thus matrix of T is given by, $m(T) = (t_{ik})$ where $T(\bar{u}_k) = \sum_{k=1}^{n} t_{ik}\bar{e}_i$.

Remark (Example). Let $T : \mathbb{R}^3 \to \mathbb{R}^2$ defined by $T(x, y, z) = (2x + y, y - z)$.

$$\begin{aligned}
T(1,2,3) =& T((1,0,0) + 2(0,1,0) + 3(0,0,1)) \\
=& T(\bar{u}_1 + 2\bar{u}_2 + 3\bar{u}_3) \\
=& T(\bar{u}_1) + 2T(\bar{u}_2) + 3T(\bar{u}_3) \\
=& \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} T(\bar{u}_1) \\ T(\bar{u}_2) \\ T(\bar{u}_3) \end{bmatrix} \\
=& \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} (2,0) \\ (1,1) \\ (0,-1) \end{bmatrix} \\
=& \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2(1,0) \\ (1,0) + (0,1) \\ -1(0,1) \end{bmatrix} \\
=& \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2\bar{e}_1 \\ \bar{e}_1 + \bar{e}_2 \\ -\bar{e}_2 \end{bmatrix} \\
=& \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \bar{e}_1 \\ \bar{e}_2 \end{bmatrix} \\
=& 4\bar{e}_1 - \bar{e}_2 = 4(1,0) - 1(0,1) = (4,-1)
\end{aligned}$$

$$\text{In the above case, } m(T) = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & -1 \end{bmatrix}$$

Using the matrix of linear function $m(T)$, we can compute the image of any point in R^3 by matrix multiplication.

Matrix of the composition of two linear functions

Let $T : \mathbb{R}^n \to \mathbb{R}^m$ and $S : \mathbb{R}^m \to \mathbb{R}^p$ be two linear functions with domain of S containing the range of T (so that $S \circ T$ is well defined). Then $S \circ T : \mathbb{R}^n \to \mathbb{R}^p$ is defined by

$$S \circ T(\bar{x}) = S(T(\bar{x})), \ \forall \bar{x} \in \mathbb{R}^n$$

Since S, T are linear, $S \circ T$ is also linear.

$$\begin{aligned}
S \circ T(a\bar{x} + b\bar{y}) =& S(T(a\bar{x} + b\bar{y})) = S(aT(\bar{x}) + bT(\bar{y})) = aS(T(\bar{x}))) + bS(T(\bar{y})) \\
=& aS \circ T(\bar{x}) + bS \circ T(\bar{y}), \ \forall a, b \in \mathbb{R}, \ \forall \bar{x}, \bar{y} \in \mathbb{R}^n
\end{aligned}$$

Let $\{\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_n\}$ be the standards basis for $\mathbb{R}^n$, $\{\bar{e}_1, \bar{e}_2, \cdots, \bar{e}_m\}$ be the standards basis for $\mathbb{R}^m$ and $\{\bar{w}_1, \bar{w}_2, \cdots, \bar{w}_p\}$ be the standards basis for $\mathbb{R}^p$. Let

$\bar{v} \in \mathbb{R}^n$, then $\bar{(}v) = \sum_{i=1}^n v_i \bar{u}_i$, and $S \circ T(\bar{v}) = \sum_{i=1}^n v_n S \circ T(\bar{u}_i)$

$$S \circ T(\bar{v}) = \begin{bmatrix} v_1 & v_2 & \vdots & v_n \end{bmatrix} \begin{bmatrix} S \circ T(\bar{u}_1) \\ S \circ T(\bar{u}_2) \\ \cdots \\ S \circ T(\bar{u}_n) \end{bmatrix}$$

$$= \begin{bmatrix} v_1 & v_2 & \vdots & v_n \end{bmatrix} \begin{bmatrix} S(t_{11}\bar{e}_1 + \cdots + t_{m1}\bar{e}_m) \\ S(t_{12}\bar{e}_1 + \cdots + t_{m2}\bar{e}_m) \\ \cdots \\ S(t_{1n}\bar{e}_1 + \cdots + t_{mn}\bar{e}_m) \end{bmatrix}$$

$$= \begin{bmatrix} v_1 & v_2 & \vdots & v_n \end{bmatrix} \begin{bmatrix} t_{11}S(\bar{e}_1) + \cdots + t_{m1}S(\bar{e}_m) \\ t_{12}S(\bar{e}_1) + \cdots + t_{m2}S(\bar{e}_m) \\ \cdots \\ t_{n1}S(\bar{e}_1) + \cdots + t_{mn}S(\bar{e}_m) \end{bmatrix}$$

$$= \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} t_{11} & t_{21} & \cdots & t_{m1} \\ t_{12} & t_{22} & \cdots & t_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1n} & t_{2n} & \cdots & t_{mn} \end{bmatrix} \begin{bmatrix} S(\bar{e}_1) \\ S(\bar{e}_2) \\ \cdots \\ S(\bar{e}_m) \end{bmatrix}$$

$$= \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} t_{11} & t_{21} & \cdots & t_{m1} \\ t_{12} & t_{22} & \cdots & t_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1n} & t_{2n} & \cdots & t_{mn} \end{bmatrix} \begin{bmatrix} s_{11}\bar{w}_1 + s_{12}\bar{w}_2 + \cdots + s_{1p}\bar{w}_p \\ s_{12}\bar{w}_1 + s_{22}\bar{w}_2 + \cdots + s_{p2}\bar{w}_p \\ \cdots \\ s_{1m}\bar{w}_1 + s_{2m}\bar{w}_2 + \cdots + s_{pm}\bar{w}_p \end{bmatrix}$$

$$= \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} t_{11} & t_{21} & \cdots & t_{m1} \\ t_{12} & t_{22} & \cdots & t_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ t_{1n} & t_{2n} & \cdots & t_{mn} \end{bmatrix} \begin{bmatrix} s_{11} & s_{21} & \cdots & s_{p1} \\ s_{12} & s_{22} & \cdots & s_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1m} & s_{2m} & \cdots & s_{pm} \end{bmatrix} \begin{bmatrix} \bar{w}_1 \\ \bar{w}_2 \\ \vdots \\ \bar{w}_p \end{bmatrix}$$

We may take transpose,

$$S \circ T(\bar{v}) = \begin{bmatrix} \bar{w}_1 & \bar{w}_2 & \vdots & \bar{w}_p \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pm} \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \cdots \\ v_n \end{bmatrix}$$

Remember : Given $T : \mathbb{R}^n \to \mathbb{R}^m$, then we may take $m(T)$ either as $m \times n$ matrix or $n \times m$ matrix. Since, we chose $m \times n$, $m(S \circ T) = m(S)m(T)$. Otherwise, $m(S \circ T) = m(T)m(S)$. This may change for different authors.

Suppose $m(S) = (s_{ij})$ and $m(T) = (t_{ij})$ respectively. Then

$$S(e_k) = \sum_{i=1}^{p} s_{ik}\bar{w}_i, \; k = 1, 2, \cdots, m \text{ and}$$

$$T(u_j) = \sum_{k=1}^{m} t_{kj}\bar{e}_k, \; j = 1, 2, \cdots, n$$

$$\begin{aligned}(S\circ T)(\bar{u}_j) =& S(T(\bar{u}_j)) = S\left(\sum_{k=1}^{m} t_{kj}\bar{e}_k\right) = \sum_{k=1}^{m} t_{kj}S(\bar{e}_k)\\ =& \sum_{k=1}^{m} t_{kj}\left(\sum_{i=1}^{p} s_{ik}\bar{w}_i\right) = \sum_{i=1}^{p}\left(\sum_{k=1}^{m} s_{ik}t_{kj}\right)\bar{w}_i\end{aligned}$$

Therefore, $m(S\circ T) = \sum_{k=1}^{m} s_{ik}t_{kj} = (s_{ik})(t_{kj}) = m(S)m(T)$.

13.2.4 The Jacobian Matrix

Let $\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_n$ be the unit co-ordinate vectors in $\mathbb{R}^n$ and $\bar{e}_1, \bar{e}_2, \cdots, \bar{e}_m$ be the unit co-ordinate vectors in $\mathbb{R}^m$. Let function $f : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at $\bar{c} \in \mathbb{R}^n$. Then there exists a linear function $T = f'(\bar{c}) : \mathbb{R}^n \to \mathbb{R}^m$ such that $f(\bar{c}+\bar{v}) = f(\bar{c}) + f'(\bar{c})(\bar{v}) + \|\bar{v}\| + E_{\bar{c}}(\bar{v})$. We have, $T(\bar{u}_k) = f'(\bar{c})(\bar{u}_k) = f'(\bar{c}, \bar{u}_k) = D_k f(\bar{c}) = D_k \sum_{i=1}^{m} f_i(\bar{c})\bar{e}_i$.

Clearly, the matrix of total derivative T, $m(T) = (t_{ik}) = (D_k f_i(\bar{c}))$. This matrix is called Jacobian matrix of f at $\bar{c}$ and is denoted by $Df(\bar{c})$.

$$Df(\bar{c}) = \begin{bmatrix} D_1 f_1(\bar{c}) & D_2 f_1(\bar{c}) & \cdots & D_n f_1(\bar{c})\\ D_1 f_2(\bar{c}) & D_2 f_2(\bar{c}) & \cdots & D_n f_2(\bar{c})\\ \vdots & \vdots & \ddots & \vdots\\ D_1 f_m(\bar{c}) & D_2 f_m(\bar{c}) & \cdots & D_n f_m(\bar{c})\end{bmatrix}$$

Properties of Jacobian matrix

1. kth row of $Df(\bar{c})$ is gradient vector of f_k
$$\nabla f_k(\bar{c}) = (D_1 f_k(\bar{c}), D_2 f_k(\bar{c}), \cdots, D_n f_k(\bar{c}))$$
2. When $m = 1$, $Df(\bar{c}) = \nabla f(\bar{c})$.
3. $f'(\bar{c})(\bar{v}) = \sum\limits_{k=1}^{m} (\nabla f_k(\bar{c})\cdot\bar{v})\,\bar{e}_k$
4. $\|f'(\bar{c})(\bar{v})\| \leq M\|v\|$ where $M = \sum\limits_{k=1}^{m}\|\nabla f_k(\bar{c})\|$, by property (3)
5. $f'(\bar{c})(\bar{v}) \to \bar{0}$ as $\bar{v} \to \bar{0}$, by property (4)

Chain Rule

Chain Rule for real function : $\frac{dF\circ G}{dx}(x) = \frac{d}{dy}F(y)\frac{d}{dx}G(x) = F'(y)\,G'(x)$

For example : $\frac{d}{dx}(ax+3)^3 = \frac{d}{dy}y^3\frac{d}{dx}\,(ax+3) = 3ay^2 = 3a(ax+3)^2$

Theorem 13.2.4. *Let g be differentiable at $\bar{a}$, with total derivative $g'(\bar{a})$ and $\bar{b} = g(\bar{a})$. Let f is differentiable at $\bar{b}$, with total derivative $f'(\bar{b})$. Then $h = f\circ g$ is differentiable at $\bar{a}$ with total derivative $h'(\bar{a}) = f'(\bar{b})\circ g'(\bar{a})$. Try to read $h'(\bar{a}) = H$, $f'(\bar{b}) = F$, $g'(\bar{a}) = G$, then $H = F\circ G \implies H(x) = F(G(x))$ In other words, $h'(\bar{a})(\bar{v}) = f'(\bar{b})\circ g'(\bar{a})\ (\bar{v}) = f'(\bar{b})(g'(\bar{a})(\bar{v}))$.*

Proof. Given $\epsilon > 0$, let $y \in \mathbb{R}^p$ such that $\|y\| < \epsilon$. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^p \to \mathbb{R}^n$, then $h = f \circ g : \mathbb{R}^p \to \mathbb{R}^m$.

We have, $h(\bar{a} + \bar{y}) - h(\bar{a}) = f(g(\bar{a} + \bar{y})) - f(g(\bar{a})) = f(\bar{b} + \bar{v}) - f(\bar{b})$ where $\bar{b} = g(\bar{a})$, and $\bar{v} = g(\bar{a} + \bar{y}) - g(\bar{a})$.

Since g is differentiable at $\bar{a}$, g satisfies first-order Taylor's formula.

$$\begin{aligned} g(\bar{a} + \bar{y}) &= g(\bar{a}) + g'(\bar{a})(\bar{y}) + \|\bar{y}\| E_{\bar{a}}(\bar{y}) \text{ where } E_{\bar{a}} \to \bar{0} \text{ as } \bar{y} \to \bar{0} \\ \implies \bar{v} &= g(\bar{a} + \bar{y}) - g(\bar{a}) = g'(\bar{a})(\bar{y}) + \|\bar{y}\| E_{\bar{a}}(\bar{y}) \end{aligned}$$

Clearly, as $\bar{y} \to \bar{0} \implies \bar{v} \to g'(\bar{a})(\bar{0}) = \bar{0}$. Again, we have f is differentiable at $\bar{b}$, thus f satisfies first-order Taylor's formula.

$$f(\bar{b} + \bar{v}) = f(\bar{b}) + f'(\bar{b})(\bar{v}) + \|\bar{v}\| E_{\bar{b}}(\bar{v}) \text{ where } E_{\bar{b}} \to \bar{0} \text{ as } \bar{v} \to \bar{0}$$

$$\begin{aligned} \implies f(\bar{b} + \bar{v}) - f(\bar{b}) &= f'(\bar{b})(\bar{v}) + \|\bar{v}\| E_{\bar{b}}(\bar{v}) \\ &= f'(\bar{b})\left(g'(\bar{a})(\bar{y}) + \|\bar{y}\| E_{\bar{a}}(\bar{y})\right) + \|\bar{v}\| E_{\bar{b}}(\bar{v}) \\ &= f'(\bar{b})(g'(\bar{a})(\bar{y})) + \|\bar{y}\| E(\bar{y}) \\ &\text{where } E(\bar{y}) = f'(\bar{b})(E_{\bar{a}}(\bar{y})) + \frac{\|\bar{v}\|}{\|\bar{y}\|} E_{\bar{b}}(\bar{v}),\ \bar{y} \neq \bar{0} \end{aligned}$$

$$\implies h(\bar{a} + \bar{y}) - h(\bar{a}) = f(\bar{b} + \bar{v}) - f(\bar{b}) = f'(\bar{b})(g'(\bar{a})(\bar{y})) + \|\bar{y}\| E(\bar{y})$$

Since $f'(\bar{b})$ and $g'(\bar{a})$ are linear, their composition is also linear. Therefore, h is differentiable at $\bar{a}$ with a linear, total derivative $h'(\bar{a}) = f'(\bar{b}) \circ g'(\bar{a})$ as it satisfies first-order Taylor's formula if $E_{\bar{y}} \to \bar{0}$ as $\bar{y} \to \bar{0}$.

We have, $\|\bar{v}\| \leq \|g'(\bar{a})(\bar{y})\| + \|\bar{y}\|\ \|E_{\bar{a}}(\bar{y})\| \leq M\|y\| + \|E_{\bar{a}}(\bar{y})\|\ \|\bar{y}\|$.

$$\implies \frac{\|\bar{v}\|}{\|\bar{y}\|} \leq M + \|E_{\bar{a}}(\bar{y})\|$$

Thus, $\bar{v} \to \bar{0}$ as $\bar{y} \to \bar{0}$. Then $f'(\bar{b})(\bar{v}) \to f'(\bar{b})(\bar{0}) = \bar{0}$. And $E_{\bar{a}}(\bar{y}) \to \bar{0}$. Therefore, $E(\bar{y}) \to \bar{0} + M\bar{0} = \bar{0}$ as $\bar{y} \to \bar{0}$. □

Matrix form of the chain rule

Let $f : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^p \to \mathbb{R}^n$. And $h = f \circ g : \mathbb{R}^p \to \mathbb{R}^m$. Suppose g is differentiable at $\bar{a} \in \mathbb{R}^p$ and f is differentiable at $g(\bar{a}) = \bar{b} \in \mathbb{R}^n$. Then h is differentiable at $\bar{a}$ and the Jacobian matrix of h is given by the chain rule,

$$Dh(\bar{a}) = Df(\bar{b})Dg(\bar{a}) \text{ where } h = f \circ g,\ \bar{b} = g(\bar{a})$$

In other words,

$$D_j h_i(\bar{a}) = \sum_{k=1}^{n} D_k f_i(\bar{b}) D_j g_k(\bar{a}),\ i = 1, 2, \cdots, m,\ j = 1, 2, \cdots, p$$

For $m = 1$, $D_j h(\bar{a}) = \sum\limits_{k=1}^{n} D_k f(\bar{b}) D_j g_k(\bar{a})$

For $m = 1$ and $p = 1$, $h'(\bar{a}) = \sum\limits_{k=1}^{n} Df(\bar{b}) g_k'(\bar{a}) = \nabla f(\bar{b}) \cdot Dg(\bar{a})$

Theorem 13.2.5. *Let f and $D_2 f$ be continuous functions on a rectangle $[a,b] \times [c,d]$. Let p and q be differentiable on $[c,d]$, where $p(y) \in [a,b]$ and $q(y) \in [c,d]$ for each $y \in [c,d]$. Define F by the equation,*

$$F(y) = \int_{p(y)}^{q(y)} f(x,y)dx, \ \ y \in [c,d]$$

Then $F'(y)$ exists for each $y \in (c,d)$ and is given by,

$$F'(y) = \int_{p(y)}^{q(y)} D_2 f(x,y)dx + f((q,y),y)q'(y) - f(p(y),y)p'(y)$$

The following two theorems are required for proving the theorem on differentiating an integral.

Theorem 13.2.6. *Let α be of bounded variation on $[a,b]$ and assume that $f \in \mathcal{R}(\alpha)$ on $[a,b]$.*

$$\textit{Define } F(x) = \int_a^x f \ d\alpha, \ \ x \in [a,b]$$

Then F is of bounded variation on $[a,b]$ and F is continuous at x if α is continuous at x. If α is increasing on $[a,b]$, then the derivative $F'(x)$ exists at each $x \in (a,b)$ where $\alpha'(x)$ exists and where f is continuous. And

$$F'(x) = f(x)\alpha'(x)$$

Theorem 13.2.7. *Let $Q = \{(x,y) : a \le x \le b, \ c \le y \le d\}$. Assume that α is of bounded variation on $[a,b]$ and for each $y \in [c,d]$, assume that the integral*

$$F(y) = \int_a^b f(x,y) \ d\alpha(x)$$

exists. If the partial derivative $D_2 f$ is continuous on Q, the derivative $F'(y)$ exists for each $y \in (c,d)$ and is given by

$$F'(y) = \int_a^b D_2 f(x,y) \ d\alpha(x)$$

Proof. Let $G(x_1,x_2,x_3) = \int_{x_1}^{x_2} f(t,x_3) \ dt$. Then we may write $F(y)$ in terms of G. That is, $F(y) = G(p(y),q(y),y)$.

Step 1 : 1-D Chain Rule

By 1-dimensional chain rule, we have

$$\begin{aligned} F'(y) = \frac{dF}{dy} &= \frac{\partial G}{\partial p}\frac{dp}{dy} + \frac{\partial G}{\partial q}\frac{dq}{dy} + \frac{\partial G}{\partial y} \\ &= D_1 G \ p'(y) + D_2 G \ q'(y) + D_3 G \end{aligned}$$

Step 2 : $D_1 G$

Since the variable of differentition is present in the limit of the integral, we use theorem 13.2.6 to compute the derivative of the integral. We may write,

$$G(p(y), q(y), y) = -\int_{q(y)}^{p(y)} f(t, y)\ dt \tag{13.6}$$

We are differentiating (partially) with respect to $p(y)$. Thus $q(y)$, y are constants for this differentiation.

$$\begin{aligned} G(x, a, y) &= -H(x) = -\int_a^x f(t, y)\ dt \\ \implies D_1G &= -H'(x) = -f(x, y). \\ \text{Thus, } D_1G &= -f(p(y), y) \end{aligned}$$

Step 3 : D_2G

Again, variable of differentiation is present in the limit of the integral. Thus, we write,

$$G(p(y), q(y), y) = \int_{p(y)}^{q(y)} f(t, y)\ dt \tag{13.7}$$

Now we are differentiating (partially) with respect to the the second component of G which is $q(y)$. Clearly, $p(y)$ and y are treated as constants.

$$\begin{aligned} G(a, x, y) &= H(x) = \int_a^x f(t, y)\ dt \\ \implies D_2G &= H'(x) = f(q(y), y) \end{aligned}$$

Step 4 : D_3G

Now the variable of integration is not affecting the limits of the integral. Also it is given that D_2f is continuous on $[a, b] \times [c, d]$. We write

$$\begin{aligned} G(a, b, x) &= H(x) = \int_a^b f(t, x)\ dt \\ \implies D_3G &= H'(x) = \int_a^b D_2f(t, x)\ dt \\ \text{Thus, } D_3G &= \int_{p(y)}^{q(y)} f(t, y)\ dt \end{aligned}$$

□

The mean-value theorem for differentiable functions

Theorem 13.2.8 (Mean-Value)**.** *Let S be an open subset of $\mathbb{R}^n$. Assume $f : S \to \mathbb{R}^m$ is differentiable at each point of S. Let $\bar{x}$, $\bar{y}$ be two points in S such that $L(\bar{x}, \bar{y}) = \{t\bar{x} + (1-t)\bar{y} : t \in [0, 1]\}$ is subset of S. Then for every $\bar{a} \in \mathbb{R}^m$, there exists a point $\bar{z} \in L(\bar{x}, \bar{y})$ such that*

$$\bar{a}.\,(f(\bar{y}) - f(\bar{x})) = \bar{a}.f'(\bar{z})(\bar{y} - \bar{x})$$

Proof. Let $\bar{u} = \bar{y} - \bar{x}$. We have S is open subset and $L(\bar{x}, \bar{y}) \subset S$, thus there exists $\delta > 0$ such that $\bar{x} + t\bar{u} \in S, \forall t \in (-\delta, 1 + \delta)$. In other words, the 'Line segment $L(\bar{x}, \bar{y})$' is properly contained in S, in such a way that extending the Line from $\bar{x}$ to $\bar{y}$ a little bit extra one either sides is still contained in S.

Let $\bar{a} \in \mathbb{R}^m$ and $F : (-\delta, 1+\delta) \to \mathbb{R}$ defined by $F(t) = \bar{a}.f(\bar{x} + t\bar{u})$. Then F is differentiable at each $t \in (-\delta, 1+\delta)$ and the derivative $F'(t) = \bar{a}.f'(\bar{x}+t\bar{u}, \bar{u})$, the directional derivative of $f(\bar{x} + t\bar{u})$ with respect to $\bar{u}$.

$$f'(\bar{x} + t\bar{u}, \bar{u}) = f'(\bar{x} + t\bar{u})(\bar{u}) \implies F'(t) = \bar{a}.f'(\bar{x} + t\bar{u})(\bar{u})$$

By 1-dimensional mean-value theorem, we have

$$\exists \theta \in (0, 1) \text{ such that } F(1) - F(0) = F'(\theta)$$

By definition of F, $F(1) = \bar{a}.f(\bar{x} + \bar{u}) = \bar{a}.f(\bar{y})$. And $F(0) = \bar{a}.f(\bar{x})$. Therefore,

$$F'(\theta) = F(1) - F(0) = \bar{a}.f(\bar{y}) - \bar{a}.f(\bar{x}) = \bar{a}.(f(\bar{y}) - f(\bar{x}))$$

We also have,

$$F'(\theta) = \bar{a}.f'(\bar{x} + \theta\bar{u})(\bar{u}) = \bar{a}.f'(\bar{z})(\bar{y} - \bar{x}), \text{ where } \bar{z} = \bar{x} + \theta\bar{u} \in L(\bar{x}, \bar{y})$$

□

Remark. Suppose S is convex in $\mathbb{R}^m$. Then for every pair of points $\bar{x}, \bar{y} \in S$, $L(\bar{x}, \bar{y}) \subset S$. Thus Mean-value theorem holds for all $\bar{x}, \bar{y} \in S$.

13.3 Multivariate Calculus

13.3.1 A sufficient condition for differentiability

Theorem 13.3.1. *Suppose one of the partial derivatives $D_1 f, D_2 f, \cdots, D_n f$ exists at $\bar{c}$. And the remaining $n-1$ partial derivatives exists in some n-ball $B(\bar{c})$ and are continuous at $\bar{c}$. Then f is differentiable at $\bar{c}$.*

Proof. Step 1 : Real-valued function

We claim that the function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\bar{c}$ iff each component f_k is differentiable at $\bar{c}$.

Suppose f is differentiable at $\bar{c}$, then there exists a linear, total derivative function $f'(\bar{c})$ satisfying first-order Taylor's formula at $\bar{c}$.

ie, $f(\bar{c} + \bar{v}) = f(\bar{c}) + f'(\bar{c})(\bar{v}) + \|\bar{v}\| E_{\bar{c}}(\bar{v})$ where $E_{\bar{c}}(\bar{v}) \to \bar{0}$ as $\bar{v} \to \bar{0}$.

$$\begin{aligned}
f(\bar{c} + \bar{v}) &= (f_1(\bar{c} + \bar{v}), f_2(\bar{c} + \bar{v}), \cdots, f_m(\bar{c} + \bar{v})) \\
f(\bar{c}) &= (f_1(\bar{c}), f_2(\bar{c}), \cdots, f_m(\bar{c})) \\
f'(\bar{c})(\bar{v}) &= (f_1'(\bar{v}), f_2'(\bar{v}), \cdots, f_m'(\bar{v})) \\
E_{\bar{c}}(\bar{v}) &= (E_1(\bar{v}), E_2(\bar{v}), \cdots, E_m(\bar{v}))
\end{aligned}$$

where each component of the error function $E_k(\bar{v}) \to 0$ as $\bar{v} \to \bar{0}$. Also since $f'(\bar{c})$ is linear, each of its components $f_k' : \mathbb{R}^n \to \mathbb{R}$ are linear.

$$\begin{aligned}
f(\bar{c} + \bar{v}) &= (f_1(\bar{c} + \bar{v}), f_2(\bar{c} + \bar{v}), \cdots, f_m(\bar{c} + \bar{v})) \\
&= (f_1(\bar{c}), f_2(\bar{c}), \cdots, f_m(\bar{c})) + (f_1'(\bar{v}), f_2'(\bar{v}), \cdots, f_m'(\bar{v})) \\
&\quad + \|\bar{v}\| \left(E_1(\bar{v}), E_2(\bar{v}), \cdots, E_m(\bar{v})\right) \text{ where } E_k(\bar{v}) \to 0 \text{ as } \bar{v} \to \bar{0} \\
&= (f_1(\bar{c}), f_2(\bar{c}), \cdots, f_m(\bar{c})) + (f_1'(\bar{v}), f_2'(\bar{v}), \cdots, f_m'(\bar{v})) \\
&\quad + (\|\bar{v}\| E_1(\bar{v}), \|\bar{v}\| E_2(\bar{v}), \cdots, \|\bar{v}\| E_m(\bar{v})) \text{ where } E_k(\bar{v}) \to 0 \text{ as } \bar{v} \to \bar{0} \\
&= (f_1(\bar{c}) + f_1'(\bar{v}) + \|\bar{v}\| E_1(\bar{v}), \cdots, f_m(\bar{c}) + f_m'(\bar{v}) + \|\bar{v}\| E_m(\bar{v}))
\end{aligned}$$

Thus first-order Taylor's forumula for f at $\bar{c}$ gives first-order Taylor's forumula for each of its components f_k. ie, $f_k(\bar{c}+\bar{v}) = f_k(\bar{c})+f'_k(\bar{v}+\|\bar{v}\|E_k(\bar{v})$ where $E_k(\bar{v}) \to 0$ as $\bar{v} \to \bar{0}$. Therefore, f_k are differentiable at $\bar{c}$ for $k = 1, 2, \cdots, m$.

Suppose each component f_k of f are differentiable at $\bar{c}$. Then there exists linear, total derivative functions f'_k satisfying first-order Taylor's formula at $\bar{c}$. ie, $f_k(\bar{c}+\bar{v}) = f_k(\bar{v}) + f'_k(\bar{v}) + \|\bar{v}\|E_k(\bar{v})$ where $E_k(\bar{v}) \to 0$ as $\bar{v} \to \bar{0}$.

Define $E_{\bar{c}}(\bar{v}) = (E_1(\bar{v}), E_2(\bar{v}), \cdots, E_k(\bar{v}))$. Then $E_{\bar{c}}(\bar{v}) \to \bar{0}$ as $\bar{v} \to \bar{0}$. Therefore, there exists a linear, total derivative function $f'(\bar{c}) = (f'_1, f'_2, \cdots, f'_m)$ satisfying first-order Taylor's formula at $\bar{c}$.

Thus, if each (real-valued) component function f_k are differentiable, then f is also differentiable. Therefore, it is sufficient to prove the theorem for a real-valued function.

Step 2: Telescopic Sum

Assume (without loss of generality) that $D_1 f$ exists at $\bar{c}$ and $D_2 f, D_3 f, \cdots, D_n f$ exist and continuous in some n-ball $B(\bar{c})$. Suppose $D_r f$ exists at $\bar{c}$ and all partial derivatives execept $D_r f$ are continuous. Then $v_0 = \bar{0}$, $v_1 = y_r\bar{u}_r$, $v_2 = y_r\bar{u}_r + y_1\bar{u}_1, \ldots$. Then the following proof can be applied without any loss of generality.

Let $\bar{v} = \lambda\bar{y}$ where $\bar{y} = \frac{\bar{v}}{\|\bar{v}\|}$. Clearly, $\|\bar{y}\| = 1$ and $\lambda = \|\bar{v}\|$. Choose $\lambda > 0$ such that $\bar{c} + \bar{v} \in B(\bar{c})$ and all the partial derivatives $D_2 f, D_3 f, \cdots, D_n f$ exists and are continuous in $B(\bar{c})$.

We have, $\bar{y} = (y_1, y_2, \cdots, y_n) = y_1\bar{u}_1 + y_2\bar{u}_2 + \cdots + y_n\bar{u}_n$.

Define $\bar{v}_0 = \bar{0},\ \bar{v}_1 = y_1\bar{u}_1, \cdots,\ \bar{v}_n = y_1\bar{u}_1 + y_2\bar{u}_2 + \cdots + y_n\bar{u}_n$.

$$\begin{aligned} f(\bar{c}+\bar{v}) - f(\bar{c}) =& (f(\bar{c}+\lambda\bar{v}_n) - f(\bar{c}+\lambda\bar{v}_{n-1})) \\ &+ (f(\bar{c}+\lambda\bar{v}_{n-1}) - f(\bar{c}+\lambda\bar{v}_{n-2})) \\ &+ \cdots + (f(\bar{c}+\lambda\bar{v}_1) - f(\bar{c}+\lambda\bar{v}_0)) \\ &= \sum_{k=1}^{n} f(\bar{c}+\lambda\bar{v}_k) - f(\bar{c}+\lambda\bar{v}_{k-1}) \\ &= \sum_{k=1}^{n} f(\bar{c}+\lambda\bar{v}_{k-1} + \lambda y_k\bar{u}_k) - f(\bar{c}+\lambda\bar{v}_{k-1}) \end{aligned}$$

Step 3 : Mean-value theorem

Define $\bar{b}_k = \bar{c} + \lambda\bar{v}_{k-1}$. Then we have

$$f(\bar{c}+\bar{v}) - f(\bar{c}) = \sum_{k=1}^{n} f(\bar{b}_k + \lambda y_k\bar{u}_k) - f(\bar{b}_k) \tag{13.8}$$

We know that all partial derivatives exists in $B(\bar{c})$. Therefore by 1-dimensional mean-value theorem we have,

$$f(\bar{b}_k + \lambda y_k\bar{u}_k) - f(\bar{b}_k) = \lambda y_k D_k f(\bar{a}_k) \text{ where } \bar{a}_k \in L(\bar{b}_k, \bar{b}_k + \lambda y_k\bar{u}_k)$$

$$f(\bar{c}+\bar{v}) - f(\bar{c}) = \lambda \sum_{k=1}^{n} y_k D_k f(\bar{a}_k) \text{ where } \bar{a}_k \in L(\bar{b}_k, \bar{b}_k + \lambda y_k\bar{u}_k) \tag{13.9}$$

Step 4 : Continuity of partial derivatives in $B(\bar{c})$

As $\lambda \to 0$, $\bar{v} \to \bar{0}$. And both $\bar{b}_k$, $\bar{b}_k + \lambda y_k\bar{u}_k \to \bar{c}$. Clearly, $\bar{a}_k$ in the line between $\bar{b}_k$ and $\bar{b}_k + \lambda y_k\bar{u}_k$ also converges to $\bar{c}$.

For $k \geq 2$, $D_k f$ are continuous in the n-ball $B(\bar{c})$. Thus $D_k f(\bar{a}_k) \to D_k f(\bar{c})$. We may write, $D_k f(\bar{a}_k) = D_k f(\bar{c}) + E_k(\lambda)$ where $E_k(\lambda) \to \bar{0}$ as $\lambda \to 0$. Also, since $D_1 f$ exists, $D_1 f(\bar{c} + \lambda y_1 \bar{u}_1) \to D_1 f(\bar{c})$ as $\lambda \to 0$.

$$\text{Remember : } D_1 f(\bar{c}) = \lim_{h \to 0} \frac{f(\bar{c} + h\bar{u}_1) - f(\bar{c})}{h}$$

$$\begin{aligned} f(\bar{c} + \bar{v}) - f(\bar{c}) =& \lambda \sum_{k=1}^{n} y_k D_k f(\bar{c}) + \lambda \sum_{k=1}^{n} y_k E_k f(\lambda) \\ =& \nabla f(\bar{c}) \cdot \bar{v} + \|\bar{v}\| E(\lambda) \\ & \text{where } E(\lambda) = \sum_{k=1}^{n} y_k E_k(\lambda) \to \bar{0} \text{ as } \bar{v} \to \bar{0} \end{aligned}$$

That is, we have a linear function which satisfies first-order Taylor's formula at $\bar{c}$. Therefore, f is differentiable. □

13.3.2 Sufficient conditions for the equality of mixed partial derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}^m$. Then $D_r f$ and $D_k f$ are two partial derivatives of f. And $D_{r,k} f = D_r(D_k f)$ and $D_{k,r} f = D_k(D_r f)$.

$$D_{r,k} f = \frac{\partial^2 f}{\partial x_r \partial x_k} = \frac{\partial}{\partial x_r} \frac{\partial f}{\partial x_k} \text{ and } D_{k,r} f = \frac{\partial^2 f}{\partial x_k \partial x_r} = \frac{\partial}{\partial x_k} \frac{\partial f}{\partial x_r}$$

There are two sufficient conditions for the equality of these mixed partial derivatives in our scope. 1. differentiability of $D_k f$ or 2. continuity of $D_{r,k} f$ and $D_{k,r}$ at $\bar{c}$ where the mixed partial derivatives are to be equal.

Differentiability

Theorem 13.3.2. *Suppose $D_r f$ and $D_k f$ exists in an n-ball about $\bar{c}$ and are both differentiable at $\bar{c}$. Then $D_{r,k} f = D_{k,r} f$.*

Proof. Step 1 : Real-valued function
It is sufficient to prove the theorem for real-valued functions. Let $f : \mathbb{R}^n \to \mathbb{R}^m$, then $f(\bar{c}) = (f_1(\bar{c}), f_2(\bar{c}), \cdots, f_m(\bar{c}))$. And

$$D_k f(\bar{c}) = (D_k f_1(\bar{c}), D_k f_2(\bar{c}), \cdots, D_k f_m(\bar{c}))$$

Thus it is sufficient to prove that $D_{r,k} f_j(\bar{c}) = D_{k,r} f_j(\bar{c}),\ j = 1, 2, \cdots, m$. That is, it is sufficient to prove equality of mixed partial derivatives of a real-valued function $f_j : \mathbb{R}^n \to \mathbb{R}$. Also, we will prove it for $n = 2$ and $\bar{c} = (0, 0)$. Now, we will prove the theorem of a real-valued function $f : \mathbb{R}^n : \mathbb{R}$.
Step 2 : $\nabla(h)$

Let $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that the partial derivatives $D_k f$, $D_r f$ exist in the n-ball $B(n)$. And let $h > 0$ such that the rectangle with vertices $(0, 0), (0, h), (h, 0), (h, h)$ lies in $B(n)$.

Suppose $n = 3$, $c = (x, y, z)$, and we want to prove equality of $D_{2,3} f$ and $D_{3,2} f$. Then we will consider the rectangle with vertices $(x, y, z), (x, y, z +$

$h), (x, y+h, z), (x, y+h, z+h)$. Again, we are taking $n = 2$ and $\bar{c} = (0, 0)$, only for the ease of notation as the same proof is applicable for any finite natural number, n and any vector $\bar{c} \in \mathbb{R}^n$.

$$\text{Define } \nabla(h) = f(h, h) - f(h, 0) - f(0, h) + f(0, 0) \tag{13.10}$$

Step 3 : $D_{1,2}f = \frac{\nabla(h)}{h^2} = D_{2,1}f$

$$\text{Define } G(x) = f(x, h) - f(x, 0) \tag{13.11}$$

Then we have, $\nabla(h) = G(h) - G(0)$ and $G'(x) = D_1 f(x, h) - D_1 f(x, 0)$. By 1-dimensional mean value theorem,

$$\begin{aligned} G(h) - G(0) =& hG'(x_1) \text{ where } x_1 \in (0, h) \\ =& h\left(D_1 f(x_1, h) - D_1 f(x_1, 0)\right) \end{aligned}$$

We have $D_1 f$ is differentiable at $(0, 0)$. There exists linear, total derivative function $(D_1 f)'(0, 0)$ where $(D_1 f)'(0, 0)(x, y) = \nabla D_1 f(0, 0) \cdot (x, y)$ satisfiying first-order Taylor's formula at $(0, 0)$.

$$\text{Remember } : f'(\bar{c})(\bar{v}) = \sum_{k=1}^{n} v_k D_k f(\bar{c}) = \nabla f(\bar{c}) \cdot \bar{v}$$

$$\begin{aligned} D_1 f((0, 0) + (x_1, h)) =& D_1 f(0, 0) + \nabla D_1 f(0, 0) \cdot (x_1, h) + \|(x_1, h)\| E_1(h) \\ & \text{where } E_1(h) \to 0 \text{ as } h \to 0 \\ D_1 f(x_1, h) =& D_1 f(0, 0) + x_1 D_{1,1} f(0, 0) + h D_{2,1} f(0, 0) + \left|\sqrt{x_1^2 + h^2}\right| E_1(h) \end{aligned}$$

Similarly,

$$\begin{aligned} D_1 f((0, 0) + (x_1, 0)) =& D_1 f(0, 0) + \nabla D_1 f(0, 0) \cdot (x_1, 0) + \|(x_1, 0)\| E_2(h) \\ & \text{where } E_2(h) \to 0 \text{ as } h \to 0 \\ D_1 f(x_1, 0) =& D_1 f(0, 0) + x_1 D_{1,1} f(0, 0) + |x_1| E_2(h) \end{aligned}$$

Therefore $\nabla(h) = h(D_1 f(x_1, h) - D_1 f(x_1, 0)) = h^2 D_{2,1} f(0, 0) + E(h)$ where $E(h) = h|\sqrt{x_1^2 + h^2}|E_1(h) - h|x_1|E_2(h)$ and $E(h) \to 0$ as $h \to 0$.

Since $0 < x_1 < h$, we have

$$0 \le E(h) \le h^2 \left(\sqrt{2} E_1(h) - E_2(h)\right)$$

Therefore,

$$\lim_{h \to 0} \frac{\nabla(h)}{h^2} \le \lim_{h \to 0} \frac{h^2 D_{2,1} f(0, 0)}{h^2} + \lim_{h \to 0} \frac{h^2(\sqrt{2} E_1(h) - E_2(h))}{h^2} = D_{2,1} f(0, 0)$$

$$\lim_{h \to 0} \frac{\nabla(h)}{h^2} = D_{2,1} f(0, 0) \tag{13.12}$$

You may skip the following part. And conclude with the last two lines.

Similarly, define $H(y) = f(h, y) - f(0, y)$. Then we have, $\nabla(h) = H(h) - H(0)$ and $H'(y) = D_2f(h, y) - D_2f(0, y)$. By 1-dimensional mean value theorem,

$$\begin{aligned} H(h) - H(0) =& hH'(y_1) \text{ where } y_1 \in (0, h) \\ =& h\left(D_2f(h, y_1) - D_2f(0, y_1)\right) \end{aligned}$$

We have D_2f is differentiable at $(0, 0)$. Thus there exists a linear, total derivative function $(D_2f)'(0, 0)$ where $(D_2f)'(0, 0)(x, y) = \nabla D_2f(0, 0) \cdot (x, y)$ satisfying first-order Taylor's formula at $(0, 0)$. That is,

$$\begin{aligned} D_2f((0,0) + (h, y_1)) =& D_2f(0, 0) + \nabla D_2f(0, 0) \cdot (h, y_1) + \|(h, y_1)\|E_3(h) \\ & \text{where } E_3(h) \to 0 \text{ as } h \to 0 \\ D_2f(h, y_1) =& D_2f(0, 0) + hD_{1,2}f(0, 0) + y_1D_{2,2}f(0, 0) + \left|\sqrt{h^2 + y_1^2}\right| E_3(h) \end{aligned}$$

Again,

$$\begin{aligned} D_2f((0, 0) + (0, y_1)) =& D_2f(0, 0) + \nabla D_2f(0, 0) \cdot (0, y_1) + |y_1|E_4(h) \\ & \text{where } E_4(h) \to 0 \text{ as } h \to 0 \\ D_2f(0, y_1) =& D_2f(0, 0) + y_1D_{2,2}f(0, 0) + |y_1|E_4(h) \end{aligned}$$

Therefore, $\nabla(h) = h(D_2f(h, y_1) - D_2f(0, y_1)) = h^2D_{1,2}f(0, 0) + E'(h)$ where $E'(h) = \left|\sqrt{h^2 + y_1^2}\right| E_3(h) - |y_1|E_4(h)$ and $E'(h) \to 0$ as $h \to 0$. And

$$\lim_{h\to 0} \frac{\nabla(h)}{h^2} = D_{1,2}f(0, 0) \tag{13.13}$$

Therefore, $D_{1,2}f(0, 0) = D_{2,1}f(0, 0)$. □

Continuity

Theorem 13.3.3. *Suppose D_rf and D_kf exists in an n-ball about $\bar{c}$. And $D_{r,k}f$ and $D_{k,r}f$ are continuous at $\bar{c}$. Then $D_{r,k}f = D_{k,r}f$.*

Proof. We have $D_rf = (D_rf_1, D_rf_2, \cdots, D_rf_m)$. Therefore, it is sufficient to prove the theorem for real-valued functions. Suppose $n = 2$, $\bar{c} = (0, 0)$ and the partial derivatives D_1f and D_2f exist and are continuous in some 2-ball about $(0, 0)$. Suppose (h, h) lies in that 2-ball, then $D_1f(h, h) \to D_1f(0, 0)$ as $h \to 0$. □

Remark. A function f such that $D_{1,2}f \neq D_{2,1}f$.

$$\text{Let, } f(x, y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} & (x, y) \neq (0, 0) \\ 0 & (x, y) = (0, 0) \end{cases}$$

$$\begin{aligned} D_1f(x, y) =& \frac{\partial}{\partial x} \frac{x^3y - xy^3}{x^2 + y^2} \\ =& \frac{(3x^2y - y^3)(x^2 + y^2) - 2x(x^3y - xy^3)}{(x^2 + y^2)^2} \\ =& \frac{x^4y + 4x^2y^3 - y^5}{(x^2 + y^2)^2} \end{aligned}$$

$$D_1f_{(x=0)} = -y \implies D_{2,1}f_{(x=0)} = \frac{\partial}{\partial y} - y = -1$$

$$\begin{aligned}
D_2 f(x,y) =& \frac{\partial}{\partial y} \frac{x^3 y - x y^3}{x^2+y^2} \\
=& \frac{(x^3 - 3xy^2)(x^2+y^2) - 2y(x^3y - xy^3)}{(x^2+y^2)^2} \\
=& \frac{x^5 - 4x^3y^2 - xy^4}{(x^2+y^2)^2} \\
D_2 f_{(y=0)} =& x \implies D_{1,2} f_{(y=0)} = \frac{\partial}{\partial x} x = 1
\end{aligned}$$

Therefore, $D_{1,2}f \neq D_{2,1}f$ in the neighbourhood of $(0,0)$. This treatment save a lot of time. After $D_1 f$, we are planning to perform $D_{2,1}f = D_2(D_1 f)$ in which the value of x is going to be treated as a constant. Therefore, we can simplify the expression by substituting $x = 0$ at this stage. If you are not confident enough to substitute that "early". You may take partial derivative with respect to y and then substitute $x = 0$ and $y = 0$. Why don't we substitute $y = 0$ before $D_2 f$ is something you should know already !

13.3.3 Implicit Functions and Extremum Problems

Definitions 13.3.1 (Implicit function)**.** Let f be a function. Consider the equation, $f(\bar{x}, \bar{y}) = \bar{0}$. If there exists a function g such that $\bar{x} = g(\bar{y})$, then g is an **implicit form** of f or g is defined implicitly by f. For example, a linear system of equations $Ax - b = 0$ implicitly defines $x = A^{-1}b$ provided A has non-zero determinant.

Definitions 13.3.2 (Jacobian Determinant)**.** Let $f : \mathbb{R}^n \to \mathbb{R}^n$, then determinant of the Jacobian matrix $Df(\bar{x})$ is the **Jacobian determinant** of f, $J_f(\bar{x})$.

What is an implicit function ?

Consider the equation $\sin(x+y) = \cos(u+v)$. We can rewrite this equation as $\sin(x+y) - \cos(u+v) = 0$. Now, the equation is of the form $f(x,y,u,v) = 0$. Therefore, we have the function $f : \mathbb{R}^4 \to \mathbb{R}$ defined by $f(x,y,u,v) = \sin(x+y) - \cos(u+v)$.

Let us play around,

$$\begin{aligned}
\sin(x+y) = \cos(u+v) &\implies x + y = \sin^{-1}\cos(u+v) \\
&\implies (x,y) = (\sin^{-1}\cos(u+v) - k, k) \\
&\implies (x,y) = g(u,v)
\end{aligned}$$

where $g_1(u,v) = \sin^{-1}\cos(u+v) - k$, and $g_2(u,v) = k$

Now we have defined a new function $g : \mathbb{R}^2 \to \mathbb{R}^2$ from the function f. Therefore, we could say that f defines g implicitly. Now we have a few questions to ask about the nature of such implicit functions (or implicitly defined functions),

1. Does there always exists a function for any such combination ?
   That is, does there exists a function h such that $(x,u) = h(y,v)$?

2. Suppose $f(x, y, u, v) = 0$. And function $h : \mathbb{R}^2 \to \mathbb{R}^2$ is implicitly defined by f. Does there always exists a neighbourhood of (u, v) in which h has continuous partial derivatives ?

3. What are the properties of f for these implicit functions to have nice properties ?

It turns out that non-zero Jacobian determinant is nice property f can have.

Theorem 13.3.4. *Let $f : \mathbb{C} \to \mathbb{C}$. Then $J_f(z) = |f'(z)|^2$.*

Proof. Suppose $f : \mathbb{C} \to \mathbb{C}$ where $f(z) = u(z) + iv(z)$ where $u : \mathbb{C} \to \mathbb{R}$ and $v : \mathbb{C} \to \mathbb{R}$. These real-valued functions u, v have respective u^*, v^* multivariate real functions such that $u^* : \mathbb{R}^2 \to \mathbb{R}$, where $u(z) = u^*(x, y)$ and $z = x + iy$. Let $f(z) = z^2 + 1$. Then $u^*(x, y) = x^2 - y^2 + 1$ and $v^*(x, y) = -2xy$. And theoretically we use derivatives of u^* when we mention derivatives of u.

Then f has a derivative at z only if the partial derivatives D_1u, D_2u, D_1v, D_2v exists at z and satisfies Cauchy-Riemann equations. ie $D_1u(z) = D_2v(z)$ and $D_1v(z) = -D_2u(z)$.[Apostol, 1973, Theorem 5.22].

Thus we have $f'(z) = D_1u + iD_1v$ [Apostol, 1973, Theorem 12.6][1].

$$\begin{aligned} f'(z) &= D_1u(z) + iD_1v(z) \\ |f'(z)|^2 &= (D_1u(z))^2 + (D_1v(z))^2 \end{aligned}$$

For ease of representation, we write D_1u instead of $D_1u(z)$

$$|f'(z)|^2 = (D_1u)^2 + (D_1v)^2$$

We also have

$$J_f(z) = |Df(z)| = \begin{vmatrix} D_1u & D_2u \\ D_1v & D_2v \end{vmatrix} = D_1uD_2v - D_1vD_2u = (D_1u)^2 + (D_1v)^2$$

Therefore, $J_f(z) = |f'(z)|^2$. □

Functions with non-zero Jacobian determinant

That is, $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $J_f \neq 0$ in an n-ball. In other words, we have an n-ball $B(\bar{x})$ such that $J_f(\bar{y}) \neq 0$, $\forall \bar{y} \in B(\bar{x})$.

Theorem 13.3.5. *Let B be an n-ball about $\bar{a}$ in $\mathbb{R}^n$, ∂B be its boundary and $\bar{B} = B \cup \partial B$ be its closure.[2] Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be continuous in $\bar{B}$ and all partial derivatives, $D_jf_i(\bar{x})$ exists for every $\bar{x} \in B$. Let $f(\bar{x}) \neq f(\bar{a})$ for every $\bar{x} \in \partial B$ and $J_f(\bar{x}) \neq 0$ for every $\bar{x} \in B$. Then $f(B)$ contains an n-ball about $f(\bar{a})$.*

$$\begin{aligned} B &= \{\bar{x} : \|\bar{x} - \bar{a}\| < r\} \\ \partial B &= \{\bar{x} : \|\bar{x} - \bar{a}\| = r\} \\ \bar{B} &= \{\bar{x} : \|\bar{x} - \bar{a}\| \leq r\} \end{aligned}$$

---

[1] Prove using first-order Taylor's formula

[2] $\bar{B}$: The line above B has a different meaning compare to $\bar{a}$ (situations like this are an abuse of language).

Proof. Define $g : \partial B \to \mathbb{R}$ where $g(\bar{x}) = \|f(\bar{x}) - f(\bar{a})\|$. We have, $f(\bar{x}) \neq f(\bar{a})$ for every $\bar{x} \in \partial B$, thus $g(\bar{x}) > 0$ for every $\bar{x} \in \partial B$. Function f is continuous on $\bar{B}$, thus g is continuous on $\bar{B}$ and thus g is continuous on its subset ∂B. Since ∂B is compact, every continuous function on ∂B attains its extrema[3] and thus g attains its minimum value $m > 0$ somewhere on ∂B.

Consider n-ball T about $f(\bar{a})$ with radius $\frac{m}{2}$,

$$T = B\left(f(\bar{a}), \frac{m}{2}\right) = \left\{\bar{y} \in \mathbb{R}^n : \|f(\bar{a}) - \bar{y}\| < \frac{m}{2}\right\}$$

Therefore, it is sufficient to prove that $T \subset f(B)$.

Let $\bar{y} \in T$. Define $h : \bar{B} \to \mathbb{R}$ where $h(\bar{x}) = \|f(\bar{x}) - \bar{y}\|$. Again this continuous function h on compact set $\bar{B}$ attains its extrema somewhere on $\bar{B}$. Since $\bar{y} \in T$, $h(\bar{a}) = \|f(\bar{a}) - \bar{y}\| < \frac{m}{2}$. Thus, the minimum of h on $\bar{B}$ is less than $\frac{m}{2}$, since $\bar{a} \in \bar{B}$.

Let $\bar{x} \in \partial B$, then

$$\begin{aligned} h(\bar{x}) =& \|f(\bar{x}) - \bar{y}\| \\ =& \|f(\bar{x}) - f(\bar{a}) + f(\bar{a}) - \bar{y}\| \\ \geq& \|f(\bar{x}) - f(\bar{a})\| + \|f(\bar{a}) - \bar{y}\| \\ =& g(\bar{x}) - h(\bar{a}) \\ >& \frac{m}{2} \text{ since } g(\bar{x}) \geq m \text{ and } h(\bar{a}) < \frac{m}{2} \end{aligned}$$

Thus h doesn't attain its minimum on ∂B, but at an interior point $\bar{c} \in B$. Consider

$$h^2(\bar{x}) = \|f(\bar{x}) - \bar{y}\|^2 = \sum_{r=1}^{n} (f_r(\bar{x}) - y_r)^2$$

The function h^2 also has minimum at the same point $\bar{c}$. Thus all partial derivatives of h^2 at $\bar{c}$ are zero. ie,

$$D_k h^2(\bar{c}) = \sum_{r=1}^{n} (f_r(\bar{c}) - y_r) D_k f_r(\bar{c}) = 0$$

This is a system of linear equations with non-zero determinant since $\bar{c} \in B$ and we have $J_f(\bar{c}) \neq 0$. Therefore, $f_r(\bar{c}) = y_r$. That is, $f(\bar{c}) = \bar{y} \in f(B)$. Since $\bar{y} \in T$ is arbitrary, $T \subset f(B)$. □

Theorem 13.3.6. *Let A be an open subset of $\mathbb{R}^n$ and $f : A \to \mathbb{R}^n$ is continuous and has continuous partial derivatives $D_j f_i$ on A. If f is one-to-one on A and $J_f(\bar{x}) \neq 0$, $\forall \bar{x} \in A$, then $f(A)$ is open.*

Proof. Let $\bar{b} \in f(A)$. Then $\bar{b} = f(\bar{a})$ for some $\bar{a} \in A$. We have, f is continuous, f has continuous partial derivatives on A and $J_f(\bar{x}) \neq 0$ for every $\bar{x} \in A$. Therefore, there exists an open ball $B \subset A$ containing $\bar{a}$ such that $f(B) \subset f(A)$ contains an n-ball about $f(\bar{a})$. Since $\bar{b} \in f(A)$ is arbitary, every point in $f(A)$ has an n-ball containing it in $f(A)$. Therefore, $f(A)$ is open.

Two assumption in above theorem are trivial. 1. f is continuous in the closed ball, $\bar{B}$. Set B so chosen that $\bar{B} \subset A$ and f is continuous in A.

[3] "Every continuous function on a compact set attains its extrema"

Thus, f is continuous in $\bar{B}$. 2. f has different value at boundary compared to center. ie, $f(\bar{a}) \neq f(\bar{x})$, $\forall x \in \partial B$. We have, f is injective on A, and $\bar{B} \subset A$. Thus f has different values for any two distinct points in it. Thus, $\forall \bar{x}, \bar{y} \in \bar{B}$, $\bar{x} \neq \bar{y} \implies \bar{x}, \bar{y} \in A$, and $\bar{x} \neq \bar{y} \implies f(\bar{x}) \neq f(\bar{y})$ ☐

Theorem 13.3.7. *Let S be an open subset of $\mathbb{R}^n$ and $f : S \to \mathbb{R}^n$. Let components of f has continuous partial derivatives on S, $D_j f_i$ and $J_f(\bar{a}) \neq 0$ for some point $\bar{a} \in S$. Then there is an n-ball B about $\bar{a}$ on which f is injective.*

Proof. Let $\bar{z} = (\bar{z}_1, \bar{z}_2, \cdots, \bar{z}_n)$ where $\bar{z}_i \in \mathbb{R}^n$. ie, $\bar{z} \in \mathbb{R}^{n^2}$. Define function $h : \mathbb{R}^{n^2} \to \mathbb{R}$ by $h(\bar{z}) = \det[D_j f_i(\bar{z}_i)]$. Since f has continuous partial derivatives on S, each component of f has continuous partial derivatives in S and thus h is continuous on S^n which is a subset of $\mathbb{R}^{n^2}$ since S is an open subset of $\mathbb{R}^n$.

Let $\bar{a} \in S$ such that $J_f(\bar{a}) \neq 0$. Existence of such a point in S is assumed. Consider,$\bar{z}_i = \bar{a}$, $\forall i$. Then $\bar{z} = (\bar{a}, \bar{a}, \cdots, \bar{a})$. And $h(\bar{z}) = \det[D_j f_i(\bar{a})] = J_f(\bar{a}) \neq 0$.

Since h is continous and $h(\bar{z}) \neq 0$. There exists an n-ball B about $\bar{a}$ in S such that $h(\bar{z}) \neq 0$ for $\bar{z}_i \in B$. We claim that f is injective on B.

Suppose f is not injective. ie, There exists $\bar{x}, \bar{y} \in B(\bar{a})$ such that $\bar{x} \neq \bar{y}$ and $f(\bar{x}) = f(\bar{y})$. Open ball $B(\bar{a})$ is a convex set. And the line segment $L(\bar{x}, \bar{y}) \subset B(\bar{a})$. The function f is differentiable on S. On applying mean-value theorem to each component of f, we get

$$0 = f_i(\bar{y}) - f_i(\bar{x}) = \nabla f_i(\bar{Z}_i) \cdot (\bar{y} - \bar{x}),\ i = 1, 2, \cdots$$

where $\bar{Z}_i \in L(\bar{x}, \bar{y}) \subset B(\bar{a})$. Therefore, We have

$$\sum_{k=1}^{n} D_k f_i(\bar{Z}_i)(y_k - x_k) = 0$$

The determinant of this system of linear equations is nonzero, as the function f has nonzero jacobian determinant at $\bar{Z}_i \in B(\bar{a})$ for $i = 1, 2, \cdots$. Thus, $y_i = x_i$ for $i = 1, 2, \cdots$. This contradicts $\bar{x} \neq \bar{y}$. Hence, the function f is injective. ☐

Theorem 13.3.8. *Let A be an open subset of $\mathbb{R}^n$ and assume that $f : A \to \mathbb{R}^n$ has continuous partial derivatives $D_j f_i$ on A. If $J_f(\bar{x}) \neq 0$ for all $\bar{x} \in A$, then f is an open mapping.*

Proof. Let S be an open subset of A. Let $\bar{x} \in S$. Clearly, f has continuous partial derivatives on S and $J_f(\bar{x}) \neq 0$ for all $\bar{x} \in S$. Thus, there is an n-ball $B(\bar{x})$ in which f is injective. Therefore, $f(B(\bar{x})$ is open in $\mathbb{R}^n$. Since $\bar{x} \in S$ is arbitrary, $S = \cup_{\bar{x} \in S} B(\bar{x})$. And $f(S) = \cup_{\bar{x} \in S} f(B(\bar{x}))$. Therefore, $f(S)$ is open. Since open set S is arbitrary, f is an open mapping. ☐

Remark (Properties). Functions with non-zero Jacobian determinant has following properties :

1. If $J_f \neq 0$ in n-ball B about $\bar{a}$ which has different values at its boundaries, then $f(B)$ has an n-ball about $f(\bar{a})$.

2. If $J_f \neq 0$, f has continuous partial derivatives in S, and f is injective in an open set A, then $f(A)$ is open.

3. Let S be an open set in $\mathbb{R}^n$, f has continuous partial derivatives in S, and $J_f(\bar{a}) \neq 0$ for some $\bar{a} \in S$, then f is injective on an n-ball $B(\bar{a})$ in S.

4. Let A be an open set in $\mathbb{R}^n$, f has continuous partial derivatives in A, and $J_f \neq 0$ in A, then f is an open mapping.

Inverse function Theorem

Theorem 13.3.9 (Inverse function)**.** *Let S be an open subset of $\mathbb{R}^n$ and f be a continuously differentiable function*[4] *$f : S \to \mathbb{R}^n$. If $J_f(\bar{a}) \neq 0$ for some $\bar{a} \in S$, then there are two open sets $X \subset S$, and $Y \subset f(S)$ such that*

1. *$\bar{a} \in X$ and $f(\bar{a}) \in Y$*
2. *$Y = f(X)$*
3. *f is injective*
4. *there exists another function $g : Y \to X$ such that $g(f(\bar{x})) = \bar{x}, \ \forall \bar{x} \in X$*
5. *g is continuously differentiable on Y*

In other words, if $f \in C'$ and there exists $\bar{a} \in S$ such that $J_f(\bar{a}) \neq 0$, then f has an inverse f^{-1} in a neighbourhood of $f(\bar{a})$ and $f^{-1} \in C'$.

Proof. Step 1 : Construction of open sets X and Y.
Given that, $J_f(\bar{a}) \neq 0$ and $f \in C'$. Thus all partial derivatives of f are continuous on S. Then J_f is continuous on S, By the continuity of J_f at $\bar{a}$, there exists a neighbourhood of $\bar{a}$, say $B_1(\bar{a})$ in which $J_f \neq 0$. That is, $\forall \bar{x} \in B_1(\bar{a}), \ J_f(\bar{x}) \neq 0$. Therefore,(by theorem) there exists an n-ball $B(\bar{a})$ on which f is injective. Let B be an n-ball with center $\bar{a}$ contained in $B(\bar{a})$. Then f is injective on B. Therefore,(by theorem) $f(B)$ contains an n-ball with center $f(\bar{a})$. Let Y be the n-ball contained in $f(B)$. And $X = f^{-1}(Y) \cap B$. That is, the inverse image of Y on B. Since f is continuous, $f^{-1}(Y)$ is open. Thus, X is an intersection of open sets. And therefore, X is open.
Step 2 : The inverse of f, say g.
Clearly $\bar{a} \in X$ and $f(\bar{a}) \in Y$. Also $Y = f(X)$ and f is injective on X(since, $X \subset B$).
The closure of B, $\bar{B}$ is compact and f is injective and continuous on $\bar{B}$. Then[5] there exists a continuous function g defined on $f(\bar{B})$ such that $g \circ f$ is the identity function on $\bar{B}$. That is, $\forall x \in \bar{B}, \ g(f(\bar{x})) = x$. Thus, $g(X) = Y$ and g is unique.
Step 3 : g has continuous partial derivatives.
Define a real-valued function $h : S^n \to \mathbb{R}$ by $h(\bar{Z}) = \det[D_j f_i(\bar{Z}_i)]$ where $\bar{Z}_1, \bar{Z}_2, \cdots, \bar{Z}_n \in S$ and $\bar{Z} = (\bar{Z}_1, \bar{Z}_2, \cdots, \bar{Z}_n)$. Now, let $\bar{Z} = (\bar{a}, \bar{a}, \cdots, \bar{a})$. Then $h(\bar{Z}) \neq 0$ and h is continuous on S^n. Therefore, $\bar{Z}$ has a neighbourhood on which h does not vanish (that is, nonzero). Let $B_2(\bar{a})$ be the corresponding n-ball with center $\bar{a}$ such that $\bar{Z}_i \in B_2(\bar{a}) \implies h(\bar{Z}) \neq 0$.
Let B be an n-ball with center $\bar{a}$ contained in $B_2(\bar{a})$. Now $\bar{B} \subset B_2(\bar{a})$. And $h(\bar{Z}) \neq 0, \ \forall \bar{Z}_i \in \bar{B}$.

---
[4] $f \in C'(S)$: f is continuously differentiable on S
[5] Existence of inverse of a continuous function on a compact set in metric spaces.

We have, $g = (g_1, g_2, \cdots, g_n)$. It is enough to prove that $g_k \in C'$ for $k = 1, 2, \cdots, n$. Again, it is enough to prove that $D_r g_k$ exists and is continuous for $1 \le r \le n$. (Fix some r and prove that $D_r g_k$ is continuous.)

Let $\bar{y} \in Y$. Define $\bar{x} = g(\bar{y})$ and $\bar{x}' = g(\bar{y} + t\bar{u}_r)$ where t is sufficiently small such that $\bar{y} + t\bar{u}_r \in Y$. Then $\bar{x}, \bar{x}' \in X$. And $f(\bar{x}') - f(\bar{x}) = t\bar{u}_r$. Therefore $f_i(\bar{x}) - f_i(\bar{x}') = 0$ when $i \neq r$. And $f_i(\bar{x}') - f_i(\bar{x}) = t$ when $i = r$. By mean-value theorem,

$$\frac{f_i(\bar{x}') - f(\bar{x})}{t} = \nabla f_i(\bar{Z}_i) \cdot \frac{\bar{x}' - \bar{x}}{t}$$

where $\bar{Z}_i \in L(\bar{x}, \bar{x}')$, the line segment joining $\bar{x}$ and $\bar{x}'$. Since $\det[D_j f_i(\bar{Z}_i)] = h(\bar{Z}) \neq 0$, this system of linear equations in n unknowns, $\frac{x_j' - x_j}{t}$ has a unique solution. As $t \to 0$, $\bar{x}' \to \bar{x}$. And $\bar{Z}_i \to \bar{x}$. Since $J_f(\bar{x}) \neq 0$, the limit

$$\lim_{t \to 0} \frac{g_k(\bar{y} + t\bar{u}_r) - g_k(\bar{y})}{t}$$

exists. Thus, $D_r g_k(\bar{y})$ exists $\forall y \in Y$ and every r. By Cramer's rule, this limit is a quotient of two determinants of partial derivatives of f, which are all continuous since $f \in C'$. Therefore, $D_r g_k$ are all continuous and $g \in C'$. $\square$

Implicit function Theorem

Theorem 13.3.10. *Let S be an open set in $\mathbb{R}^{n+k}$ and f be a function $f : S \to \mathbb{R}^n$. Suppose f is continuously differentiable on S. Let $(\bar{x}_0, \bar{t}_0) \in S$ such that $\bar{x}_0 \in \mathbb{R}^n$, $\bar{t}_0 \in \mathbb{R}^k$, $f(\bar{x}_0, \bar{t}_0) = \bar{0}$ and $J_f(\bar{x}_0, \bar{t}_0) \neq 0$. Then there exists an open set T_0 containing $\bar{t}_0$ in $\mathbb{R}^k$ and a unique function $g : T_0 \to \mathbb{R}^n$ such that*

1. *g is continuously differentiable on T_0*
2. *$g(\bar{t}_0) = \bar{x}_0$*
3. *$f(g(\bar{t}), \bar{t}) = \bar{0}, \ \forall \bar{t} \in T_0$*

Proof. Given $f : S \to \mathbb{R}^n$ where $S \subset \mathbb{R}^{n+k}$. We have $f = (f_1, f_2, \cdots, f_m)$. Also given that $f \in C'(S)$, $f(\bar{x}_0; \bar{t}_0) = 0$, and $J_f(\bar{x}_0; \bar{t}_0) \neq 0$. Define a function $F : S \to \mathbb{R}^{n+k}$ defined by $F = (F_1, F_2, \cdots, F_{n+k})$.

$$F_m(\bar{x}; \bar{t}) = \begin{cases} f_m(\bar{x}; \bar{t}) & 1 \le m \le n \\ t_{m-n} & n < m \le n + k \end{cases}$$

For example, let $n = 3$, $k = 2$. Let $\bar{x} = (1, 2, 3)$ and $\bar{t} = (4, 5)$. Suppose $f_k(1, 2, 3, 4, 5) = a_k$. Then $F(\bar{x}; \bar{t}) = (a_1, a_2, a_3, 4, 5)$.

Then the Jacobian determinant of $F'(\bar{x}; \bar{t})$ is given by

$$J_F(\bar{x}; \bar{t}) = \begin{vmatrix} \begin{matrix} D_1 f_1 & D_2 f_1 & \cdots & D_m f_1 \\ D_1 f_2 & D_2 f_2 & \cdots & D_m f_2 \\ \vdots & \vdots & \ddots & \vdots \\ D_1 f_m & D_2 f_m & \cdots & D_m f_m \end{matrix} & \begin{matrix} 0 & \cdots & 0 \end{matrix} \\ \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & \begin{matrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{matrix} \end{vmatrix}$$

Thus $J_F(\bar{x}_0;\bar{t}_0) = J_f(\bar{x}_0;\bar{t}_0) \neq 0$. Also, $F(\bar{x}_0;\bar{t}_0) = (\bar{0};\bar{t}_0)$. Therefore, by inverse function theorem, there exists open sets X, Y containing $(\bar{x}_0;\bar{t}_0)$ and $(\bar{0};\bar{t}_0)$ such that F is injective on X, $X = F^{-1}(Y)$ and there exists a unique local inverse function G such that $G(F(\bar{x};\bar{t})) = (\bar{x};\bar{t})$ and $G \in C'(Y)$.

Let $G = (v;w)$. That is, $v_i = G_i$ for $1 \leq i \leq n$. And $w_j = G_{n+j}$ for $1 \leq j \leq k$. We have, $G(F(\bar{x};\bar{t})) = (\bar{x};\bar{t})$. Therefore, $v(F(\bar{x};\bar{t})) = \bar{x}$ and $w(F(\bar{x};\bar{t})) = \bar{t}$. Since $X \subset F^{-1}(Y)$ and F is one-to-one on X, for every $(\bar{x};\bar{t}) \in Y$, there exists $(\bar{x}';\bar{t}') \in X$ such that $F(\bar{x}';\bar{t}') = (\bar{x};\bar{t})$. By the definition of F, $\bar{t}' = \bar{t}$. Therefore, $v(\bar{x};\bar{t}) = v(F(\bar{x}';\bar{t}) = \bar{x}'$ and $w(\bar{x};\bar{t}) = w(F(\bar{x}';\bar{t}) = \bar{t}$. Now, we have $G : Y \to X$ defined by $G(\bar{x};\bar{t}) = (\bar{x}';\bar{t})$.

Let T_0 be a subset of $\mathbb{R}^k$ defined by $T_0 = \{\bar{t} \in \mathbb{R}^k : (\bar{0};\bar{t}) \in Y\}$. For each $\bar{t} \in T_0$ let $g : T_0 \to \mathbb{R}^n$ is defined by $g(\bar{t}) = v(\bar{0};\bar{t})$. The set T_0 is open. And $g \in C'(T_0)$ since g is constructed from the components of G which has continuous partial derivatives on Y. (ie, $G \in C'(Y)$.)

Clearly, $g(\bar{t}_0) = v(\bar{0};\bar{t}_0) = \bar{x}_0$. And $(\bar{0};\bar{t}) = F(\bar{x}_0;\bar{t}_0)$. Therefore, we have $f(v(\bar{x};\bar{t});\bar{t}) = \bar{x}$. Let $\bar{x} = \bar{0}$, then $f(g(\bar{t});\bar{t}) = \bar{0}$. It is enough to prove that the function g is unique. Suppose $f(g(\bar{t});\bar{t}) = f(h(\bar{t});\bar{t})$. Since f is one-to-one on X, $(g(\bar{t});\bar{t}) = (h(\bar{t});\bar{t})$ for every $\bar{t} \in T_0$. And $g(\bar{t}) = h(\bar{t})$, $\forall \bar{t} \in T_0$. ☐

Extrema of function of one variable

The function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^3$ has derivative $f'(x) = 3x^2$. Thus $f'(0) = 0$. However, 0 is not a local extrema for the function f. Thus, derivative of the function vanishing at point is not sufficient for a local extrema at that point.

Theorem 13.3.11 (sufficient condition for local extrema)**.** *Let $n \geq 1$ and function f has nth partial derivative in open interval (a,b). Suppose for some $c \in (a,b)$,*

$$f'(c) = f''(c) = \cdots = f^{(n-1)}(c) = 0 \text{ and } f^{(n)}(c) \neq 0$$

If n is even,

1. *f has a local minimum at c if $f^{(n)}(c) > 0$*
2. *f has a local maximum at c if $f^{(n)}(c) < 0$*

and If n is odd, there is neither a local minimum nor a local maximum at c.

Proof. We have $f^{(n)}(c) \neq 0$. Thus, there exists an open interval $B(c)$ such that for each $x \in B(c)$, $f^{(n)}(x)$ has the same sign as $f^{(n)}(c)$. By Taylor's theorem, we have

$$f(x) = f(c) + \sum_{k=1}^{n-1} \frac{f^{(k)}(c)}{k!}(x-c)^k + \frac{f^{(n)}(x_1)}{n!}(x-c)^n \tag{13.14}$$

where $x_1 \in L(x,c)$, the line connecting x and c. We have, $f^{(k)}(c) = 0$ for $k = 1, 2, \cdots, n-1$. Thus,

$$f(x) - f(c) = \frac{f^{(n)}(x_1)}{n!}(x-c)^n$$

Case 1 : n is even.

If n is even, then $(x-c)^n > 0$. Therefore, $f(x) - f(c)$ has the same sign as $f^{(n)}(x_1)$. If $f^{(n)}(c) > 0$, then $f^{(n)}(x_1)$ has a positive value and $f(x) - f(c) > 0$ for every $x \in B(c)$. Therefore, f has a local minimum at c. Similarly, if $f^{(n)}(c) < 0$, then $f(x) - f(c) < 0$ for every $x \in B(c)$. Therefore, f has a local maximum at c.

Case 2 : n is odd.

If n is odd, then $(x-c)^n$ takes both positive and negative values. Therefore, $f(x) - f(c)$ has both positive and negative values in $B(c)$. Thus f has neither local minimum nor local maximum at c. □

Extrema of function of several variables

Definitions 13.3.3. If function f is differentiable at $\bar{a}$ and $\nabla f(\bar{a}) = \bar{0}$, the point $\bar{a}$ is a **stationary point** of f.

Definitions 13.3.4. A stationary point is a **saddle point** if every n-ball $B(\bar{a})$ contains points $\bar{x}$ such that $f(\bar{x}) > f(\bar{a})$ and other points such that $f(\bar{x}) < f(\bar{a})$.

Definitions 13.3.5. A function $Q : \mathbb{R}^n \to \mathbb{R}$ defined by $Q(\bar{x}) = \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} a_{ij} x_i x_j$ where $a_{ij} \in \mathbb{R}$ is a function of the **quadratic form** or simply a quadratic form.

symmetric quadratic form $a_{ij} = a_{ji}, \ \forall i, j$.

positive definite quadratic form $Q(\bar{x}) > 0, \forall \bar{x} \neq \bar{0}$.

negative definite quadratic form $Q(\bar{x}) < 0, \forall \bar{x} \neq \bar{0}$.

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$. And second derivative of f exists if the derivative of f, f' is differentiable. Thus, $f'(\bar{a}+h\bar{t}) = f'(\bar{a}) + hf''(\bar{a})(\bar{t}) + |h|\|t\|E_{\bar{a}}(h\bar{t})$ where $E_{\bar{a}} \to \bar{0}$ as $h \to 0$. Writing the Taylor's first order formula using matrices, we can see that $f''(\bar{a})(\bar{t})$ is of the quadratic form.

Theorem 13.3.12. *Let f be a function $f : \mathbb{R}^n \to \mathbb{R}^m$. Suppose that the second order partial derivatives $D_{i,j}f$ exists in an n-ball $B(\bar{a})$ and are continuous at $\bar{a}$ where $\bar{a}$ is a stationary point of f.*

$$\text{Let } Q(\bar{t}) = \frac{1}{2} f''(\bar{a}, \bar{t}) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{i,j} f(\bar{a}) t_i t_j$$

1. *If $Q(\bar{t}) > 0$ for all $\bar{t} \neq \bar{0}$, f has a relative minimum at $\bar{a}$.*

2. *If $Q(\bar{t}) < 0$ for all $\bar{t} \neq \bar{0}$, f has a relative maximum at $\bar{a}$.*

3. *If $Q(\bar{t})$ takes both positive and negative values, then f has a saddle point at $\bar{a}$.*

Proof. Define $Q : \mathbb{R}^n \to \mathbb{R}$ by $Q(\bar{t}) = \frac{1}{2} f''(\bar{a}; \bar{t})$. Then Q is continuous for every $\bar{t} \in \mathbb{R}^n$. Let S be the boundary of the n-ball $B(\bar{0}, 1)$. That is, $S = \{\bar{t} \in \mathbb{R}^n : \|t\| = 1\}$.

Case 1 : Suppose that $Q(\bar{t}) > 0, \ \forall \bar{t} \neq \bar{0}$.

We have, Q is a continuous real-valued function on a compact interval, S. Therefore, Q attains its extrema. Thus Q has a minimum value at point in S, say m. Clearly $Q(\bar{t}) > 0 \implies m > 0$.

By Taylor's formula, $f(\bar{a}+\bar{t}) - f(\bar{a}) = \nabla f(\bar{a}) \cdot \bar{t} + \dfrac{1}{2} f''(\bar{z};\bar{t})$ where $\bar{z} \in L(\bar{a}+\bar{t}, \bar{a})$

For every $\bar{a} \in S$, $\nabla f(\bar{a}) = \bar{0}$. And $f(\bar{a}+\bar{t}) - f(\bar{a}) \to \frac{1}{2} f''(\bar{a};\bar{t})$ as $\bar{t} \to \bar{0}$.

$$\begin{aligned} \text{Thus, } f(\bar{a}+\bar{t}) - f(\bar{a}) &= \frac{1}{2} f''(\bar{a};\bar{t}) + \|t\|^2 E(\bar{t}) \\ &= Q(\bar{t}) + \|t\|^2 E(\bar{t}) \text{ where } E(\bar{t}) \to \bar{0} \text{ as } \bar{t} \to \bar{0} \end{aligned}$$

Let $c = \frac{1}{\|\bar{t}\|}$. Then $c\bar{t} \in S$ and $Q(c\bar{t}) = c^2 Q(\bar{t}) \geq m$. Thus, $Q(\bar{t}) \geq m\|\bar{t}\|^2$. Therefore, $f(\bar{a}+\bar{t}) - f(\bar{a}) \geq m\|\bar{t}\|^2 + \|\bar{t}\|^2 E(\bar{t})$

Choose n-ball $B(\bar{0}, r)$ such that $|E(\bar{t})| < \frac{m}{2}$ for every $\bar{t} \in B(\bar{0}, r)$. Thus,

$$-\frac{m}{2}\|\bar{t}\|^2 \leq -\|\bar{t}\|^2 |E(\bar{t})| \leq 0$$

Therefore, $f(\bar{a}+\bar{t}) - f(\bar{a}) \geq m\|\bar{t}\|^2 - \frac{m}{2}\|\bar{t}\|^2 = \frac{m}{2}\|t\|^2$ for every $\bar{t} \in B(\bar{0}, r)$. Clearly, f has a local minimum at $\bar{a}$.
Case 2 : Suppose $Q(\bar{t}) < 0$, $\forall \bar{t} \neq \bar{0}$, then consider $-f$.

Clearly, function $-f$ has a local minimum at $\bar{t}$. Thus f has a local maximum at $\bar{t}$.
Case 3 : Suppose $Q(\bar{t})$ takes both positive and negative values.

$$\begin{aligned} \text{We have, } f(\bar{a}+\lambda\bar{t}) - f(\bar{a}) &= Q(\lambda\bar{t}) + \lambda^2\|t\|^2 E(\lambda\bar{t}). \\ &= \lambda^2[Q(\bar{t}) + \|\bar{t}\|^2 E(\lambda\bar{t})] \end{aligned}$$

Choose n-ball $B(\bar{0}, r)$ such that $\|\bar{t}\|^2 E(\lambda\bar{t}) < \frac{1}{2}|Q(\bar{t})|$, $\forall \bar{t} \in B(\bar{0}, r)$. We have $\bar{t} \in B(\bar{0}, r) \implies \lambda < r$. Then, error function $E(\bar{t})$ on the RHS is small enough, not to affect the sign of the RHS. Thus $f(\bar{a}+\lambda\bar{t}) - f(\bar{a})$ has the same sign as $Q(\bar{t})$. Therefore, $\bar{a}$ is a saddle point. □

Theorem 13.3.13. *Let f be a real-valued function $f : \mathbb{R}^2 \to \mathbb{R}$ with continuous second order partial derivatives at a stationary point $\bar{a} \in \mathbb{R}^2$. Let $A = D_{1,1}f(\bar{a})$, $B = D_{1,2}f(\bar{a}) = D_{2,1}f(\bar{a})$, and $C = D_{2,2}f(\bar{a})$. And let $\Delta = \begin{vmatrix} A & B \\ B & C \end{vmatrix} = AC - B^2$. Then we have,*

1. *If $\Delta > 0$ and $A > 0$, then f has a relative minimum at $\bar{a}$.*

2. *If $\Delta > 0$ and $A < 0$, then f has a relative maximum at $\bar{a}$.*

3. *if $\Delta < 0$, then f has a saddle point at $\bar{a}$.*

Proof. We have, function $f : \mathbb{R}^2 \to \mathbb{R}$ with $\nabla f(\bar{a}) = (0, 0)$. Consider the quadratic form $Q(x, y) = \frac{1}{2}[Ax^2 + Bxy + Cy^2]$ where $A = D_{1,1}f(\bar{a})$, $B =$

$D_{1,2}f(\bar{a})$, and $C = D_{2,2}f(\bar{a})$. Therefore, $Q(x,y) = \frac{1}{2}f''(\bar{a};\bar{t})$.
Case 1 : Suppose $A \neq 0$.

$$\begin{aligned} Q(x,y) &= \frac{1}{2A}[A^2x^2 + ABxy + ACy^2] \\ &= \frac{1}{2A}[(Ax+By)^2 - B^2y^2 + ACy^2] \\ &= \frac{1}{2A}[(Ax+By)^2 + \Delta y^2] \end{aligned}$$

If $\Delta > 0$, then $Q(x,y)$ has the same sign as A. Therefore, f has a local minimum/maximum at $\bar{a}$ depending on the sign of A.
Case 2 : Suppose $A = 0$. Then $Q(x,y) = \frac{1}{2}[Bxy + Cy^2] = \frac{1}{2}(Bx + Cy)y$.

Now we have two lines in $\mathbb{R}^2$, $y = 0$ and $Bx + Cy = 0$. These two lines divides $\mathbb{R}^2$ into four regions. The value of $Q(x,y)$ is positive in two of those regions and negative in the other two regions. Therefore, $f(\bar{a}+\bar{t}) - f(\bar{a})$ assumes both positive and negative values in any neighbourhood $B(\bar{a},r)$. Therefore, $\bar{a}$ is a saddle point. □

13.4 Integration on Differential Forms

Definitions 13.4.1. A k**-cell** in R^k is given by,

$$I^k = \{\bar{x} \in \mathbb{R}^k : a_i \le x_i \le b_i,\ \forall i\}$$

where $\bar{a}, \bar{b} \in \mathbb{R}^k$. Let f be a continuous, real-valued function on I^k. Then, the **integral** of f over I^k is given by,

$$\int_{I^k} f(\bar{x})d\bar{x} = f_0 \text{ where } f_k = f \text{ and}$$

$$f_{r-1} = \int_{a_r}^{b_r} f_r(x_0, x_1, \cdots, x_r)dx_r,\ \ r = 1, 2, \cdots, k$$

In other words,

$$\int_{I^k} f(\bar{x})\ d\bar{x} = \int \cdots \int_{a_k}^{b_k} [f(x_0, x_1, \cdots, x_k)\ dx_k]\ dx_{k-1} \cdots dx_1$$

Theorem 13.4.1. *For every $f \in \mathscr{C}(I^k)$, $L(f) = L'(f)$.*

In other words, integral of a function over a k-cell is independent of the order in which those k integrations are carried out.

Proof. Step 1 : "Separable" Functions ie, $h(\bar{x}) = \prod h_i(x_i)$.

("separable" is not standard. It is only for the purpose of understanding.)

Let $h(\bar{x}) = h_1(x_1)h_2(x_2)\cdots h_k(x_k)$ where $h_j \in [a_j, b_j]$.

$$L(h) = \int_{I^k} \left(\prod_{i=1}^{k} h_i(x_i)\right) d\bar{x} = \prod_{i=1}^{k} \int_{a_k}^{b_k} h_i(x_i)\ dx_i = L'(h)$$

Step 2 : Algebra of "separable" functions, $\mathscr{A}$.

Let $\mathscr{A}$ be all finite sums of functions such as h. Let $g \in \mathscr{A}$.

$$\begin{aligned} L(g) &= \int_{I^k} \left(\sum_j \prod_i h_{i,j}(x_i) \right) d\bar{x} \\ &= \sum_j \int_{I^k} \prod_i h_{(j)}(\bar{x})\ d\bar{x} \\ &= \sum_j \prod_i \int_{a_k}^{b_k} h_{i,j}(x_i)\ d(x_i) \\ &= L'(g) \end{aligned}$$

Step 3 : All functions continuous on I^k.

Let $f \in \mathscr{C}(I^k)$. ie, a function which is continuous in I^k.

Stone-Weierstrass theorem - Let $\mathscr{A}$ be an algebra of real, continuous functions on a compact set K. If $\mathscr{A}$ separates points on K and if $\mathscr{A}$ vanishes at no point of K, then the uniform closure $\mathscr{B}$ of $\mathscr{A}$ consists of all real, continuous functions on K.

Clearly, the algebra of functions $\mathscr{A}$ separates points on I^k and vanishes nowhere on I^k. Suppose $\bar{x} \neq \bar{y}$, then there exists m such that $x_m \neq y_m$. Thus $h(\bar{x}) = x_m$ separates $\bar{x}$ and $\bar{y}$. Again $h(\bar{x}) = 1$ vanishes nowhere on I^k. Thus every function which is continuous on I^k is the limit of some uniformly convergent sequence of functions in $\mathscr{A}$.

Let $V = \prod_{j=1}^{k}(b_j - a_j)$. Then by Stone-Weierstrass theorem, for any $\epsilon > 0$, there exists a function $g \in \mathscr{A}$ such that $\|f - g\| < \frac{\epsilon}{V}$ where the norm of a function f is defined by $\|f\| = \max\{f(\bar{x}) : \bar{x} \in I^k\}$.

Therefore, it is sufficient to prove that $\|L(f) - L'(f)\| < \epsilon$. Since $\|f - g\| < \frac{\epsilon}{V}$, $|L(f - g)| < \epsilon$ and $|L'(f - g)| < \epsilon$. Thus,

$$\begin{aligned} L(f) - L'(f) &= L(f) - L(g) + L'(g) - L'(f) \\ &= L(f - g) + L'(g - f) \\ |L(f) - L'(f)| &< 2\epsilon \end{aligned}$$

Therefore, $L(f) = L'(f)$. □

Definitions 13.4.2. Let $f : \mathbb{R}^k \to \mathbb{C}$. The **support** of f is the closure of the set of all points $\bar{x} \in \mathbb{R}^k$ such that $f(\bar{x}) \neq 0$.

Remark. Let f be a continuous function with compact support. And I^k be any k-cell containing the support of f. Then, $\int_{\mathbb{R}^k} f\ d\bar{x} = \int_{I^k} f\ d\bar{x}$.

$L(f) = L'(f)$ where f is the limit function of a sequence of functions with compact support. [Apostol, 1973, §10.4 Example]

Definitions 13.4.3. Let E be an open subset in $\mathbb{R}^n$. Then function $G : E \to \mathbb{R}^n$ is **primive** if it satisfies

$$G(\bar{x}) = \sum_{i \neq m} x_i \bar{e}_i + g(\bar{x})\bar{e}_m \tag{13.15}$$

for some integer m and some function $g : E \to \mathbb{R}$ (where $\bar{e}_i$ are the unit co-ordinate vectors).

$$G(\bar{x}) = \bar{x} + [g(\bar{x}) - x_m]\bar{e}_m \tag{13.16}$$

If g is differentiable at $\bar{a}$, then G is also differentiable at $\bar{a}$. The matrix $[\alpha_{i,j}]$ of $G'(\bar{a})$ is given by

$$\alpha_{i,j} = \begin{cases} D_j g(\bar{a}) & i = m \\ 1 & i \neq m, j = i \\ 0 & i \neq m, j \neq i \end{cases}$$

The Jacobian of G at $\bar{a}$ is given by, $J_G(\bar{a}) = \det[G'(\bar{a})] = D_m g(\bar{a})$.

Total derivative $G'(\bar{a})$ is invertible if and only if $D_m g(\bar{a}) \neq 0$.

Definitions 13.4.4. A linear operator B on $\mathbb{R}^n$ that interchanges some pair of members of the standard basis and leaves the others fixed is a **flip**.

Theorem 13.4.2. *Suppose F is a $\mathscr{C}'$-mapping of an open set $E \subset \mathbb{R}^n$ into $\mathbb{R}^n$, $\bar{0} \in E$, $F(\bar{0}) = \bar{0}$, and $F'(\bar{0})$ is invertible. Then there is a neighbourhood of $\bar{0}$ in $\mathbb{R}^n$ in which the representation $F(\bar{x}) = B_1 B_2 \cdots B_{n-1} G_n \circ G_{n-1} \circ \cdots G_1(\bar{x})$ is valid where G_i are primitive $\mathscr{C}'$-mapping in some neighbourhood of $\bar{0}$, $G_i(\bar{0}) = \bar{0}$, $G'(\bar{0})$ is invertible and each B_i is either a flip or identity operator.*

In other words, locally F is a composition of primitive mappings and flips.

Proof. Proof by mathematical induction on m.

$$\text{There exists a neighbourhood of } \bar{0},\ V_m \text{ such that } F_m \in \mathscr{C}'(V_m) \tag{13.17}$$

$$F_m(\bar{0}) = \bar{0} \tag{13.18}$$

$$F'_m(\bar{0}) \text{ is invertible and} \tag{13.19}$$

$$P_{m-1} F_m(\bar{x}) = P_{m-1}\bar{x},\ \bar{x} \in V_m \tag{13.20}$$

where the kth projection $P_k : \mathbb{R}^n \to \mathbb{R}^n$ is defined by $P_k(\bar{x}) = x_1\bar{e}_1 + x_2\bar{e}_2 + \cdots + x_k\bar{e}_k + 0\bar{e}_{k+1} + \cdots + 0\bar{e}_n$. Clearly, $P_0(\bar{x}) = \bar{0}$.

The essence of this proof lies in the fourth statement which is designed to construct a seqeunce of functions $F_1, F_2, \cdots, F_n$ such that the other three statements remains true for every function in this sequence.

Step 1 : Initial Case, prove that all the four statements are true for $m = 1$.

Define $F_1 = F$ and $V_1 = E$. Thus, $F \in \mathscr{C}'(E) \implies F_1 \in \mathscr{C}'(V_1)$. Also we have, $F_1(\bar{0}) = F(\bar{0}) = 0$, and $F'_1(\bar{0}) = F'(\bar{0})$ is invertible. Obviously the trivial projection, $P_0(F_1(\bar{x})) = P_0(F(\bar{x})) = \bar{0} = P_0(\bar{x})$ for every $\bar{x} \in V_1$.

Step 2 : Induction Hypothesis, suppose all the four statements are true for $m = 1, 2, \cdots, n-1$.

Suppose that for each $m = 1, 2, \cdots, n-1$, there exists a neighbourhood of $\bar{0}$, say V_m such that $F_m \in \mathscr{C}'(V_m)$, $F_m(\bar{0}) = \bar{0}$, $F'_m(\bar{0})$ is invertible and $P_{m-1}F_m(\bar{x}) = P_{m-1}(\bar{x})$ for every $\bar{x} \in V_m$.

Step 3 : Induction Step, prove that all the four statements are true for $m = n$.

$$\begin{aligned} P_{m-1}F_m(\bar{x}) &= P_{m-1}\bar{x},\ \bar{x} \in V_m \\ &= x_1\bar{e}_1 + x_2\bar{e}_2 + \cdots + x_{m-1}\bar{e}_{m-1} \end{aligned}$$

We may write remaining components of $F_m(\bar{x})$ using real-valued functions. That is, kth component of the function F_m, say $F_{m_k} = \alpha_k$.

$$F_m(\bar{x}) = P_{m-1}(\bar{x}) + \sum_{i=m}^{n} \alpha_i(\bar{x})\bar{e}_i$$

Since $F_m \in \mathscr{C}'(V_m)$, the functions $\alpha_i \in \mathscr{C}'(V_m)$. Taking mth partial derivative on either sides, we get

$$D_m F_m(\bar{0}) = F'_m(\bar{0})\bar{e}_m = \sum_{i=m}^{n} D_m \alpha_i(\bar{0})\bar{e}_i$$

We have, $F'_m(\bar{0})$ is invertible. Thus $F'_m(\bar{0})\bar{e}_m \neq 0$. Thus there exists some integer k such that $m \leq k \leq n$ and $D_m\alpha_k(\bar{0})\bar{e}_k \neq 0$. Let B_m be the flip that interchanges m and k. If $m = k$, then B_m is identity map.

$$\text{Define } G_m(\bar{x}) = \bar{x} + [\alpha_k(\bar{x}) - x_m]\bar{e}_m \tag{13.21}$$

Then $G_m \in \mathscr{C}'(V_m)$, G_m is a primitive mapping, and $G'_m(\bar{0})$ is invertible since $D_m\alpha_k(\bar{0}) \neq 0$.

By inverse function theorem, there exists an open set U_m, $0 \in U_m$ and $U_m \subset V_m$ such that G_m is a bijection from U_m onto a neighbourhood of $\bar{0}$, say V_{m+1} in which G_m^{-1} is continuously differentiable.

$$\text{Define } F_{m+1}(\bar{y}) = B_m F_m \circ G_m^{-1}(\bar{y}),\ \bar{y} \in V_{m+1} \tag{13.22}$$

Then $F_m \in \mathscr{C}'(V_{m+1})$, $F_{m+1}(\bar{0}) = 0$ and $F'_{m+1}(\bar{0})$ is invertible by the chain rule. Also, for $\bar{x} \in U_m$, we have

$$\begin{aligned} P_m F_{m+1}(G_m(\bar{x})) &= P_m B_m F_m(\bar{x}) \\ &= P_m(P_{m-1}\bar{x} + \alpha_k(\bar{x}\bar{e}_m + \cdots) \\ &= P_{m-1}\bar{x} + \alpha_k(\bar{x})\bar{e}_m \\ &= P_m G_m(\bar{x}) \\ \implies P_m F_{m+1}(\bar{y}) &= P_m(\bar{y}) \end{aligned}$$

Thus, by mathematical induction, we have characterised a finite sequence of function $F_1, F_2, \cdots, F_n$ such that all the four statements are true.
Step 4 : Using the finite sequence of functions $F_1, F_2, \cdots, F_n$ constructed in the proof, we can represent F in terms of flips and primitive mappings.

Let $\bar{x} \in U_m$ and $\bar{y} = G_m(\bar{x})$. We have,

$$\begin{aligned} B_m F_m \circ G_m^{-1}(\bar{y}) &= F_{m+1}(\bar{y}) \\ B_m B_m F_m(\bar{x}) &= B_m F_{m+1}(G_m(\bar{x})) \\ I F_m(\bar{x}) &= B_m F_{m+1}(G_m(\bar{x})) \\ F_m(\bar{x}) &= B_m(F_{m+1}(G_m(\bar{x})) \end{aligned}$$

$$\text{Thus, } F_m = B_m F_{m+1} \circ G_m,\ m = 1, 2, \cdots, n \tag{13.23}$$

Therefore, we have

$$\begin{aligned} F &= F_1 \\ &= B_1 F_2 \circ G_1 \\ &= B_1(B_2 F_3 \circ G_2) \circ G_1 = B_1 B_2 F_3 \circ G_2 \circ G_1 \\ &\ \vdots \\ F &= B_1 B_2 \cdots B_n F_n \circ G_{n-1} \circ G_{n-2} \circ \cdots \circ G_1 \end{aligned}$$

Since $P_nF_n(\bar{x}) = P_{n-1}(\bar{x})$, F_n is a primitive mapping, say G_n.

$$F = B_1B_2\cdots B_nG_n \circ G_{n-1} \circ G_{n-2} \circ \cdots \circ G_1$$

□

Remark. Let K be a compact subset of $\mathbb{R}^n$. Then a family of function $\psi_1, \psi_2, \cdots, \psi_s$ where $\psi_j : K \to \mathbb{R}$ is a partition of unity if it satisfies

1. $0 \leq \psi_j(\bar{x}) \leq 1$ for every $\bar{x} \in K,\ j = 1, 2, \cdots, s$.
2. $\sum\limits_{j=1}^{s} \psi_j(\bar{x}) = 1$ for every $\bar{x} \in K$.

Theorem 13.4.3 (partitions of unity)**.** *Suppose K is a compact subset of $\mathbb{R}^n$, and $\{V_\alpha\}$ is an open cover of K. Then there exists functions $\psi_1, \psi_2, \cdots, \psi_s \in \mathscr{C}(\mathbb{R}^n)$ such that*

1. *$0 \leq \psi_i \leq 1$ for $1 \leq i \leq s$*
2. *each ψ_i has its support in some V_α and*
3. *$\psi_1(\bar{x}) + \psi_2(\bar{x}) + \cdots + \psi_s(\bar{x}) = 1$ for every $\bar{x} \in K$.*

In other words, every open cover of a compact subset of $\mathbb{R}^n$ has a partition of unity with compact support in a finite subcover.

Proof. Step 1 : Construction of ϕ_j.

We have, $\{V_\alpha\}$ is a cover of K. Therefore, every $\bar{x} \in K$ belongs to some $V_{\alpha(\bar{x})}$. And there exists open balls $B(\bar{x})$ and $W(\bar{x})$ such that

$$\bar{x} \in B(\bar{x}) \subset \bar{B}(\bar{x}) \subset W(\bar{x}) \subset \bar{W}(\bar{x}) \subset V_{\alpha(\bar{x})} \tag{13.24}$$

The family $\{B(\bar{x})\}$ is an open cover of K. Since K is compact, the open cover $\{B(\bar{x})\}$ has a finite subcover, say $\{B(\bar{x}_1), B(\bar{x}_2), \cdots, B(\bar{x}_s)\}$. That is, there are points $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_s$ in K such that

$$K = B(\bar{x}_1) \cup B(\bar{x}_2) \cup \cdots \cup B(\bar{x}_s) \tag{13.25}$$

Since $\mathbb{R}^n$ is a metric space,[6] there exists continuous functions $\phi_j : \mathbb{R}^n \to [0, 1]$ such that $\phi_j(B(\bar{x}_j)) = \{1\}$, $\phi_j(\mathbb{R}^n - W(\bar{x}_j)) = \{0\}$ and $0 \leq \phi_j(\bar{x}) \leq 1$ for every $\bar{x} \in \mathbb{R}^n$.

Step 2 : Construction of ψ_j.

$$\begin{aligned} \text{Define, } \psi_1 &= \phi_1 \\ \psi_2 &= (1 - \phi_1)\phi_2 \\ &\vdots \\ \psi_s &= (1 - \phi_1)(1 - \phi_2)\cdots(1 - \phi_{s-1})\phi_s \end{aligned}$$

Clearly, $0 \leq \psi_j \leq 1$ for $j = 1, 2, \cdots, s$. And ψ_j has a compact support in $W(\bar{x}_j) \subset V_{\alpha(\bar{x}_j)}$.

Step 3 : $\{\psi_j\}$ is a partition of unity.

[6]metric spaces have a simpler version of Urysohn's lemma or apply Urysohn's lemma

We claim that,

$$\psi_1 + \psi_2 + \cdots + \psi_j = 1 - (1-\phi_1)(1-\phi_2)\cdots(1-\phi_j),\ j = 1, 2, \cdots, s \quad (13.26)$$

It is proved using mathematical induction on j. Clearly, it is true for $j = 1$. $\psi_1 = \phi_1 = 1 - 1 + \phi = 1 - (1-\phi)$. Suppose the claim is true for $j = k$ for some $1 \le k < s$. Then we have, $\psi_1 + \psi_2 + \cdots + \psi_k = 1 - (1-\phi_1)(1-\phi_2)\cdots(1-\phi_k)$.

$$\begin{aligned}\text{Thus, } \psi_1 + \psi_2 + \cdots + \psi_k + \psi_{k+1} =& 1 - (1-\phi_1)(1-\phi_2)\cdots(1-\phi_k)\\ &+ (1-\phi_1)(1-\phi_2)\cdots(1-\phi_k)\phi_{k+1}\\ =& 1 - (1-\phi_1)(1-\phi_2)\cdots(1-\phi_{k+1})\end{aligned}$$

Thus, the claim is true for $j = 1, 2, \cdots, s$. Let $\bar{x} \in K$. Then $\bar{x} \in B(\bar{x}_j)$ for some j. By definition of ϕ_j, $\phi_j(\bar{x}) = 1$. That is, $(1 - \phi_j(\bar{x})) = 0$.

$$\begin{aligned}\implies (1-\phi_1(\bar{x}))(1-\phi_2(\bar{x}))\cdots(1-\phi_s(\bar{x})) = 0\\ \implies \psi_1(\bar{x}) + \psi_2(\bar{x}) + \cdots + \psi_s(\bar{x}) = 1\end{aligned}$$

Therefore, $\psi_1(\bar{x}) + \psi_2(\bar{x}) + \cdots + \psi_s(\bar{x}) = 1$ for every $\bar{x} \in K$. □

Definitions 13.4.5. By theorem, every open cover $\{V_\alpha\}$ of a compact subset K has a partition of unity $\{\psi_j\}$ with compact support in a finite subcover $\{V_{\alpha_j}\}$. Then $\{\psi_j\}$ is **subordinate** to the cover $\{V_\alpha\}$.

Corollary 13.4.3.1. *If $f \in \mathscr{C}(\mathbb{R}^n)$ and the support of f lies in K, then $f = \sum_{i=1}^{s} \psi_i f$. Each ψ_i has its support in some V_α.*

Any continuous function with support in a compact set can be represented as sum of continuous functions $\psi_j f$ with small supports.

Proof. Let K be compact subset of $\mathbb{R}^n$. And support of a function f lies in K. Then,

$$\begin{aligned}f(\bar{x}) = I(\bar{x})f(\bar{x}) &= \left(\sum_{j=1}^{s} \psi_j(\bar{x})\right) f(\bar{x})\\ &= \sum_{j=1}^{s} \psi_j(\bar{x})f(\bar{x}) = \sum_{j=1}^{s} (\psi_j f)(\bar{x})\end{aligned}$$

Let $\{\psi_j\}$ be a partition of unity. Let K' be the support of f and V_{α_j} be support of each ψ_j. Then $K' \cap V_{\alpha_j}$ is the support of each $\psi_j f$. □

Theorem 13.4.4 (effect of change of variable on multiple integral)**.** *Suppose T is a one-to-one $\mathscr{C}'$-mapping of an open set $E \subset \mathbb{R}^k$ into $\mathbb{R}^k$ such that $J_T(\bar{x}) \neq 0$ for all $\bar{x} \in E$. If f is a continuous function on $\mathbb{R}^k$ whose support is compact and lies in $T(E)$, then*

$$\int_{\mathbb{R}^k} f(\bar{y})\ d\bar{y} = \int_{\mathbb{R}^k} f(T(\bar{x}))|J_T(\bar{x})|\ d\bar{x} \quad (13.27)$$

Proof. Step 1 : 'separable' functions

Let E be an open subset of $\mathbb{R}^k$. We claim that[7] the theorem is true for functions $h : E \to \mathbb{R}$ of the form $h(\bar{y}) = h_1(y_1)h_2(y_2)\cdots h_k(y_k)$. We know that, every continuous function T is locally a composition of primitives and flips. Therefore it is enough to prove that the theorem is true for functions h with transformations T primitives, flips and their compositions.

Step 2 : Trasformation T is a primitive

Let G be a primitive with change in mth coordinate. Then

$$G(\bar{x}) = x_1\bar{e}_1 + x_2\bar{e}_2 + \cdots + g(\bar{x})\bar{e}_m + \cdots + x_k\bar{e}_k \tag{13.28}$$

We have,

$$J_G(\bar{x}) = \begin{vmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \iddots & \vdots & 0 \\ D_1g(\bar{x}) & D_2g(\bar{x}) & \cdots & D_mg(\bar{x}) & \cdots & D_{k-1}g(\bar{x}) & D_kg(\bar{x}) \\ \vdots & \vdots & \iddots & 0 & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 & 1 \end{vmatrix} \tag{13.29}$$

Therefore, $J_G(\bar{x}) = D_mg(\bar{x})$.

$$\begin{aligned} \int_{\mathbb{R}^k} h(\bar{y})\ d\bar{y} &= \prod_{j=1}^{k} \int h_j(y_j)\ dy_j \\ &= \left(\prod_{j\neq m} \int h_j(y_j)\ dy_j \right) \int h_m(y_m)\ dy_m \\ G(\bar{x}) = \bar{y}\ &\Longrightarrow (x_1, x_2, \cdots, g(\bar{x}), \cdots, x_k) = (y_1, y_2, \cdots, y_m, \cdots, y_k) \\ &\Longrightarrow x_j = y_j \text{ for } j \neq m \text{ and } g(\bar{x}) = y_m \\ \int_{\mathbb{R}^k} h(\bar{y})\ d\bar{y} &= \left(\prod_{j\neq m} \int h_j(x_j)\ 1\ dx_j \right) \int h_m(g(\bar{x}))\ D_mg(\bar{x})\ dx_m \\ \int_{\mathbb{R}^k} h(\bar{y})\ d\bar{y} &= \int_{\mathbb{R}^k} h(G(\bar{x}))\ |J_G(\bar{x})|\ d\bar{x} \end{aligned}$$

Step 3 : Transformation T is a flip.

Let B be a flip that interchanges m and n coordinates. When $m = n$, B is an identity map and the theorem is trivially true.

$$B(\bar{x}) = x_1\bar{e}_1 + \cdots + x_n\bar{e}_m + \cdots + x_m\bar{e}_n + \cdots + x_k\bar{e}_k \tag{13.30}$$

The Jacobian matrix is an identity matrix with mth and nth rows interchanged.

[7] This proof is my own work. There may exist simpler proofs.

For example,

$$J_B(\bar{x}) = \begin{vmatrix} 0 & 1 & \cdots & 0 \\ 1 & 0 & \vdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{vmatrix} \tag{13.31}$$

Therefore, $J_B(\bar{x}) = \pm 1$.

$$\begin{aligned}
\int_{\mathbb{R}^k} h(\bar{y})\, d\bar{y} &= \prod_{j=1}^{k} \int h_j(y_j)\, dy_j \\
&= \left(\prod_{j \neq m,n} \int h_j(y_j)\, dy_j \right) \int h_n(y_n)\, dy_n \int h_m(y_m)\, dy_m \\
B(\bar{x}) = \bar{y} &\implies (x_1, \cdots, x_n, \cdots, x_m, \cdots, x_k) = (y_1, \cdots, y_m, \cdots, y_n, \cdots, y_k) \\
\int_{\mathbb{R}^k} h(\bar{y})\, d\bar{y} &= \left(\prod_{j \neq m,n} \int h_j(x_j)\, dx_j \right) \int h_n(x_m)\, dx_n \int h_m(x_n)\, dx_m \\
&= \int_{\mathbb{R}^k} h(B(\bar{x}))\ 1\ d\bar{x} = \int_{\mathbb{R}^k} h(B(\bar{x}))\ |J_B(\bar{x}|\ d\bar{x}
\end{aligned}$$

Notice that, substituting $y_n = x_m$ gives $\int h_n(x_m)\, dx_n$. This is due to the fact that $\bar{y} = B(\bar{x})$ have $x_m \bar{e}_n$ in place of $y_n \bar{e}_n$. Therefore, parameter of h_n is on the same axis $\bar{e}_n$.

Step 4 : If the theorem is true for two transformations, then it is true for their composition.

Suppose the theorem is true for transformations P and Q. And $S = P \circ Q$. Let $\bar{z} = P(\bar{y})$ and $\bar{y} = Q(\bar{x})$. Then $\bar{z} = S(\bar{x})$. Since $m(P)m(Q) = m(S)$, we have $J_P(\bar{y})J_Q(\bar{x}) = J_S(\bar{x})$.

$$\begin{aligned}
\int_{\mathbb{R}^k} f(\bar{z})\, d\bar{z} &= \int_{\mathbb{R}^k} f(P(\bar{y}))\ |J_P(\bar{y})|\ d\bar{y} \\
&= \int_{\mathbb{R}^k} f(P(Q(\bar{x})))\ |J_P(Q(\bar{x}))|\ |J_Q(\bar{x})|\ d\bar{x} \\
&= \int_{\mathbb{R}^k} f(S(\bar{x}))\ |J_S(\bar{x})|\ d\bar{x}
\end{aligned}$$

Therefore, the theorem is true for their composition.

Now for any function of the form $h(\bar{x}) = \prod_{j=1}^{k} h_j(x_j)$, the theorem is true for any continuous transformation T. Then it is true for the algebra of functions $\mathscr{A}$. And by Stone-Weierstrass theorem,[8] the theorem is true for every continuous function f. □

13.4.1 Differential Forms

Definitions 13.4.6. Suppose E is an open set in $\mathbb{R}^n$. A k**-surface** in E is a $\mathscr{C}'$-mapping Φ from a compact set $D \subset \mathbb{R}^k$ into E. D is the **parameter**

[8] I haven't checked whether the application of Stone-Weierstrass theorem causes any problem. It is the same proof technique as we have seen in $L(f) = L'(f)$.

domain of Φ.

Definitions 13.4.7. Suppose E is an open set in $\mathbb{R}^n$. A **differential form** of order $k \geq 1$ in E, a k-form in E is a function ω, symbolically represented by the sum

$$\omega = \sum a_{i1\cdots ik}(\bar{x})\ dx_{i1} \wedge dx_{i2} \wedge \cdots dx_{ik} \tag{13.32}$$

which assigns to each k-surface Φ in E a number $\omega(\Phi) = \int_\Phi \omega$, according to the rule

$$\int_\Phi \omega = \int_D \sum a_{i1\cdots ik}(\Phi(\bar{u}))\frac{\partial(x_{i1}, x_{i2}, \cdots, x_{ik})}{\partial(u_1, u_2, \cdots, u_k)} d\bar{u} \tag{13.33}$$

where D is the paramter domain of Φ and

$$\frac{\partial(x_{i1}, x_{i2}, \cdots, x_{ik})}{\partial(u_1, u_2, \cdots, u_k)} = \begin{vmatrix} \frac{\partial x_{i1}}{\partial u_1} & \frac{\partial x_{i1}}{\partial u_2} & \cdots & \frac{\partial x_{i1}}{\partial u_k} \\ \frac{\partial x_{i2}}{\partial u_1} & \frac{\partial x_{i2}}{\partial u_2} & \cdots & \frac{\partial x_{i2}}{\partial u_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_{ik}}{\partial u_1} & \frac{\partial x_{ik}}{\partial u_2} & \cdots & \frac{\partial x_{ik}}{\partial u_k} \end{vmatrix} \tag{13.34}$$

Let $\omega = 4\ dx_1 \wedge dx_3 + 3x_2^2\ dx_2 \wedge dx_1$. Then ω is a 2-form with $a_{1,3}(x,y,z) = 4$, $a_{2,1}(x,y,z) = 3y^2$ and all other $a_{i_1 i_2}(x,y,z) = 0$.

Consider the upper hemi-sphere of unit radius, $S^2_{y\geq 0}$ in $\mathbb{R}^3$ and the closure of the unit disc $\bar{S}^1 = D \subset \mathbb{R}^2$. Then D is compact. Let $\Phi : D \to S^2_{y\geq 0}$ defined by $\Phi(x,y) = (x, +\sqrt{1-x^2-y^2}, y)$. Clearly, $\Phi \in \mathscr{C}'(D)$. Thus, Φ is a 2-surface with parameter domain D.

Clearly, $\Phi_1(x,y) = x$, $\Phi_2(x,y) = \sqrt{1-x^2-y^2}$ and $\Phi_3(x,y) = y$. We have, $\omega(\Phi) = \int_\Phi \omega$.

$$\omega(\Phi) = \int_\Phi \omega = \iint_D \omega_\Phi(x,y)\frac{\partial(\Phi_{i_1}, \Phi_{i_2})}{\partial(x,y)} dx\ dy \tag{13.35}$$

where $\omega_\Phi(x,y)$ is obtained by evaluating each function $a_{i_1 i_2 \cdots i_k}(\Phi(x,y))$.

$$\begin{aligned}
\omega(\Phi) &= \iint_D a_{1,3}(\Phi(x,y)) \begin{vmatrix} \frac{\partial \Phi_1}{\partial x} & \frac{\partial \Phi_1}{\partial y} \\ \frac{\partial \Phi_3}{\partial x} & \frac{\partial \Phi_3}{\partial y} \end{vmatrix} dxdy + \iint_D a_{2,1}(\Phi(x,y)) \begin{vmatrix} \frac{\partial \Phi_2}{\partial x} & \frac{\partial \Phi_2}{\partial y} \\ \frac{\partial \Phi_1}{\partial x} & \frac{\partial \Phi_1}{\partial y} \end{vmatrix} dxdy \\
&= \int_0^1 \int_0^{\sqrt{1-y^2}} 4 \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} dxdy \\
&+ \int_0^1 \int_0^{\sqrt{1-y^2}} 3(1-x^2-y^2) \begin{vmatrix} \frac{1}{2}\frac{-2x}{\sqrt{1-x^2-y^2}} & \frac{1}{2}\frac{-2y}{\sqrt{1-x^2-y^2}} \\ 1 & 0 \end{vmatrix} dxdy \\
&= \int_0^1 \int_0^{\sqrt{1-y^2}} 4dxdy + \int_0^1 \int_0^{\sqrt{1-y^2}} \frac{3}{\sqrt{1-x^2-y^2}} dx\ dy
\end{aligned}$$

Remark (example 1). Integrals of 1-forms are line integrals. And $\omega(\gamma) = 0$ for every closed curve γ.[Rudin, 1976, 10.12a]

Remark (example 2). Let $\gamma : [0, 2\pi] \to \mathbb{R}^2$ defined by $\gamma(t) = (a \cos t, b \sin t)$. Then γ is a 1-surface with parameter domain $[0, 2\pi]$.[Rudin, 1976, 10.12b] Let ω be a 1-form defined by $\omega = xdy$. Then

$$\omega(\gamma) = \int_\gamma \omega = \int_\gamma x \ dy = \int_0^{2\pi} ab \cos^2 t \ dt = \pi ab$$

Similarly, $\omega = ydx$ gives

$$\omega(\gamma) = \int_\gamma \omega = \int_\gamma y \ dx = \int_0^{2\pi} -ab \sin^2 t \ dt = -\pi ab$$

Remark (example 3). Let $0 \le r \le 1$, $0 \le \theta \le \pi$ and $0 \le \phi \le 2\pi$. Then $D \subset \mathbb{R}^n$ defined by $\{(r, \theta, \phi)\}$ is compact. Define $\Phi : D \to \mathbb{R}^3$ by $\Phi(r, \theta, \phi) = (r \sin\theta \cos\phi, r \sin\theta \sin\phi, r \cos\theta)$. Then Φ is a 3-surface in $\mathbb{R}^3$. We have

$$J_\Phi(r, \theta, \phi) = \begin{vmatrix} \sin\theta\cos\phi & r\cos\theta\cos\phi & -r\sin\theta\sin\phi \\ \sin\theta\sin\phi & r\cos\theta\sin\phi & r\sin\theta\cos\phi \\ \cos\theta & -r\sin\theta & 0 \end{vmatrix} = r^2 \sin\theta$$

Let $\omega = dx_1 \wedge dx_2 \wedge dx_3$. Then $\omega(\Phi) = \int_\Phi \omega = \int_D J_\Phi = \frac{4\pi}{3}$ is the volume of the unit ball $\Phi(D)$.

Remark. A k-form ω is of class $\mathscr{C}'$ or $\mathscr{C}''$ if the functions $a_{i1 \cdots ik}$ are all of class $\mathscr{C}'$ or $\mathscr{C}''$. A 0-form in E is defined to be a continuous function in E. And 0 is the only k-form in any open set $E \subset \mathbb{R}^n$.

Definitions 13.4.8. Let $\omega = a(\bar{x}) \ dx_{i1} \wedge dx_{i2} \wedge \cdots \wedge dx_{ik}$. Then $\bar{\omega}$ is the k-form obtained by interchanging a pair subscripts of ω. ie, $\bar{\omega} = a(\bar{x}) \ dx_{i2} \wedge dx_{i1} \wedge \cdots \wedge dx_{ik}$.

Elementary properties of k-forms

Definitions 13.4.9. Let E be an open set in $\mathbb{R}^n$. And Φ be a k-surface in E. And let ω_1, ω_2 be k-forms in E, then

1. $\omega_1 = \omega_2 \iff \omega_1(\Phi) = \omega_2(\Phi) \iff \int_\Phi \omega_1 = \int_\Phi \omega_2, \ \forall \Phi \in E$
2. $\omega = 0 \iff \omega(\Phi) = 0 \iff \int_\Phi \omega = 0, \ \forall \Phi \in E$
3. k-form Addition, $\omega_1 + \omega_2$
$\omega = \omega_1 + \omega_2 \iff \omega(\Phi) = \omega_1(\Phi) + \omega_2(\Phi) \iff \int_\Phi \omega = \int_\Phi \omega_1 + \int_\Phi \omega_2$.
4. Scalar multiplication, $c\omega$
$c\omega(\Phi) = c(\omega(\Phi)) \iff \int_\Phi c\omega = c \int_\Phi \omega$.
5. Inverse k-form, $-\omega$
$-\omega(\Phi) = -(\omega(\Phi)) \iff \int_\Phi -\omega = -\int_\Phi \omega$
6. $\bar{\omega} = -\omega$.

Remark. Let $\omega = a(\bar{x}) \ dx_{i1} \wedge dx_{i2} \wedge \cdots \wedge dx_{ik}$. If $\bar{\omega} = \omega$, then $\omega = 0$. Since $\wedge$ is anticommutative.

Thus, differential k-forms with repeated subscripts are 0. For example, $\omega = dx_i \wedge dx_j \wedge dx_i = 0$.

Definitions 13.4.10. Let $\bar{I} = (i_1, i_2, \cdots, i_k)$ be an **increasing k-index**. That is, $1 \leq i_1 \leq i_2 \leq \cdots i_k \leq n$. Then $dx_{\bar{I}}$ of the form $dx_{\bar{I}} = dx_{i1} \wedge dx_{i2} \cdots dx_{ik}$ is **basic k-form** in $\mathbb{R}^n$.

Remark. List of all basic k-forms

0-form 0

1-forms dx_1, dx_2, $\cdots$, dx_n

2-forms $dx_i \wedge dx_j$ for every $1 \leq i, j \leq n$.

3-forms $dx_i \wedge dx_j \wedge dx_k$ for every $1 \leq i, j, k \leq n$.

Remark. There are $\binom{n}{k}$ basic k-forms in $\mathbb{R}^n$.

Every k-form can be represented in terms of basic k-forms. For every k-tuple $(j_1, j_2, \cdots, j_k)$, $dx_{j1} \wedge \cdots dx_{jk} = \sigma(j_1, j_2, \cdots, j_k) dx_{\bar{J}}$ where $\bar{J}$ is the increasing k-index obtained by interchanging pairs. And σ maps odd permutations to -1 and even permutations to 1.

Standard representation of a k-form

$$\omega = \sum_I b_I(\bar{x}) dx_I \tag{13.36}$$

For example : $x_1\ dx_2 \wedge dx_1 - x_2\ dx_3 \wedge dx_2 + x_3\ dx_2 \wedge dx_3 + dx_1 \wedge dx_2 = (1 - x)\ dx_1 \wedge dx_2 + (x_2 + x_3)\ dx_2 \wedge dx_3$ is a 2-form in $\mathbb{R}^3$.

$dx_1 \wedge dx_2 \wedge dx_3 = dx_2 \wedge dx_3 \wedge dx_1$, since $(1\ 2\ 3) \in A_3 \implies \sigma(1\ 2\ 3) = 1$. And, $dx_1 \wedge dx_2 \wedge dx_3 = -dx_1 \wedge dx_3 \wedge dx_2$, $(2\ 3) \notin A_3 \implies \sigma(2\ 3) = -1$. Here, A_3 is an alternating group of all even permutation on $\{1, 2, 3\}$.

Subject 14

ME010304 Functional Analysis

Module 1 - Banach Space

14.1 Metric Spaces

The following are a few metric spaces and their distance functions,

1. $\mathbb{R}, d(x, y) = |x - y|$
2. $\mathbb{R}^2, d(x, y) = \left(|\xi_1 - \eta_1|^2 + |\xi_2 - \eta_2|^2\right)^{\frac{1}{2}}$ where $x = (\xi_1, \xi_2)$
3. $\mathbb{R}^3, d(x, y) = \left(|\xi_1 - \eta_1|^2 + |\xi_2 - \eta_2|^2 + |\xi_3 - \eta_3|^2\right)^{\frac{1}{2}}$ where $x = (\xi_1, \xi_2, \xi_3)$
4. Euclidean space, $\mathbb{R}^n, d(x, y) = \left(\sum_{j=1}^{n} |\xi_j - \eta_j|^2\right)^{\frac{1}{2}}$ where $x = (\xi_1, \xi_2, \ldots, \xi_n)$
5. $\mathbb{C}, d(x, y) = |x - y|$ where $x = a + ib$ and $|x| = (a^2 + b^2)^{\frac{1}{2}}$
6. Unitary space, $\mathbb{C}^n, d(x, y) = \left(\sum_{j=1}^{n} |\xi_j - \eta_j|^2\right)^{\frac{1}{2}}$ where $x = (\xi_1, \xi_2)$ and $\xi_j \in \mathbb{C}$.
7. $C[a, b], d(x, y) = \max_{t \in [a,b]} \{|x(t) - y(t)|\}$ where x is a continuous, complex-valued function defined on closed interval $[a, b] \subset \mathbb{R}$.
8. $B(A), d(x, y) = \sup_{t \in A} \{|x(t) - y(t)|\}$ where x is a bounded, complex-valued function defined on $A \subset \mathbb{R}$
9. $l^p, d(x, y) = \left(\sum_{j=1}^{\infty} |\xi_j - \eta_j|^p\right)^{\frac{1}{p}}$ where $x = \{\xi_j\}$ and $\sum_{j=1}^{\infty} |\xi_j|^p < \infty$. That is, set of all sequences such that the pth power series is convergent.

10. $l^\infty, d(x,y) = \sup_{j\in\mathbb{N}}\{|\xi_j - \eta_j|\}$ where $x = \{\xi_j\}$ and $|\xi_j| \le c$. That is, l^∞ is the space of all bounded sequences of complex numbers.

11. $c, d(x,y) = \sup_{j\in\mathbb{N}}\{|\xi_j - \eta_j|\}$ where $\{\xi_j\} = \xi_1, \xi_2, \dots$ is the space of all convergent sequences of complex numbers.

12. $s, d(x,y) = \sum_{j=1}^{\infty} \frac{1}{2^j}\frac{|\xi_j - \eta_j|}{1+|\xi_j - \eta_j|}$ where $x = \{\xi_j\} = \xi_1, \xi_2, \dots$ is the space of all sequence of complex numbers.

Clearly, $l^p \subset l^\infty \subset s$. But, $B(A), C[a,b]$ are non-comparable since there exists discontinuous, bounded functions and continuous, unbounded functions.

Definitions 14.1.1 (conjugate exponents)**.** Real numbers p, q are conjugates, if $p > 1, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$.

Note : if p, q are conjugates, then $p + q = pq$ and $(p-1)(q-1) = 1$. Also if $u = t^{p-1}$, then $t = u^{q-1}$.

Theorem 14.1.1 (Hölder[1] Inequality for sums)**.** *Let p, q be conjugates and $(\xi_j), (\eta_j)$ be two complex sequences. $(\xi_j) \in l^p$ and $(\eta_j) \in l^q$. Then*

$$\sum_{j=1}^{\infty} |\xi_j \eta_j| \le \left(\sum_{j=1}^{\infty} |\xi_j|^p\right)^{\frac{1}{p}} \left(\sum_{j=1}^{\infty} |\eta_j|^q\right)^{\frac{1}{q}} \tag{14.1}$$

Corollary 14.1.1.1 (Cauchy-Schwarz Inequality)**.** *When $p = 2$, we have its conjugate $q = 2$. Then Hölder inequality reduces to the following*

$$\sum_{j=1}^{\infty} |\xi_j \eta_j| \le \left(\sum_{j=1}^{\infty} |\xi_j|^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^{\infty} |\eta_j|^2\right)^{\frac{1}{2}} \tag{14.2}$$

Theorem 14.1.2 (Minkowski[2] Inequality)**.** *Let p, q be conjugates and $(\xi_j), (\eta_j)$ be two complex sequences in l^p. Then*

$$\left(\sum_{j=1}^{\infty} |\xi_j + \eta_j|^p\right)^{\frac{1}{p}} \le \left(\sum_{j=1}^{\infty} |\xi_j|^p\right)^{\frac{1}{p}} + \left(\sum_{j=1}^{\infty} |\eta_j|^q\right)^{\frac{1}{q}} \tag{14.3}$$

14.1.1 A few more Concepts

product of metric spaces Let $(X_1, d_1), (X_2, d_2)$ be two metric spaces. Let $x, y \in X_1 \times X_2$ where $x = (\xi_1, \xi_2)$ and $y = (\eta_1, \eta_2)$. Then, $X_1 \times X_2$ is a metric space with following metrics,

1. $d(x,y) = d_1(\xi_1, \eta_1) + d_2(\xi_2, \eta_2)$
2. $d(x,y) = \left(d_1(\xi_1, \eta_1)^2 + d_2(\xi_2, \eta_2)^2\right)^{\frac{1}{2}}$

[1] O.Hölder
[2] H. Minkowski

3. $d(x, y) = \max\{d_1(\xi_1, \eta_1), d_2(\xi_2, \eta_2)\}$

separable A space is separable if it has a countable, dense subset.
Examples : l^∞ is not separable. But, l^p spaces are separable ($1 \leq p < \infty$). $\mathbb{R}, \mathbb{C}$ are separable. $B[a, b]$ not seprable.

14.1.2 Exercises

1. Find a sequence $\{\xi_j\}$ which converges to 0 but does not belong to any l^p space. ($1 \leq p < \infty$).
2. Find a sequence $\{\xi_j\}$ which belong to l^p space with $p > 1$, but not in l^1.

14.1.5 Completeness

Convergent sequences in a metric space are Cauchy sequences. But, Cauchy sequences are not necessarily convergent.

Definitions 14.1.2 (complete)**.** A metric space is complete if every Cauchy sequence in it is convergent.

Theorem 14.1.3. $\mathbb{R}$ *is complete.*

Proof. Let $\{\xi_n\}$ be a Cauchy sequence in $\mathbb{R}$. Let $M_0 > 0$. Then, there exists $N \in \mathbb{N}$ such that $\forall n, m > N,\ d(x_n, x_m) < M_0$. Let $M = \max\{M_0, M_1, \ldots, M_N\}$ where $\forall n \leq N,\ |\xi_j| < M_j$. Then $\{\xi_n\}$ is bounded by M.

By Bolzano Weierstrass†[3] theorem, every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence. Let $x \in \mathbb{R}$ be the limit of a convergent subsequence $\{\xi_{n_k}\}$ of $\{\xi_n\}$. Then $\xi_n \to x$ since $\{\xi_n\}$ is a Cauchy sequence. ☐

Theorem 14.1.4. $\mathbb{C}$ *is complete.*

Proof. Let $\{\xi_n\}$ be a Cauchy sequence in $\mathbb{C}$. Then the real, imaginary parts of ξ_n are a Cauchy sequences in $\mathbb{R}$. We know that $\mathbb{R}$ is complete. Let $\Re(\xi_n) \to a$ and $\Im(\xi_n) \to b$. Then $\xi_n \to a + ib$ since $\{\xi_n\}$ is a Cauchy sequence. ☐

Theorem 14.1.5. *Finite dimensional Euclidean space, $\mathbb{R}^n$ is complete.*

Proof. Let $\{x_k\}$ be Cauchy sequence in $\mathbb{R}^n$. Let $\varepsilon > 0$. There exists $N \in \mathbb{N}$ such that $\forall m, r > N,\ d(x_m, x_r) < \varepsilon$. Let $x_m = (\xi_{1,m}, \xi_{2,m}, \ldots, \xi_{n,m})$ and $x_r = (\xi_{1,r}, \xi_{2,r}, \ldots, \xi_{n,r})$.

$$
\begin{aligned}
d(x_m, x_r) = \left(\sum_{j=1}^{n} |\xi_{j,m} - \xi_{j,r}|^2\right)^{\frac{1}{2}} &< \varepsilon \\
\implies \sum_{j=1}^{n} |\xi_{j,m} - \xi_{j,r}|^2 &< \varepsilon^2 \\
\implies |\xi_{j,m} - \xi_{j,r}| &< \varepsilon^2,\ j = 1, 2, \ldots, n
\end{aligned}
$$

[3]Every sequence has a monotone subsequence. And by monotone converges theorem, every monotone bounded sequence has a convergent subsequence.

$$\begin{array}{cccccc}
x_1 = & \xi_{1,1} & \xi_{1,2} & \xi_{1,3} & \cdots \\
x_2 = & \xi_{2,1} & \xi_{2,2} & \xi_{2,3} & \cdots \\
x_3 = & \xi_{3,1} & \xi_{3,2} & \xi_{3,3} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
x = & \xi_1 & \xi_2 & \xi_3 & \cdots
\end{array}$$

Figure 14.1: Construction of limit in Sequence spaces

Therefore, $\{\xi_{j,k}\}$'s are Cauchy sequences in $\mathbb{R}$ for $j = 1, 2, \dots, n$. Since $\mathbb{R}$ is convergent, $\xi_{j,k} \to \xi_j$ for each j. Let $x = (\xi_1, \xi_2, \dots, \xi_n)$. Let $\varepsilon > 0$. Then there exists $N_j \in \mathbb{N}$ such that $\forall m > N$, $|\xi_{j,m} - \xi_j| < \frac{\varepsilon}{n}$. Let $N = \max\{N_1, N_2, \dots, N_n\}$. Then $\forall m > N$,

$$d(x_m, x) = \left(\sum_{j=1}^{n} |\xi_{j,m} - \xi_j|^2 \right)^{\frac{1}{2}} = \left(\sum_{j=1}^{n} \frac{\varepsilon^2}{n^2} \right)^{\frac{1}{2}} = \frac{\varepsilon}{\sqrt{n}} < \varepsilon$$

Therefore, $\{x_n\}$ converges to x in $\mathbb{R}^n$. ☐

Theorem 14.1.6. *Finite dimensional Unitary space, $\mathbb{C}^n$ is complete.*

Proof. Same proof as above. ☐

Theorem 14.1.7. *The complex sequence space of all bounded sequences, l^∞ is complete.*

Proof. Let $\{x_k\}$ be a Cauchy sequence in l^∞ where $x_k = \xi_{k,1},\ \xi_{k,2},\ \dots$ are bounded sequences in $\mathbb{C}$ for each k.
Step 1 : Construct x
Let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that $\forall m, n > N$,

$$d(x_m, x_n) = \sup_{k \in \mathbb{N}} \{|\xi_{m,k} - \xi_{n,k}|\} < \varepsilon$$

From the metric, we have $\{\xi_{j,k}\}$'s are Cauchy sequences in $\mathbb{C}$ for each k. Thus, $\xi_{j,k} \to \xi_k \in \mathbb{C}$. Let $x = \xi_1,\ \xi_2,\ \dots$. Now we need to prove that $x_k \to x$ and $x \in l^\infty$.
Step 2 : $x \in l^\infty$
We have, $x_j \in l^\infty \implies |\xi_{m,j}| < c_j,\ \forall m \in \mathbb{N}$. Therefore,

$$|\xi_j| \le |\xi_j - \xi_{m,j}| + |\xi_{m,j}| \le \varepsilon + c_j, \quad \forall j \in \mathbb{N}$$

Thus, $x \in l^\infty$.
Step 3 : $x_n \to x$
Since $|\xi_{m,j} - \xi_j| < \varepsilon$ for $m \in \mathbb{N}$,

$$d(x_m, x) = \sup_{j \in \mathbb{N}} |\xi_{m,j} - \xi_j| \le \varepsilon$$

☐

Theorem 14.1.8. *The complex sequence space of all convergent sequences, c is complete*

Proof. Since every convergent sequence is bounded, $c \subset l^\infty$. And the sequence space c is complete if c is a closed subset of the complete space l^∞. Therefore, it is sufficient to show that $c = \overline{c}$

Let $x = \{\xi_k\} \in \overline{c}$. Then there exists a convergent sequence $\{x_k\} = \{\xi_{k,j}\}$ in c converging to x. That is, $\xi_{k,j} \to \xi_j$.

Let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that $\forall n \geq N$ and $\forall j, k \in \mathbb{N}$, we have $d(x_N, x) = \sup_{r \in \mathbb{N}}\{|\xi_{n,r} - \xi_r|\} < \frac{\varepsilon}{3}$. Therefore,

$$|\xi_{N,j} - \xi_j| < \frac{\varepsilon}{3} \text{ and } |\xi_{N,k} - \xi_k| < \frac{\varepsilon}{3} \tag{14.4}$$

Every convergent sequence in complete metric space l^∞ is also a Cauchy sequence. Since $x_N = \{\xi_{N,k}\}$ is a Cauchy sequence, there exists $N_1 \in \mathbb{N}$ and $\forall j, k \geq N_1$,

$$|\xi_{N,j} - \xi_{N,k}| < \frac{\varepsilon}{3} \tag{14.5}$$

Therefore,

$$|\xi_j - \xi_k| \leq |\xi_j - \xi_{N,j}| + |\xi_{N,j} - \xi_{N,k}| + |\xi_{N,k} - \xi_k| \leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$$

Clearly, $x = \{\xi_k\}$ is a Cauchy sequence in l^∞. And since l^∞ is complete, every Cauchy sequence in l^∞ is convergent. Therefore, x is a convergent sequence and $x \in c$. Thus, $c = \overline{c}$ and the induced†[4] metric space c is complete. □

Theorem 14.1.9. *For any $p \geq 1$, the sequence space l^p is complete.*

Proof. Let $\{x_k\}$ be a Cauchy sequence l^p where $x_k = \{\xi_{k,j}\}$ and

$$\sum_{j=1}^{\infty} |\xi_{k,j}|^p < \infty, \ \forall k \in \mathbb{N}$$

Since $\{x_k\}$ is a Cauchy sequence, for $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$d(x_m, x_n) = \left(\sum_{j=1}^{\infty} |\xi_{m,j} - \xi_{n,j}|^p \right)^{\frac{1}{p}} < \varepsilon, \quad \forall m, n > N \text{ since } x_k \in l^p \tag{14.6}$$

Thus, $\forall m, n > N, \ |\xi_{m,j} - \xi_{n,j}| < \varepsilon$ for each j.

Step 1 : Construction of x

Define $x = \{\xi_j\}$ where ξ_j is an accumulation point of $\{\xi_{k,j}\}$.

Step 2 : $x \in l^p$

Let $j > N$. Then, $\{\xi_{k,j}\}$ is a Cauchy sequence of complex numbers. Then,

[4] Let (X, d) be a complete metric space. And Y is a closed subset of X. Then the induced metric space $(Y, d_{|_Y})$ is complete.

$\xi_{k,j} \to \xi_j$. Applying limit $n \to \infty$ on equation 14.6, we get

$$\left(\sum_{j=1}^{\infty} |\xi_{m,j} - \xi_j|^p\right)^{\frac{1}{p}} < \varepsilon, \quad \forall m > N$$
$$\implies \sum_{j=1}^{\infty} |\xi_{m,j} - \xi_j|^p < \varepsilon^p$$

Let $m > N$. Then $x_m - x \in l^p$. By Minkowski inequality, we have

$$\sum_{j=1}^{\infty} |\xi_j|^p \leq \sum_{j=1}^{\infty} |\xi_{m,j} - \xi_j|^p + \sum_{j=1}^{\infty} |\xi_{m,j}|^p < \infty \tag{14.7}$$

Thus, $x \in l^p$.
Step 3: $x_m \to x$
Since $\sum_{j=1}^{\infty} |\xi_{m,j} - \xi_j|^p < \varepsilon^p$, we have $\xi_{k,j} \to \xi_j$. Therefore, $x_m \to x$. □

Theorem 14.1.10. *The function space of all continuous functions defined on closed interval $[a, b]$, $C[a, b]$ is complete.*

Proof. Let $\{x_k\}$ be a Cauchy sequence in $C[a, b]$.
Let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that $m, n > N$,

$$d(x_m, x_n) = \max_{t\in[a,b]} \{x_m(t) - x_n(t)\} < \varepsilon$$

Thus, for each $t \in [a, b]$, sequence $\{x_k(t)\}$'s are Cauchy sequences. We know that $\mathbb{C}$ is complete. Thus every Cauchy sequence in $\mathbb{C}$ is convergent.

Define $x : [a, b] \to \mathbb{C}$ defined by $x(t) =$ the limit of the sequence $\{x_k(t)\}$. As $n \to \infty$, $d(x_m, x_n) \to d(x_m, x)$ and for any $m > N$, $d(x_m, x) < \varepsilon$ uniformly. It is evident from the construction that, the convergence $x_k(t) \to x(t)$ is uniform.

Since the function x_k's are continuous and the convergence is uniform, the limit function x is also continuous. Therefore $x \in C[a, b]$. □

Definitions 14.1.3 (uniform metric)**.** Let (X, d) be a metric space in which every convergence $x_k \to x$ is uniform. Then the metric d is a uniform metric.

For example, the metric of $C[a, b]$ is a uniform metric.

A few examples of incomplete metric spaces

The following metric spaces are not complete,

1. The space of rational numbers, $\mathbb{Q}$ with usual metric $d(x, y) = |x-y|$ is not complete since the rational approximations of π is sequence in $\mathbb{Q}$ which doesn't converge in $\mathbb{Q}$.

$$3,\ 3.1,\ 3.14,\ 3.141,\ \cdots \to \pi \notin \mathbb{Q}$$

2. The space of polynomial functions with metric $d(x,y) = \sup\{x(t) - y(t)|$ is not complete since taylor approximations of $\sin x$ is a sequence of polynomial funcitons which doesn't converge to a polynomial function.
$$x,\ x + \frac{-x^3}{3!},\ x + \frac{-x^3}{3!} + \frac{x^5}{5!},\ \cdots \to \sin x \notin p$$
3. The space of continuous function on unit interval $[0,1]$ with a different†[5] metric
$$d(x,y) = \int_0^1 |x(t) - y(t)|\ dt$$
is not complete since $\{x_k\}$ where $x_k : [0,1] \to \mathbb{R}$ is defined by
$$x_k(t) = \begin{cases} 0 & t \in \left[0, \frac{1}{2}\right) \\ (t - \frac{1}{2})k & t \in \left[\frac{1}{2}, \frac{1}{2} + \frac{1}{k}\right) \\ 1 & t \in \left[\frac{1}{2} + \frac{1}{k}, 1\right] \end{cases}$$
doesn't converge to a continuous polynomial function.

14.1.6 Completion of Metric Space

Let Y be an incomplete metric space. Completion of Y is a construction of a complete metric space X in which Y is dense.

For example, $\mathbb{Q}$ is not complete. However, $\mathbb{R}$ is a complete metric space in which $\mathbb{Q}$ is complete. Therefore, $\mathbb{R}$ is a completion of $\mathbb{Q}$.

Definitions 14.1.4 (Isometry)**.** Isometric functions are distance preserving functions. And two metric spaces are isometric if there exists a bijective isometry between them.

For example, function $f : X \to Y$ is an isometry if
$$\forall x, y \in X, \quad \hat{d}(f(x), f(y)) = d(x,y)$$
where $d, \hat{d}$ are metrics in X, Y respectively. And if f is a bijection, then X, Y are isometric space.

Two metric spaces are isometric is another way of saying that the spaces are identical (same) from metric point of view.

Theorem 14.1.11 (completion)**.** *Let (X,d) be a metric space. Then, there exists a complete metric space, $(\hat{X}, \hat{d})$ such that it has a dense subspace W which is isometric with X. And $\hat{X}$ is unique upto isomerties.*

In other words, for any metric space X there exists a unique complete metric space, $\hat{X}$ in which X is dense.

Proof. **Step 1 : Construction of** $(\hat{X}, \hat{d})$
Let (X,d) be a metric space. Let C be the set of all Cauchy sequence in X. Define relation $\{x_k\} \sim \{y_k\}$ if and only if $\lim\limits_{k\to\infty} d(x_k, y_k) = 0$. This is an equivalence relation.(proof not required)

[5] Area under the graph of difference function is well-defined for integrable functions.

1. Reflexive - $\hat{x} \sim \hat{x}$ since $\lim_{k\to\infty} d(x_k, x_k) = 0$.
2. Symmetric - $\hat{x} \sim \hat{y} \implies \lim\limits_{k\to\infty} d(x_k, y_k) = \lim\limits_{k\to\infty} d(y_k, x_k) \implies \hat{y} \sim \hat{x}$.
3. Transitive - Suppose, $\hat{x} \sim \hat{y}$ and $\hat{y} \sim \hat{z}$.
   $\hat{x} \sim \hat{y} \implies \hat{d}(\hat{x}, \hat{y}) = \lim\limits_{k\to\infty} d(x_k, y_k) = 0$, $\hat{y} \sim \hat{z} \implies \hat{d}(\hat{y}, \hat{z}) = \lim\limits_{k\to\infty} d(y_k, z_k) = 0$.
   Then $\hat{d}(\hat{x}, \hat{z}) = \lim\limits_{k\to\infty} d(x_n, z_n) \le \lim\limits_{k\to\infty} d(x_n, y_n) + d(y_n, z_n) = \hat{d}(\hat{x}, \hat{y}) + \hat{d}(\hat{y}, \hat{z}) = 0$.
   Therefore $\hat{x} \sim \hat{z}$.

Let $\hat{X}$ be the set of all equivalent classes in C. Define $\hat{d} : \hat{X} \times \hat{X} \to \mathbb{R}$ given by $\hat{d}(\hat{x}, \hat{y}) = \lim\limits_{k\to\infty} d(x_k, y_k)$ where the Cauchy sequences $\{x_k\} \in \hat{x}$ and $\{y_k\} \in \hat{y}$. Then $\hat{d}$ is metric in $\hat{X}$.

1. $\hat{d}$ is well-defined
   Suppose $\{x_k\}, \{x'_k\} \in \hat{x}$ and $\{y_k\}, \{y'_k\} \in \hat{y}$. Then $\{x_k\} \sim \{x'_k\}$ and $\{y_k\} \sim \{y'_k\}$. In other words, $\lim\limits_{k\to\infty} d(x_k, x'_k) = 0$ and $\lim\limits_{k\to\infty} d(y_k, y'_k) = 0$.

   By triangular inequality, we have

$$d(x_k, y_k) \le d(x_k, x'_k) + d(x'_k, y'_k) + d(y'_k, y_k)$$
$$\implies d(x_k, y_k) - d(x'_k, y'_k) \le d(x_k, x'_k) + d(y_k, y'_k)$$

   Similarly,

$$d(x'_k, y'_k) \le d(x'_k, x_k) + d(x_k, y_k) + d(y_k, y'_k)$$
$$\implies d(x'_k, y'_k) - d(x_k, y_k) \le d(x_k, x'_k) + d(y_k, y'_k)$$

   Therefore,

$$|d(x_k, y_k) - d(x'_k, y'_k)| \le d(x_k, x'_k) + d(y_k, y'_k)$$

   Apply the limit $k \to \infty$ on either sides, we get

$$\lim_{k\to\infty} |d(x_k, y_k) - d(x'_k, y'_k)| \le \lim_{k\to\infty} d(x_k, x'_k) + \lim_{k\to\infty} d(y_k, y'_k) = 0$$

   Thus, $\hat{d}(\hat{x}, \hat{y})$ depends only on the equivalent class $\hat{x}, \hat{y}$ to which x_k, y_k belongs and is independent of the representative from these equivalent classes. Therefore, $\hat{d} : \hat{X} \times \hat{X} \to \mathbb{R}$ is well-defined.

2. $\hat{d}(\hat{x}, \hat{y}) = 0 \iff \hat{x} = \hat{y}$

$$\hat{d}(\hat{x}, \hat{y}) = 0 \iff \forall \{x_k\} \in \hat{x}, \forall \{y_k\} \in \hat{y}, \lim_{k\to\infty} d(x_k, y_k) = 0 \iff \{x_k\} \sim \{y_k\}$$

3. $\hat{d}(\hat{x}, \hat{y}) = \hat{d}(\hat{y}, \hat{x})$ is trivial since $d(x_k, y_k) = d(y_k, x_k)$.

4. $\hat{d}(\hat{x}, \hat{y}) \le \hat{d}(\hat{x}, \hat{z}) + \hat{d}(\hat{z}, \hat{y})$

   Let $\{z_k\}$ be a Cauchy sequence in X. Then $\{z_k\} \in \hat{z}$ for some $\hat{z} \in \hat{X}$,

$$d(x_k, y_k) \le d(x_k, z_k) + d(z_k, y_k)$$

   Applying limit $k \to \infty$ on either sides, we get

$$\hat{d}(\hat{x}, \hat{y}) = \lim_{k\to\infty} d(x_k, y_k) \le \lim_{k\to\infty} d(x_k, z_k) + \lim_{k\to\infty} d(z_k, y_k) = \hat{d}(\hat{x}, \hat{z}) + \hat{d}(\hat{z}, \hat{y})$$

Step 2: Construction of Isometry $T : X \to W,\ W \subset \hat{X}$.
Let $b \in X$. Then $b, b, \ldots$ is a Cauchy sequence in $\hat{X}$. Let $\hat{b} \in \hat{X}$ be the equivalent class of Cauchy sequences containing the constant sequence $\{b\}$. Define $T : X \to \hat{X}$ given by $T(b) = \hat{b}$.
Let $a, b \in X$, then

$$\begin{aligned} Ta = Tb &\implies \{a\} \sim \{b\} \\ &\implies \lim_{k\to\infty} d(a,b) = 0 \\ &\implies d(a,b) = 0 \implies a = b \end{aligned}$$

Thus $T : X \to T(X)$ is a bijection. Also T is an isometry since

$$\hat{d}(\hat{a}, \hat{b}) = \hat{d}(Ta, Tb) = \lim_{k\to\infty} d(a,b) = d(a,b)$$

Step 3: $W = T(X)$ is dense in $\hat{X}$
Let $\hat{x} \in \hat{X}$, then $\{x_k\} \in \hat{x}$. Then, $\{x_k\}$ is a Cauchy sequence.
Let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that $\forall m, n \geq N,\ d(x_n, x_m) < \frac{\varepsilon}{2}$.
Thus,

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N} \text{ such that } d(x_n, x_N) < \frac{\varepsilon}{2}$$

Clearly, $x_N \in X$ and the constant sequence $\{x_N\} \in \hat{x_N} \in T(X) = W$.
Thus,

$$\hat{d}(\hat{x}, \hat{x_N}) = \lim_{k\to\infty} d(x_k, x_N) \leq \frac{\varepsilon}{2} < \varepsilon$$

For every $\varepsilon > 0$, there exists $\hat{x}_N \in W$ such that $\hat{d}(\hat{x}, \hat{x}_N) < \varepsilon$. Therefore, W is dense in $\hat{X}$.

Step 4: $\hat{X}$ is complete
Let $\{\hat{x}_n\}$ be any Cauchy sequence in $\hat{X}$. Since W is dense in $\hat{X}$, for each $\hat{x}_n$ there exists $\hat{z}_n \in W$ such that $\hat{d}(\hat{x}_n, \hat{z}_n) < \frac{1}{n}$. Then, $\{\hat{z}_n\}$ is a Cauchy sequence since,

$$\hat{d}(\hat{z}_m, \hat{z}_n) \leq \hat{d}(\hat{z}_m, \hat{x}_m) + \hat{d}(\hat{x}_m, \hat{x}_n) + \hat{d}(\hat{x}_n, \hat{z}_n) \leq \frac{1}{m} + \hat{d}(\hat{x}_m, \hat{x}_n) + \frac{1}{n}$$

We have, $T : X \to W$ defined by $T(z_n) = \hat{z}_n$ is a bijective, isometry. Thus, sequence $\{z_n\}$ is a Cauchy sequence in X, since $\hat{d}(\hat{z}_n, \hat{z}_m) = d(z_n, z_m)$. Let $\hat{x} \in \hat{X}$ be the equivalent class of Cauchy sequences containing $\{z_n\}$. Then,

$$\hat{d}(\hat{x}_n, \hat{x}) \leq \hat{d}(\hat{x}_n, \hat{z}_n) + \hat{d}(\hat{z}_n, \hat{x}) = \frac{1}{n} + \hat{d}(\hat{z}_n, \hat{x})$$

Apply limit $n \to \infty$ on either sides, we get

$$\lim_{n\to\infty} \hat{d}(\hat{x}_n, \hat{x}) \leq \lim_{n\to\infty} \frac{1}{n} + \lim_{n\to\infty} \hat{d}(\hat{z}_n, \hat{x}) = 0$$

Thus, $\hat{x}_n \to \hat{x}$ as $n \to \infty$. Therefore $\hat{X}$ is complete, since every Cauchy sequence in $\hat{X}$ converges.

Step 5: Uniqueness of $\hat{X}$
Suppose $(\tilde{X}, \tilde{d})$ is a complete metric space with dense subset $\tilde{W}$ which is isometric with X. Let $\tilde{x}, \tilde{y} \in \tilde{X}$. Since $\tilde{W}$ is dense in $\tilde{X}$, there exists sequences $\{\tilde{x}_n\}, \{\tilde{y}_n\}$ in $\tilde{W}$ such that $\tilde{x}_n \to \tilde{x}$ and $\tilde{y}_n \to \tilde{y}$.
We have,

$$\tilde{d}(\tilde{x}, \tilde{y}) \leq \tilde{d}(\tilde{x}, \tilde{x}_n) + \tilde{d}(\tilde{x}_n, \tilde{y}_n) + \tilde{d}(\tilde{y}_n, \tilde{y})$$

$$\tilde{d}(\tilde{x}_n, \tilde{y}_n) \leq \tilde{d}(\tilde{x}_n, \tilde{x}) + \tilde{d}(\tilde{x}, \tilde{y}) + \tilde{d}(\tilde{y}, \tilde{y}_n)$$

Thus,

$$|\tilde{d}(\tilde{x}_n, \tilde{y}_n) - \tilde{d}(\tilde{x}, \tilde{y})| \leq \tilde{d}(\tilde{x}, \tilde{x}_n) + \tilde{d}(\tilde{y}, \tilde{y}_n)$$

Applying limit $n \to \infty$ on either sides, we get

$$\lim_{n\to\infty} |\tilde{d}(\tilde{x}_n, \tilde{y}_n) - \tilde{d}(\tilde{x}, \tilde{y})| \leq \lim_{n\to\infty} \tilde{d}(\tilde{x}, \tilde{x}_n) + \lim_{n\to\infty} \tilde{d}(\tilde{y}, \tilde{y}_n) = 0$$

$$\implies \tilde{d}(\tilde{x}, \tilde{y}) = \lim_{n\to\infty} \tilde{d}(\tilde{x}_n, \tilde{y}_n)$$

We have, $\tilde{W}$ is isometric†[6] with W. Also $\tilde{W}, W$ are dense in $\tilde{X}, \hat{X}$ respectively. Therefore, $\hat{X}$ and $\tilde{X}$ are isometric. In other words, the completion of X is unique except for isometries. □

14.2 Banach Spaces

Definitions 14.2.1 (vector space)**.** A set of vectors X, a field of scalars K together with vector addition $+ : X \times X \to X$ and scalar multiplication $\times : K \times X \to X$ is a vector(linear) space if it satisfies

1. $x + y = y + x$
2. $x + (y + z) = (x + y) + z$
3. $\exists 0 \in X$ such that $x + 0 = x$
4. $\forall x \in X,\ \exists -x \in X$ such that $x + (-x) = 0$
5. $\alpha(\beta x) = (\alpha\beta)x$
6. $1x = x$
7. $\alpha(x + y) = \alpha x + \alpha y$
8. $(\alpha + \beta)x = \alpha x + \beta x$

Note : $0x = 0$, $\alpha 0 = 0$ and $(-1)x = -x$.

For example, $\mathbb{R}^n, \mathbb{C}^n, C[a, b], l^p, l^\infty$ are vector spaces.

[6]Suppose $\tilde{W}$ is isometric with X. Then there exists an isometry $S : \tilde{W} \to X$ such that $\tilde{d}(\tilde{x}_n, \tilde{y}_n)) = d(S(\tilde{x}_n), S(\tilde{y}_n))$. Since $T : X \to W$ is also an isometry, $T^{-1} \circ S : \tilde{W} \to W$ is an isometry where $\tilde{d}(\tilde{x}_n, \tilde{y}_n) = \hat{d}(T^{-1} \circ S(\tilde{x}_n), T^{-1} \circ S(\tilde{y}_n))$.

14.2.1 A few definitions

Unlike other texts†[7]. And the **subspace** Y **generated a subset** M of a vector space X is the set of all linear combinations of vectors in M. That is,

$$Y = \text{span } M$$

A set M is **linear independent** if there exists a non-zero r-tuple $(\alpha_1, \alpha_2, \ldots, \alpha_r)$ such that $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_r x_r = 0$ where $x_1, x_2, \ldots, x_r \in M$.

A vector space X is **finite dimensional** if there exists an integer n which is the maximum cardinality for any linearly independent subset of X, say **dimension** of X, dim X.

Let X be a finite dimenstional vector space of dimension n. Then **basis** of X is a linearly independent subset of cardinality n.
Note : Let B be a basis of vector space X. Then $X = \text{span } B$. And every element $x \in X$ has a unique representation $x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$ where $B = \{x_1, x_2, \ldots, x_n\}$.

Let X be an infinite dimensional vector space. Let B be a linearly independent subset of X which span X, then B is a **Hamel basis** for X.

By Zorn's lemma, every vector space has a basis. And for any vector space, all the bases are of same cardinality (**dimension**).

Let X be a finite dimensional vector space of dimension n. Then any proper subspace Y of X has dimension less than n.

14.2.2 Banach Space

Definitions 14.2.2 (norm)**.** A real-valued function $\|\cdot\|$ on a vector space X, $\|\cdot\| : X \to \mathbb{R}$ is a norm if it satisfies

1. $\|x\| \geq 0, \quad \forall x \in X$
2. $\|x\| = 0 \iff x = 0$
3. $\|\alpha x\| = |\alpha|\|x\|, \quad \forall x \in X, \quad \forall \alpha \in K$
4. $\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in X$

Note : Let $\|\cdot\|$ be a norm on a vector space X. Then $d(x, y) = \|x - y\|$ is a metric induced by the norm.

Proof. Let $d(x, y) = \|x - y\|$.

1. $d(x, y) = \|x - y\| \geq 0$ since $x - y \in X$, $\|x - y\| \geq 0$
2. $d(x, y) = \|x - y\| = 0 \iff x - y = 0 \iff x = y$
3. $d(x, y) = \|x - y\| = \| - 1(y - x)\| = | - 1|\|y - x\| = d(y, x)$

[7]Usually, we define **space generated by a set** as the intersection of all spaces containing that set.

4. $d(x,y) = \|x - z + z - y\| \leq \|x - z\| + \|z - y\| = d(x,z) + d(z,y)$

□

normed space is a vector space with a norm defined on it.

Banach space is a complete, normed space.

Note : Norm is continuous.

Proof. Let X be a normed space. We have,

$$\lim_{h\to 0} \|x + hy\| \leq \|x\| + \lim_{h\to 0} |h|\|y\| = \|x\|$$

$$\|x\| \leq \lim_{h\to 0} \|x + hy\| + \lim_{h\to 0} |-h|\|y\| = \lim_{h\to 0} \|x + hy\|$$

Therefore, $\lim_{h\to 0} \|x + hy\| = \|x\|$. And the norm is continuous. □

Challenge 5. Characterise metric spaces, which are normed space ?
By defining $\|x\| = d(x, 0)$, we may construct a norm on various metric spaces so that the induced metric is the same. We observe the following conditions are necessary and sufficient for this function to be a norm.

Proof. Let (X, d) be a metric space. Let $d(x+z, y+z) = d(x, y)$ and $d(\alpha x, \alpha y) = |\alpha| d(x, y)$ for any $x, y, z \in X$ and for any $\alpha \in K$. Then,

1. $\|x\| = d(x, 0) \geq 0$

2. $\|x\| = d(x, 0) = 0 \iff x = 0$

3. $\|\alpha x\| = d(\alpha x, 0) = |\alpha| d(x, 0)$

4. $\|x + y\| \leq d(x + y, 0) \leq d(x + y, y) + d(y, 0) = d(x, 0) + d(y, 0) = \|x\| + \|y\|$

□

Lemma 14.2.1 (translational invariance)**.** *Let d be a metric induced by a norm on a normed space X. Then*

1. $d(x + z, y + z) = d(x, y), \quad \forall x, y, z \in X$

2. $d(\alpha x, \alpha y) = |\alpha| d(x, y), \quad \forall x, y \in X, \forall \alpha \in K$

Proof. Let $(X, \|\cdot\|)$ be a normed space. Define $d(x, y) = \|x - y\|$. Then,

1. $d(x + z, y + z) = \|x + z - (y + z)\| = \|x - y\| = d(x, y)$

2. $d(\alpha x, \alpha y) = \|\alpha x - \alpha y\| = \|\alpha(x - y)\| = |\alpha| \|x - y\| = |\alpha| d(x, y)$

□

Normed spaces

The following normed spaces can be easily constructed using the above construction.

1. $\mathbb{R}, \|x\| = |x|$
2. $\mathbb{R}^2, \|x\| = \left(|\xi_1|^2 + |\xi_2|^2\right)^{\frac{1}{2}}$ where $x = (\xi_1, \xi_2)$
3. $\mathbb{R}^3, \|x\| = \left(|\xi_1|^2 + |\xi_2|^2 + |\xi_3|^2\right)^{\frac{1}{2}}$ where $x = (\xi_1, \xi_2, \xi_3)$
4. Euclidean space, $\mathbb{R}^n, \|x\| = \left(\sum_{j=1}^{n} |\xi_j|^2\right)^{\frac{1}{2}}$ where $x = (\xi_1, \xi_2, \ldots, \xi_n)$
5. $\mathbb{C}, \|x\| = |x|$ where $x = a + ib$ and $|x| = (a^2 + b^2)^{\frac{1}{2}}$
6. Unitary space, $\mathbb{C}^n, \|x\| = \left(\sum_{j=1}^{n} |\xi_j|^2\right)^{\frac{1}{2}}$ where $x = (\xi_1, \xi_2)$ and $\xi_j \in \mathbb{C}$.
7. $C[a,b], \|x\| = \max_{t\in[a,b]} \{|x(t)|\}$ where x is a continuous, complex-valued function defined on closed interval $[a,b] \subset \mathbb{R}$.
8. $B(A), \|x\| = \sup_{t\in A} \{|x(t)|\}$ where x is a bounded, complex-valued function defined on $A \subset \mathbb{R}$
9. $l^p, \|x\| = \left(\sum_{j=1}^{\infty} |\xi_j|^p\right)^{\frac{1}{p}}$ where $x = \{\xi_j\}$ and $\sum_{j=1}^{\infty} |\xi_j|^p < \infty$. That is, set of all sequences such that the pth power series is convergent.
10. $l^\infty, \|x\| = \sup_{j\in\mathbb{N}}\{|\xi_j|\}$ where $x = \{\xi_j\}$ and $|\xi_j| \le c$. That is, l^∞ is the space of all bounded sequences of complex numbers.
11. $c, \|x\| = \sup_{j\in\mathbb{N}}\{|\xi_j|\}$ where $\{\xi_j\} = \xi_1, \xi_2, \ldots$ is the space of all convergent sequences of complex numbers.

Test for norm which can be induced from a metric

There exists metric which cannot be induced by any norm. Suppose (X, d) is a metric space which doesn't feature translational invariance. Suppose there exists a norm on X which induces d. Then d must have translational invariance which is a contradition.

For example, metric space of all sequence of complex numbers with metric

$$d(x,y) = \sum_{j=1}^{\infty} \frac{1}{2^j} \frac{|\xi_j - \eta_j|}{1 + |\xi_j - \eta_j|}$$

where $x = \{\xi_j\} = \xi_1, \xi_2, \ldots$ is not translationally invariant since,

$$d(\alpha x, \alpha y) = \sum_{j=1}^{\infty} \frac{1}{2^j} \frac{|\alpha\xi_j - \alpha\eta_j|}{1 + |\alpha\xi_j - \alpha\eta_j|} = \sum_{j=1}^{\infty} \frac{1}{2^j} \frac{|\xi_j - \eta_j|}{|\frac{1}{\alpha}| + |\xi_j - \eta_j|} \neq |\alpha| d(x, y)$$

Therefore, above metric cannot be induced from any norm defined on the sequence space s.

Completion of incomplete normed space

Consider the set of all continuous functions defined on closed interval $[a, b]$ together with the norm $\| \cdot \| : C[a, b] \to \mathbb{R}$ defined by

$$\|x\| = \left(\int_a^b |x(t)|^2 \, dt \right)^{\frac{1}{2}}$$

is an incomplete normed space. The completion of this space is $L^2[a, b]$, the collection of all Lebesgue measurable function x defined on closed interval $[a, b]$ such that $|x|^2$ is Lebesgue integrable .

Consider the collection of continuous functions defined on $[a, b]$ with norm,

$$\|x\| = \left(\int_a^b |x(t)|^p \, dt \right)^{\frac{1}{p}}$$

We have its completion $L^p[a, b]$, the collection of all Lebesgue measurable function x defined on closed $[a, b]$ such that $|x|^p$ are Lebesgue integrable.

14.2.3 Further properties of normed spaces

1. Let X be a Banach space. And Y is a subset of X. The subspace Y is Banach if and only if Y is closed in X.

2. Every normed space X has a Banach space $\hat{X}$ in which X is dense.

Theorem 14.2.2. *Let X be a Banach space. And Y is a subset of X. The subspace Y is complete if and only if Y is closed in X.*

Proof. Let $(X, \| \cdot \|)$ be a Banach space. Then (X, d) is a metric space with induced metric $d : X \times X \to \mathbb{R}$ defined by $d(x, y) = \|x - y\|$. Let $Y \subset X$. Clearly, Y is a normed space with induced norm. And if Y is closed then every Cauchy sequence in Y does converge to a point in Y. Therefore, Y is a Banach space.

Suppose Y is a Banach space with induced norm. Since Y is complete, every convergent sequence in Y converges to a point in Y itself. Therefore, Y is a closed subset of X. $\square$

Convergence in normed spaces

We have the notion of convergence defined for metric spaces. Now, we may extend that notion to normed spaces.

Definitions 14.2.3 (convergence)**.** A sequence $\{x_n\}$ is a normed X converges to $x \in X$ if
$$\lim_{n\to\infty} \|x_n - x\| = 0$$

A series in a normed space $(X, \|\cdot\|)$ is convergent if the sequence of partial sums converges. If it converges, the limit of the sequence of partial sums is the **sum** of the series. And series is **absolutely convergent** if the series of norms converges.

Schauder basis

Definitions 14.2.4 (Schauder basis)**.** A sequence $\{e_n\}$ of normed space $(X, \|\cdot\|)$ is a Schauder basis of X if for any $x \in X$, there exists a unique sequence of scalars $\{\alpha_n\}$ such that the series $\sum_{n=1}^{\infty} \alpha_n e_n$ converges to x.

$$\lim_{n\to\infty} \left\| x - \sum_{k=1}^{n} \alpha_k e_k \right\| = 0$$

Note : Every normed space with a Schauder basis is separable. But there exists separable, Banach spaces without a Schauder basis.

Completion of normed spaces

Theorem 14.2.3 (completion)**.** *Let $(X, \|\cdot\|)$ be a normed space. Then there exists a Banach space $\hat{X}$ with a dense subset W and an isometry A from W onto X. And $\hat{X}$ is unique except for isometries.*

Proof. Let $(X, \|\cdot\|)$ be a normed space. Then (X, d) where $d : X \times X \to \mathbb{R}$ defined by $d(x, y) = \|x - y\|$ is a metric space. By the completion of metric spaces, there exists a unique, complete metric space $\hat{X}$ with a dense subset W and an isometry from W onto X.

Step 1 : $\hat{X}$ is a vector space
Let $\hat{x}, \hat{y} \in \hat{X}$. Let sequence $\{x_n\}$ be some Cauchy sequence in $\hat{x}$ and sequence $\{y_n\}$ be some Cauchy sequence in $\hat{y}$.

1. The vector set, $\hat{X}$ is the equivalent classes of Cauchy sequence in X.

2. The scalar field is real/complex field K of numbers.

3. Vector addition $+ : \hat{X} \times \hat{X} \to \hat{X}$ is defined by $\hat{x} + \hat{y} = \hat{z}$. Let $\{x_n\} \in \hat{x}$, $\{y_n\} \in \hat{y}$. Then $\{x_n + y_n\} = \{z_n\} \in \hat{z}$.

   Suppose $\hat{x}, \hat{y} \in \hat{X}$. We need prove that vector addition is well-defined as well as closed. In other words, $\hat{x} + \hat{y}$ is independent of the choice of

$\{x_n\} \in \hat{x}$, $\{y_n\} \in \hat{y}$ and is an equivalent class of Cauchy sequences.

Step 1.3a : Vector Sum is well-defined
Let $\{x_n\}, \{x'_n\} \in \hat{x}$ and $\{y_n\}, \{y'_n\} \in \hat{y}$. Then, $\{x_n\} \sim \{x'_n\}$ and $\{y_n\} \sim \{y'_n\}$. By triangular inequality,

$$\|x_n + y_n - (x'_n + y'_n)\| \leq \|x_n - x'_n\| + \|y_n - y'_n\|$$

Applying limit $n \to \infty$ on either sides we get,

$$\lim_{n\to\infty} \|x_n + y_n - (x'_n + y'_n)\| \leq \lim_{n\to\infty} \|x_n - x'_n\| + \lim_{n\to\infty} \|y_n - y'_n\| = 0$$

Clearly, $\{x_n + y_n\} \sim \{x'_n + y'_n\}$ and the vector sum $\hat{x} + \hat{y}$ is consistent, independent of the choice of the sequences.

Step 1.3b: Vector addition is closed
Let $\hat{x}, \hat{y} \in \hat{X}$. Then every sequence $\{x_n\} \in \hat{x}$ and every sequence $\{y_n\} \in \hat{y}$ are Cauchy sequences. Then,

$$\forall \varepsilon > 0,\ \exists N_1 \in \mathbb{N} \text{ such that } \forall n, m \geq N_1,\ \|x_n - x_m\| < \frac{\varepsilon}{2}$$

$$\forall \varepsilon > 0,\ \exists N_2 \in \mathbb{N} \text{ such that } \forall n, m \geq N_2,\ \|y_n - y_m\| < \frac{\varepsilon}{2}$$

We need to prove that $\hat{x}+\hat{y}$ is also an equivalent class of Cauchy sequences. Let $N = \max\{N_1, N_2\}$. Then $\forall \varepsilon > 0$, $\forall n, m \geq N$ we have,

$$\|x_n + y_n - (x_m + y_m)\| \leq \|x_n - x_m\| + \|y_n - y_m\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

4. Scalar multiplication $\times : K \times \hat{X} \to \hat{X}$ defined by $\alpha\hat{x}$ is the collection of Cauchy sequences equivalent to $\{\alpha x_n\}$ where $\{x_n\} \in \hat{x}$ and $\alpha \in K$.

We need to prove that scalar product is well-defined and is closed.
Let $\hat{x} \in \hat{X}$ and $\alpha \in K$.

Step 1.4a : Scalar product is well-defined
Let $\{x_n\}, \{x'_n\} \in \hat{x}$. Then $\{x_n\} \sim \{x'_n\}$. And both sequences are Cauchy sequences.

$$\lim_{n\to\infty} \|\alpha x_n - \alpha x'_n\| = |\alpha| \lim_{n\to\infty} \|x_n - x'_n\| = 0$$

Therefore, $\alpha\hat{x}$ is the equivalent class of sequences which are equivalent to sequence $\{\alpha x_n\}$ where sequence $\{x_n\} \in \hat{x}$. Clearly, $\alpha\hat{x}$ is independent of the choice of sequence $\{x_n\} \in \hat{x}$.

Step 1.4b: Scalar product is closed
We have, $\{x_n\}$ is a Cauchy sequence. Then,

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N} \text{ such that } \forall n, m \geq N,\ \|x_n - x_m\| < \frac{\varepsilon}{|\alpha|}$$

Clearly, $\forall n, m > N$

$$\|\alpha x_n - \alpha x_m\| = |\alpha| \ \|x_n - x_m\| < \varepsilon$$

Therefore, $\alpha\hat{x}$ is an equivalent class of Cauchy sequences.

Therefore, $\hat{X}$ is vector space.

Step 2 : Construction of norm

We introduce a norm on $\hat{X}$ such that the metric $\hat{d}$ coincides with the induced metric.

Define $\|\cdot\| : W \to \mathbb{R}$ as $\|\hat{x}\| = \hat{d}(\hat{x}, \hat{0})$ for every $\hat{x} \in W$. We know that $A : W \to X$ is an isometry. Therefore, $\hat{d}(\hat{x}, \hat{y}) = d(x, y)$ for every $\hat{x}, \hat{y} \in W$ where $A(\hat{x}) = x$ and the constant sequence $\{x\} \in \hat{x}$. Clearly, for each $\hat{x} \in W$, $\|\hat{x}\| = \hat{d}(\hat{x}, \hat{0}) = d(x, 0) = \|x\|$ where $x \in X$.

Let $\hat{x} \in \hat{X}$. We know that, W is dense in $\hat{X}$, there exists a Cauchy sequence $\{\hat{x}_n\}$ in W such that $\hat{d}(\hat{x}_n, \hat{x}) < \frac{1}{n}$. Thus, there exists a sequence $\{\hat{x}_n\}$ in W such that $\hat{x}_n \to \hat{x}$. Now we may extend the norm function to $\hat{X}$

$$\|\cdot\| : \hat{X} \to \mathbb{R} \text{ defined by } \|\hat{x}\| = \lim_{n\to\infty} \|\hat{x}_n\| \tag{14.8}$$

where $\{\hat{x}_n\} \in W$ and $\lim_{n\to\infty} \hat{d}(\hat{x}_n, \hat{x}) = 0$.

Now, we show that $\|\cdot\|$ defined on $\hat{X}$ is a norm

1. We have, $\hat{d}$ is metric on $\hat{X}$. Thus, $\hat{d}(\hat{x}, \hat{y}) \geq 0, \ \forall \hat{x}, \hat{y} \in \hat{X}$.

$$\|\hat{x}\| = \lim_{n\to\infty} \|\hat{x}_n\| = \lim_{n\to\infty} \hat{d}(\hat{x}_n, \hat{0}) \geq 0$$

2. Suppose $\hat{x} = \hat{0}$. Then $\hat{x} \in W$ and $\|\hat{x}\| = \hat{d}(\hat{x}, \hat{0}) = 0$.
Suppose $\|\hat{x}\| = 0$. Then,

$$\|\hat{x}\| = \lim_{n\to\infty} \|\hat{x}_n\| = \lim_{n\to\infty} \hat{d}(\hat{x}_n, \hat{0}) = \lim_{n\to\infty} d(x_n, 0) = 0$$

Clearly, sequence $\{x_n\}$ is a Cauchy sequence in X converging to $0 \in X$. Therefore, $\{x_n\} \in \hat{0}$ and $\hat{x}_n \to \hat{0}$. But, we know that $\hat{x}_n \to \hat{x}$. Since $\hat{X}$ is a metric space, $\hat{x} = \hat{0}$.

3. We have,

$$\|\alpha\hat{x}\| = \lim_{n\to\infty} \|\alpha\hat{x}_n\| = |\alpha| \lim_{n\to\infty} \|\hat{x}_n\| = |\alpha|\|\hat{x}\|$$

4. Suppose $\hat{y} \in \hat{X}$. Since W is dense in $\hat{X}$, there exists a Cauchy sequence $\{\hat{y}_n\}$ in W such that $\hat{y}_n \to \hat{y}$. Thus,

$$\|\hat{x} + \hat{y}\| = \lim_{n\to\infty} \|\hat{x}_n + \hat{y}_n\| \leq \lim_{n\to\infty} \|\hat{x}_n\| + \lim_{n\to\infty} \|\hat{y}_n\| = \|\hat{x}\| + \|\hat{y}\|$$

Thus, $(\hat{X}, \|\cdot\|)$ is complete, normed space in which W is dense.

Step 3 : $\hat{X}$ is unique
Suppose $\tilde{X}$ is another Banach space with dense subset $\tilde{W}$ which has a isometry $\tilde{A}$ from $\tilde{W}$ onto X. Then, the norm $\|\cdot\|$ defined on $\tilde{W}$,

$$\|\tilde{x}\| = \tilde{d}(\tilde{x}, \tilde{0}) = d(x, 0),\ \forall \tilde{x} \in \tilde{W}$$

And we may extend this norm to $\tilde{X}$,

$$\|\tilde{x}\| = \lim_{n\to\infty} \|\tilde{x}_n\|$$

where sequence $\{\tilde{x}_n\}$ is a Cauchy sequence in $\tilde{W}$ and $\tilde{x}_n \to \tilde{x}$.

Clearly, $\tilde{W}$ is isometric with W and the topological closure of $\tilde{W}$ is isometric with the topological closure of W. That is, $\tilde{X}$ is isometric with $\hat{X}$. □

Exercise

Definitions 14.2.5 (seminorm)**.** A real-valued function p on a vector space X, $p : X \to \mathbb{R}$ is a seminorm if it satisfies

1. $p(x) \geq 0, \quad \forall x \in X$
2. $x = 0 \implies p(x) = 0$
3. $p(\alpha x) = |\alpha| p(x), \quad \forall x \in X, \quad \forall \alpha \in K$
4. $p(x + y) \leq p(x) + p(y), \quad \forall x, y \in X$

Lemma 14.2.4. *Let $(X, \|\cdot\|)$ be a normed space. Then, the vector addition $+(x, y) = x + y$ and scalar multiplication $\times(\alpha, x) = \alpha x$ are continuous with respect to norm.*

Proof. Let $x, y \in X$ and $\alpha \in K$. Let sequences $\{x_n\}, \{y_n\}$ converges to x, y respectively. Let sequence $\{\alpha_n\}$ converges to $\alpha \in K$. Then,

$$\lim_{n\to\infty} \|x_n + y_n - (x + y)\| \leq \lim_{n\to\infty} \|x_n - x\| + \lim_{n\to\infty} \|y_n - y\| = 0$$

And,

$$\begin{aligned}
\lim_{n\to\infty} \|\alpha_n x_n - \alpha x\| &= \lim_{n\to\infty} \|\alpha_n x_n - \alpha x_n + \alpha x_n - \alpha x\| \\
&\leq \lim_{n\to\infty} \|(\alpha_n - \alpha) x_n\| + \lim_{n\to\infty} \|\alpha(x_n - x)\| \\
&= \lim_{n\to\infty} |\alpha_n - \alpha| \lim_{n\to\infty} \|x_n\| + |\alpha| \lim_{n\to\infty} \|x_n - x\| \\
&= 0
\end{aligned}$$

□

Lemma 14.2.5. *Let $(X, \|\cdot\|)$ be a normed space. Let sequence $\{x_n\}$ converges to $x \in X$ and sequences $\{y_n\}$ converges to $y \in X$. Then, sequence $\{x_n + y_n\}$ converges to $x + y$. Let sequence $\{\alpha_n\}$ converges to α. Then sequence $\{\alpha_n x_n\}$ converges to αx.*

Proof. Suppose $x_n \to x$, $y_n \to y$ and $\alpha_n \to \alpha$. We have, vector addition and scalar multiplication are continuous with repsect to norm.

$$\lim_{n\to\infty} \|x_n + y_n - (x+y)\| = 0 \text{ and } \lim_{n\to\infty} \|\alpha_n x_n - \alpha x\| = 0$$

Therefore, $x_n + y_n \to x + y$ and $\alpha_n x_n \to \alpha x$. □

Lemma 14.2.6 (separable)**.** *Let $(X, \|\cdot\|)$ be normed space. Suppose X has a Schauder basis. Then, X is separable.*

Proof. Seminar - Akshaya Peter
refer : Exercise 2.3.10
hint : Let $B = \{e_1, e_2, \dots\}$ be a Schauder basis of normed space $(X, \|\cdot\|)$. Then B is a dense subset of X. □

Lemma 14.2.7 (quotient space)**.** *Let $(X, \|\cdot\|)$ be a normed space. Let Y be a closed subset of X. Then $(X/Y, \|\cdot\|_0)$ is a normed space with norm defined by*

$$\|\hat{x}\|_0 = \inf_{x\in\hat{x}} \|x\|, \qquad \textit{for any coset } \hat{x} \in X/Y$$

Proof. Seminar - Amala Mathew
refer : Exercise 2.3.14
hint : You can skip algebraic details. N1,N2 are trivial. N3 - use $\alpha\hat{x} = \alpha x + Y$ where $\hat{x} = x + Y$. □

Lemma 14.2.8 (product of normed spaces)**.** *Let $(X_1, \|\cdot\|_1)$, $(X_2, \|\cdot\|_2)$ be normed spaces. Let $X = X_1 \times X_2$. Then $(X, \|\cdot\|)$ is a normed space with norm defined by,*

$$\|x\| = \max\{\|x_1\|_1, \|x_2\|_2\} \; \textit{where } x = (x_1, x_2)$$

Proof. Seminar - Rinkum Susan Punnoose
refer : Exercise 2.3.15
hint : Let $x = (x_1, x_2)$. WLOG suppose $\|x_1\|_1 \geq \|x_2\|_2$. Then $\|x\| = \|x_1\|_1$ and the rest is obvious. □

14.2.4 Finite dimensional normed spaces and subspaces

Lemma 14.2.9. *Let $(X, \|\cdot\|)$ be a normed space. Let $\{x_1, x_2, \dots, x_n\}$ be any linearly independent subset of X. Then there exists a real-number $c > 0$ such that for any subset $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ from field K*

$$\left\| \sum_{j=1}^{n} \alpha_j x_j \right\| \geq c \sum_{j=1}^{n} |\alpha_j| \tag{14.9}$$

Proof. Let $\{x_1, x_2, \dots, x_n\}$ be a linearly independent subset of a normed space X.

$$\left\| \sum_{j=1}^{n} \alpha_j x_j \right\| \geq c \sum_{j=1}^{n} |\alpha_j|$$

Let $s = \sum_{j=1}^{n} |\alpha_j|$. Suppose $s \neq 0$. Consider $\beta_j = \alpha_j/s$. Then,

$$\left\| \sum_{j=1}^{n} \beta_j x_j \right\| \geq c \text{ where } \sum_{j=1}^{n} |\beta_j| = 1$$

It is sufficient to prove that for any real-number $c > 0$, there does not exists an n-tuple of scalars violating the above inequality.

On the contrary, suppose that for every positive real number $c \in \mathbb{R}$ there exists a subset $\{\beta_1, \beta_2, \ldots, \beta_n\}$ such that $\|\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n\| < c$. Then, for each $m \in \mathbb{N}$, there exist a subset of n scalars $\{\beta_{m,1}, \beta_{m,2}, \ldots, \beta_{m,n}\}$ such that

$$\|\beta_{m,1} x_1 + \beta_{m,2} x_2 + \cdots + \beta_{m,n} x_n\| < \frac{1}{m}$$

Define, $y_m = \beta_{m,1} x_1 + \beta_{m,2} x_2 + \cdots + \beta_{m,n} x_n$. Then, $\|y_m\| \to 0$ as $m \to \infty$.

Now, every $\beta_{j,k}$ is bounded since for each m, $|\beta_{m,1}| + |\beta_{m,2}| + \cdots + |\beta_{m,n}| = 1$. The sequence $\{\beta_{m,1}\}$ of first terms of the sequence $\{y_m\}$ is a bounded sequence. By Bolzano-Weierstrass theroem, sequence $\{\beta_{m,1}\}$ has a convergent subsequence, say $\beta_{m,1} \to 1$ as $m \to \infty$. Let sequence $\{y_{1,m}\}$ be subsequence of sequence $\{y_m\}$ such that the sequence of first terms is convergent.

Again, sequence $\{\beta_{m,2}\}$ of second terms of the sequence $\{y_{1,m}\}$ is a bounded sequence. And $\{y_{1,m}\}$ has a subsequence $\{y_{2,m}\}$ such that both the sequence of first terms $\{\beta_{m,1}\}$ and $\{\beta_{m,2}\}$ are convergent, say $\beta_{2,m} \to \beta_2$ as $m \to \infty$.

Continuing like this n times, we get sequence $\{y_{n,m}\}$ such that sequence corresponding each term is convergent. That is, $\beta_{m,j} \to \beta_j$ for $j = 1, 2, \ldots, n$.

Define $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$. Clearly, the sequence $\{y_{n,m}\}$ is a convergent subsequence of $\{y_m\}$. And $y_{n,m} \to y$ as $m \to \infty$. Suppose $\{y_{n,m}\}$ is convergent, then $y_{n,m} \to y$ as the limit is unique. By the continuity of norm, we have $\|y_m\| \to \|y\|$.

Since $\|y\| = 0$, we have $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = 0$. We have $\{x_1, x_2, \ldots, x_n\}$ is linearly independent. Then, $\beta_j = 0$ for each j. And, for each m, sequence $\{|\beta_{m,n}|\}$ converges to 0 and $|\beta_{m,1}| + |\beta_{m,2}| + \cdots + |\beta_{m,n}| = 1$. This is a contradition. Therefore, there exists such a positive real-number c satisfying the lemma. □

Theorem 14.2.10. *Every finite dimensional subspace of a normed space is complete.*

Proof. Let Y be a finite dimensional subspace of a normed space X. Let *dim* $Y = n$ and $B = \{y_1, y_2, \ldots, y_n\}$ be a basis for Y. Let sequence $\{y_m\}$ be a Cauchy sequence in Y. Then,

$$y_m = \alpha_{m,1} y_1 + \alpha_{m,2} y_2 + \cdots + \alpha_{m,n} y_n$$

Since $\{y_m\}$ is a Cauchy sequence, we have $N \in \mathbb{N}$ such that $\forall s, t > N$

$$\|y_s - y_t\| = \left\| \sum_{j=1}^{n} \alpha_{s,j} y_j - \sum_{j=1}^{n} \alpha_{t,j} y_j \right\| < \varepsilon$$

Since basis of Y is a linearly independent subset of X, applying lemma we get

$$c \sum_{j=1}^{n} |\alpha_{s,j} - \alpha_{t,j}| \leq \left\| \sum_{j=1}^{n} (\alpha_{s,j} - \alpha_{t,j}) y_j \right\| < \varepsilon$$

Therefore, for each $j \in \mathbb{N}$, sequence $\{\alpha_{m,j}\}$ are Cauchy sequences in K. Suppose $\alpha_{m,j} \to \alpha_j$. Define $y = \alpha_1 y_1 + \alpha_2 y_2 + \alpha_n y_n$. Then, $y_m \to y$ as $m \to \infty$. Therefore, Y is complete. □

Theorem 14.2.11. *Every finite dimensional subspace Y of a normed space X is closed in X.*

Proof. Let Y be a finite dimensional subspace of a normed space X. Then Y is complete. Let $\hat{X}$ be the completion of X. Now, Y is complete subspace of a Banach space $\hat{X}$. Then, Y is closed $\hat{X}$. And $Y \subset X \subset \hat{X}$. Therefore, Y is closed in X. □

Equivalent Norm

Consider normed space $(X, \|\cdot\|)$. Then (X, d) is a metric space with induced metric $d(x, y) = \|x - y\|$. Since every metric space is a topological space, there exists a unique topology $(X, \mathcal{T})$ with respect to its norm. Let $\|\cdot\|_0, \|\cdot\|_1$ be two different norms defined a vector space X. Then, the norms are equivalent if their topologies coincide.

Alternatively, we have the following inequality to characterise the equivalent of norms,

Definitions 14.2.6. Let X be a vector space. Let $\|\cdot\|_0$, $\|\cdot\|_1$ be two normed defined on X. These norms are equivalent if there exists positive, real numbers a, b such that for any $x \in X$,

$$a\|x\|_0 \leq \|x\|_1 \leq b\|x\|_0 \tag{14.10}$$

For a topologist, it will interesting to prove that the topologies coincide if and only if the above inequality holds.

Theorem 14.2.12. *Every norm on any finite dimensional normed spaces are equivalent.*

Let $(X, \|\cdot\|)$ be a finite dimensional normed space. Then there exists a unique norm on X upto equivalence. That is, the topology induced by norm is independent of the norm for finite dimensional normed spaces.

Proof. Let X be a vector space with $dim\ X = n$ and basis $B = \{x_1, x_2, \ldots, x_n\}$. Let $\|\cdot\|_1$, $\|\cdot\|_2$ be two norms defined on X. Let $x \in X$. Then,

$$x = \sum_{j=1}^{n} \alpha_j x_j$$

We know that, basis is a linearly independent subset of X itself. By the lemma for finite dimensional, normed spaces, there exists a positive, real number c such that

$$\|x\|_1 = \left\| \sum_{j=1}^{n} \alpha_j x_j \right\|_1 \geq c \sum_{j=1}^{n} |\alpha_j|$$

By triangular inequality, we have

$$\|x\|_1 = \left\| \sum_{j=1}^{n} \alpha_j x_j \right\| \leq \sum_{j=1}^{n} \|\alpha_j x_j\| = \sum_{j=1}^{n} |\alpha_j| \|x_j\| \leq k \sum_{j=1}^{n} |\alpha_j|$$

where $k = \max\{\|x_j\|\}$.

Let $a = \frac{c}{k}$. Then $a\|x\|_1 \leq \|x\|_2$. Similarly, we can obtain a positive, real number b such that $b\|x\|_2 \leq \|x\|_1$. Therefore, the norms are equivalent. □

14.2.5 Compactness and Finite dimension

Definitions 14.2.7 (compact)**.** Let X be a metric space. The space X is **compact** if every sequences in X has a convergent subsequence. A subset M is a **compact subset** if every sequence in M has a convergent subsequence.

Lemma 14.2.13 (compact)**.** *Compact subsets of metric spaces are closed and bounded.*

Proof. Let C be a compact subset of metric space X. Let $\tilde{X}$ be the completion of X. Let $\{x_n\}$ be a Cauchy sequence in C. Since, C is compact $\{x_n\}$ has a convergent subsequence which converges to $x \in C$. Clearly, $\{x_n\}$ is a Cauchy sequence in $\tilde{X}$. Then, $\{x_n\} \to x \in \tilde{X}$. Thus, every Cauchy seqeuence in C converges. Therefore, C is closed, since every complete space is closed.

Let $y \in C$. Suppose C is unbounded. Then, there exists a sequence $\{y_n\}$ in C such that $B(y_n, y) > n$. Clearly, $\{y_n\}$ doesn't have any convergent subsequence which is a contradiction. Therefore, C is bounded. □

Theorem 14.2.14. *In a finite dimensional metric space, a subset M is compact if and only if M is closed and bounded.*

Proof. Let X be a metric space with $\dim X = n$. Let $C \subset X$. Suppose C is compact. Then, C is closed and bounded.

Suppose C is closed and bounded. Let $\{y_n\}$ be a sequence in C. Let $\{x_1, x_2, \ldots, x_n\}$ be a basis for X. Then $y_j = \sum_{k=1}^{n} \xi_{j,k} x_k$. By lemma of linear combinations, we have

$$\|y_j\| = \| \sum_{k=1}^{n} \xi_{j,k} x_k \| \geq c \sum_{k=1}^{n} |\xi_{j,k}|$$

Clearly, for each j, sequences $\{\xi_{j,k}\}$'s are bounded. By Bolzano-Weierstrass theorem, there exists a convergent subsequence converging to ξ_j. Define, $y = \sum_{k=1}^{n} \xi_k x_k$. Then $\{y_n\}$ has a subsequence which converges to y (obtained by successive construction of subsequences.) Since C is closed, $y \in C$. Let $\tilde{X}$ be the completion of X. Then $\{y_n\}$ has a subsequence converging to $y \in \tilde{X}$. Thus, every sequence in C has a convergent subseqeunce. Therefore, C is compact. $\square$

Lemma 14.2.15 (Riesz)**.** *Let Y, Z be subspaces of a normed space X. Suppose Y is closed and is a proper subset of Z. Then for any real number $\theta \in (0,1)$, there exists $z \in Z$ such that*

$$\|z\| = 1 \text{ and } \|z - y\| \geq \theta, \ \forall y \in Y$$

Proof. Let X be a normed space. Let Y be a closed subspace of X. And Z is a subspace of X properly containing Y. Thus, $Z - Y$ is non-empty. Let $v \in Z - y$. Define $a = \inf\{\|v - y\| : y \in Y\}$. †[8] Let $\theta \in (0,1)$. Then, $a < \frac{a}{\theta}$. And there exists $y_0 \in Y$ such that $a < \|v - y_0\| < \frac{a}{\theta}$. Define $z = \frac{v - y_0}{\|v - y_0\|}$. Then $\|z\| = 1$. And $\|z - y\| = c\|v - y_0 - c^{-1}y\|$ where $c^{-1} = \|v - y_0\|$. Clearly, $y_0 + c^{-1}y \in Y$ since Y is subspace and $y, y_0 \in Y$. Thus, $\|z - y\| = c\|v - (y_0 + c^{-1}y)\| \geq ca \geq \theta$ since $c^{-1} = \|v - y_0\| \leq \frac{a}{\theta}$. $\square$

Theorem 14.2.16 (compact set characterisation of finiteness)**.** *Let X be normed space in which closed unit ball is compact. Then X is a finite dimensional normed space.*

Proof. Let X be normed space. Let $M = \{x : \|x\| = 1\}$ be a compact subset of X. Suppose $\dim X = \infty$. Choose $x_1 \in X$ such that $\|x_1\| = 1$. For any $x \in X$, if $x \neq 0$, then there exists $x/\|x\| \in X$ has norm 1.

The subspace X_1 spanned by x_1 is a proper, closed subspace of X. Since X is infinite dimensional, from Riesz's lemma there exists $x_2 \in X$ such that $\|x_2\| = 1$ and $\|x_2 - x_1\| \geq \frac{1}{2}$. Again, the subspace X_2 spanned by $\{x_1, x_2\}$ is a proper, closed subspace of X. Continuing like this, we get a sequence $\{x_n\}$ in M such that $\|x_n - x_m\| \geq \frac{1}{2}$ for any $n \neq m$. Since $\|x_n - x_m\| \geq \frac{1}{2}$, the sequence doesn't have a convergent subsequence. This is a contradiction since M is compact. Therefore, X is finite dimensional. $\square$

Theorem 14.2.17. *Let X, Y be metric spaces. Let $T : X \to Y$ be a continuous function. Then, the image of compact subset of X under T is a compact subset of Y.*

Proof. Let $T : X \to Y$ be a continuous function where X, Y are metric spaces. Let M be a compact subset of X. Let sequence $\{y_n\}$ be a sequence in $T(M)$. Then $y_n = Tx_n$ for some $x_n \in M$. Thus, we have a sequence $\{x_n\}$ in M corresponding to each sequence $\{y_n\}$ in $T(M)$. Since M is compact, every sequence in M has a convergent subsequence, say $\{x_{n_k}\} \to x$. Then, the corresponding sequence $\{Tx_{n_k}\}$ is subsequence of sequence $\{y_n\}$.

Since T is continuous, $\{Tx_{n_k}\}$ converges to Tx in $T(M)$. Thus, every sequence in $T(M)$ has a convergent subsequence. That is, M is compact. $\square$

[8] Simply, $a = \|v - Y\|$ where $v \in Z - Y$

Corollary 14.2.17.1. *A continuous mapping T of a compact subset M ofa metric space X into $\mathbb{R}$ assumes a maximum and a minimum value at some points of M.*

Proof. Let X be a metric space and M be a compact subset of X. Let $T : X \to \mathbb{R}$ be a continuous function. Then, $T(M)$ is compact. Since $T(M)$ is a compact subset of $\mathbb{R}$, $T(M)$ is a closed and bounded subset of $\mathbb{R}$. Thus, $\inf T(M), \sup T(M) \in T(M)$. Therefore, there exists $x, y \in M$ such that $Tx = \inf T(M) = \min T(M)$ and $Ty = \sup T(M) = \max T(M)$. $\square$

Module 2 - Linear Operators

14.2.6 Linear Operators

Definitions 14.2.8 (linear operator)**.** Let X, Y be vector spaces with the same field K. A function $T : X \to Y$ is a linear operator if it satisfies

1. $T(x + y) = Tx + Ty, \ \forall x, y \in X$
2. $T(\alpha x) = \alpha Tx, \ \forall x \in X, \forall \alpha \in K$

Usually, we address a linear, function from a vector space into another as linear transformation. And linear operators are functions from a vector space into itself. Kreyszig uses linear operator for both.

1. $T(\alpha x + \beta y) = T(\alpha x) + T(\beta y) = \alpha Tx + \beta Ty$
2. $T0 = T(0x) = 0Tx = 0$
3. $T : X \to T(X)$ is a homomorphism of vector(linear) spaces

Examples of linear operators

1. Identity operator, $I : X \to X, \ Ix = x, \forall x \in X$
$$I(\alpha x + \beta y) = \alpha Ix + \beta Iy = \alpha x + \beta y$$
2. Zero operator, $O : X \to X, \ Ox = 0, \forall x \in X$
$$0(\alpha x + \beta y) = \alpha 0(x) + \beta 0(y) = 0$$
3. Differentiation operator, $T : X \to X, \ Tx(t) = x'(t), \forall t \in [a, b]$ where X is the set of all polynomials over $[a, b]$.
$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y) = \alpha x' + \beta y'$$
4. Integration operator, $T : X \to X$, $Tx(t) = \int_a^b k(t, \tau)x(\tau)d\tau$ where $X = C[a, b]$.
$$T(\alpha x + \beta y) = \alpha \left(\int_a^b k(t, \tau)x(\tau)d\tau \right) + \beta \left(\int_a^b k(t, \tau)y(\tau)d\tau \right)$$
5. Multiplication by parameter, $T : X \to X$, $Tx(t) = tx(t)$ where $X = C[a, b]$.
$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y) = \alpha tx + \beta ty$$
6. $T : \mathbb{R}^3 \to \mathbb{R}^3$, $T(x) = x \times a$ where $a \in \mathbb{R}^3$.
$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y) = \alpha(x \times a) + \beta(y \times a)$$
7. $T : \mathbb{R}^3 \to \mathbb{R}$, $T(x) = x \cdot a$ where $a \in \mathbb{R}^3$.
$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y) = \alpha(x \cdot a) + \beta(y \cdot a)$$

8. $T : \mathbb{R}^n \to \mathbb{R}^r$, $T(x) = Ax$ where $A \in \mathbb{R}^{n\times r}$ is an $n \times r$ matrix of real-numbers.
$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y) = \alpha Ax + \beta Ay$$

Theorem 14.2.18 (Range,Null space)**.** *Let T be a linear operator. Then*

1. *Range, $\mathscr{R}(T)$ is a vector space.*

2. *If $\dim X = n < \infty$, then $\dim \mathscr{R}(T) \leq n$.*

3. *Null space, $\mathscr{N}(T)$ is a vector space.*

Proof. Let $T : X \to Y$ be a linear operator where X, Y are vector spaces over the same field K. Let $y_1, y_2 \in \mathscr{R}(T)$. Then $y_1 = Tx_1$ and $y_2 = Tx_2$ for some $x_1, x_2 \in X$. Let $\alpha, \beta \in K$. Since X is a vector space, $\alpha x_1 + \beta x_2 \in X$. We have, T is a linear operator. Thus,

$$\alpha y_1 + \beta y_2 = \alpha Tx_1 + \beta Tx_2 = T(\alpha x_1 + \beta x_2)$$

Therefore, $\alpha y_1 + \beta y_2 \in \mathscr{R}(T)$.

Let $\dim X = n$. Let $\{y_1, y_2, \ldots, y_{n+1}\}$ be elements from $\mathscr{R}(T)$. Then for each j, we have $y_j = Tx_j$. Since $\dim X = n$, the set $\{x_1, x_2, \ldots, x_{n+1}\}$ is linearly dependent. That is, there exist α_j's in K such that

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{n+1} x_{n+1} = 0$$

Therefore,

$$T(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{n+1} x_{n+1}) = \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_{n+1} y_{n+1} = 0$$

Thus, any subset of $\mathscr{R}(T)$ having $n+1$ or more elements is linearly dependent. Clearly, the dimension of $\mathscr{R}(T) \leq n = \dim X$.
In other words, linear operators preserves linear dependence.

Let $x_1, x_2 \in \mathscr{N}(T)$. Then $Tx_1 = 0$ and $Tx_2 = 0$. Let $\alpha, \beta \in K$. Since T is linear,
$$T(\alpha x_1 + \beta x_2) = \alpha T(x_1) + \beta T(x_2) = 0$$
Therefore, $\alpha x_1 + \beta x_2 \in \mathscr{N}(T)$ for any $\alpha, \beta \in K$. □

Definitions 14.2.9 (Inverse)**.** Let X, Y be vector spaces over the same field K. Let $T : X \to Y$ be an injective, linear operator. Then, $T^{-1} : \mathscr{R}(T) \to X$ defined by $y = Tx \iff T^{-1}y = x$ is the image operator of T.

Theorem 14.2.19 (inverse)**.** *Let X, Y be vector spaces over the same field K. Let $T : X \to Y$ be a linear operator. Then,*

1. *T^{-1} exists if and only if ($Tx = 0 \implies x = 0$).*

2. *If T^{-1} exists, then it is a linear operator.*

3. *If $\dim \mathscr{D}(T) = n < \infty$, and T^{-1} exists, then $\dim \mathscr{R}(T) = \dim \mathscr{D}(T)$.*

Proof. Suppose $Tx = 0 \implies x = 0$. Let $Tx_1 = Tx_2$. Then $Tx_1 - Tx_2 = T(x_1 - x_2) = 0$ That is, $x_1 - x_2 = 0$. Therefore, T is injective. Suppose T^{-1} exists, then T is injective and $Tx_1 = Tx_2 \implies x_1 = x_2$. Thus, $Tx = Tx_1 - Tx_2 = 0 \implies x = x_1 - x_2 = 0$.

Let $T : X \to Y$ be a linear operator where X, Y are vector spaces over the same field K. Suppose T^{-1} exists. Let $y_1 = Tx_1$ and $y_2 = Tx_2$. Then, $T^{-1}y_1 = x_1$ and $T^{-1}y_2 = x_2$. Let $\alpha, \beta \in K$. We have,

$$\alpha x_1 + \beta x_2 = \alpha T^{-1}(y_1) + \beta T^{-1}(y_2)$$

Also we have,

$$T(\alpha x_1 + \beta x_2) = \alpha Tx_1 + \beta Tx_2 = \alpha y_1 + \beta y_2$$

$$\implies \alpha x_1 + \beta x_2 = T^{-1}(\alpha y_1 + \beta y_2) = \alpha T^{-1}y_1 + \beta T^{-1}y_2$$

Therefore, T^{-1} is a linear operator.

Since $T : X \to Y$ is a linear transformation. If $\dim X < \infty$, then $\mathscr{R}(T) \leq \dim \mathscr{D}(T)$. Again, $T^{-1} : \mathscr{R}(T) \to X$ is a linear transformation. Thus, $\dim \mathscr{D}(T) \leq \dim \mathscr{R}(T)$. Therefore, $\dim \mathscr{D}(T) = \dim \mathscr{R}(T)$. $\square$

Theorem 14.2.20 (Invese of Product)**.** *Let $T : X \to Y$ and $S : Y \to Z$ be bijective linear operators, where X, Y, Z are vector spaces. Then the inverse $(ST)^{-1} : Z \to X$ exists and $(ST)^{-1} = T^{-1}S^{-1}$.*

Proof. The operator $ST : X \to Z$ is bijective since S, T are bijective. Therefore, $(ST)^{-1} : Z \to X$ exists and $ST(ST)^{-1} = I_Z$.

$$S^{-1}ST(ST)^{-1} = T(ST)^{-1} = S^{-1}I_Z = S^{-1} \text{ and}$$
$$T^{-1}S^{-1}ST(ST)^{-1} = T^{-1}T(ST)^{-1} = (ST)^{-1} = T^{-1}S^{-1}$$

$\square$

14.2.7 Bounded and Continuous Linear Operators

Definitions 14.2.10 (bounded operator)**.** Let X, Y be normed spaces and $T : X \to Y$ be a linear operator. Then T is bounded if there exists a real number c such that $\|Tx\| \leq c\|x\|$.

Suppose $T : X \to Y$ be a bounded linear operator. Then $\|T\| : B(X, Y) \to \mathbb{R}$ is well-defined.

$$\|T\| = \sup_{\substack{x \in \mathscr{D}(T) \\ x \neq 0}} \frac{\|Tx\|}{\|x\|}$$

Lemma 14.2.21. *Let $T : X \to Y$ be a bounded linear operator. Then,*

1. *We have an alternate formula,*

$$\|T\| = \sup_{\substack{x \in \mathscr{D}(T) \\ \|x\| = 1}} \|Tx\|$$

2. $\|T\|$ satisfies norm axioms

Proof. Suppose $x \neq 0$ and $\|x\| = a$. Then $a \neq 0$ since X is a normed space. Define $y = \frac{1}{a}x$. Then $\|y\| = 1$. Also we have,

$$\|T\| = \sup_{\substack{x \in \mathscr{D}(T) \\ x \neq 0}} \frac{1}{a}\|Tx\| = \sup_{\substack{x \in \mathscr{D}(T) \\ x \neq 0}} \left\| T\left(\frac{1}{a}x\right) \right\| = \sup_{\substack{y \in \mathscr{D}(T) \\ \|y\|=1}} \|Ty\|$$

---

1. We have
$$\|T\| = \sup_{\substack{x \in \mathscr{D}(T) \\ x \neq 0}} \frac{\|Tx\|}{\|x\|}$$
Since X, Y are normed spaces $\|Tx\|, \|x\| \geq 0$ and $\|T\| \geq 0$.

2. Also we have, Y is normed space and $Tx \in Y$. Thus, $\|Tx\| = 0$ if and only if $Tx = 0$. And, $\|T\| = 0$ if and only if $Tx = 0,\ \forall x \in \mathscr{D}(T)$. Therefore, $T = 0$. That is, $Tx = 0x = 0,\ \forall x \in \mathscr{D}(T)$.

3. Let X, Y be normed spaces over the field K. Let $\alpha \in K$ and T be a linear operator from X into Y. Then,
$$\begin{aligned}
\|(\alpha T)\| &= \sup_{\substack{x \in \mathscr{D}(\alpha T) \\ \|x\|=1}} \|(\alpha T)(x)\| \\
&= \sup_{\substack{x \in \mathscr{D}(T) \\ \|x\|=1}} \||\alpha T(x)\| \\
&= |\alpha| \sup_{\substack{x \in \mathscr{D}(T) \\ \|x\|=1}} \|Tx\| \\
&= |\alpha| \|T\|
\end{aligned}$$

4. Let X, Y be normed spaces and T_1, T_2 are linear operators from X to Y. Then,
$$\begin{aligned}
\|T_1 + T_2\| &= \sup_{\substack{x \in \mathscr{D}(T_1+T_2) \\ \|x\|=1}} \|(T_1 + T_2)(x)\| \\
&= \sup_{\substack{x \in \mathscr{D}(T_1+T_2) \\ \|x\|=1}} \|T_1(x) + T_2(x)\| \\
&\leq \sup_{\substack{x \in \mathscr{D}(T_1) \\ \|x\|=1}} \|T_1(x)\| + \sup_{\substack{x \in \mathscr{D}(T_2) \\ \|x\|=1}} \|T_2(x)\| \\
&\leq \|T_1\| + \|T_2\|
\end{aligned}$$

□

Norm of a few Linear operators

1. Identity Operator $I_X : X \to X$ defined by $I(x) = x$.

$$\|I\| = \sup_{\substack{x \in \mathscr{D}(I) \\ \|x\|=1}} \|Ix\| = 1$$

2. Zero Operator $0 : X \to Y$ defined by $0x = 0$.

$$\|0\| = \sup_{\substack{x \in \mathscr{D}(0) \\ \|x\|=1}} \|0x\| = 0$$

3. Differential Operator $D : K[t] \to K[t]$ defined by $D(x(t)) = x'(t)$ where $K[t]$ is the vector space of all polynomial functions defined on $[0, 1]$ together with norm $\|x\| = \sup_{t \in [0,1]} |x(t)|$.

   The differential operator is not bounded. Consider $x(t) = t^n$ where $n \in \mathbb{N}$. Then $D(x) = x'(t) = nt^{n-1}$. Thus, $\|x\| = 1$ and $\|D(x)\| = n$. Clearly, differential operator D is not bounded. Therefore, $\|D\|$ is not defined.

4. Integral Operator $T : C[0, 1] \to C[0, 1]$ defined by

$$T(x) = \int_0^1 k(t, \tau)x(\tau)d\tau$$

   where $k(t, \tau)$ is the kernel of the integral operator. Clearly, the integral of any continuous function is zero if only if $k(t, \tau) = 0$ for any $\tau \in [0, 1]$.

$$\begin{aligned} \|T(x)\| &= \left\| \int_0^1 k(t, \tau)\ x(\tau)\ d\tau \right\| \\ &\leq \int_0^1 \|k(t, \tau)\| \|x(\tau)\| d\tau \end{aligned}$$

   Let $k_0 = \max_{\tau \in [0,1]} \{\|k(t, \tau)\}\}$. Replacing $\|k(t, \tau)\|$ with k_0, we get

$$\|T(x)\| \leq k_0 \|x\| \int_0^1 d\tau = k_0 \|x\|$$

   Therefore, integral operator is bounded and $\|T\| \leq k_0$.

5. Matrix multiplication $T : \mathbb{R}^n \to \mathbb{R}^r$ defined by $y = Ax$ where $A = (\alpha_{jk})$ where $\alpha_{jk} \in K$. Let $x = (\xi_k)$ and $y = (\eta_j)$ where $\xi_k, \eta_j \in K$. Then

$$\eta_j = \sum_{k=1}^{n} \alpha_{jk} \xi_k$$

Therefore,

$$\begin{aligned}\|Tx\|^2 &= \sum_{j=1}^{r} \eta_j^2 \\ &= \sum_{j=1}^{r} \left(\sum_{k=1}^{n} \alpha_{jk} \xi_k \right)^2\end{aligned}$$

By Cauchy-Schwarz inequality

$$\begin{aligned}\|T(x)\|^2 &\leq \sum_{j=1}^{r} \left(\sum_{k=1}^{n} \alpha_{jk}^2 \sum_{k=1}^{n} \xi_k^2 \right) \\ &\leq \sum_{j=1}^{r} \sum_{k=1}^{n} \alpha_{jk}^2 \ \sum_{j=1}^{r} \sum_{k=1}^{n} \xi_k^2 \\ &\leq c^2 \|x\|^2 \\ \|Tx\| &\leq c\|x\|\end{aligned}$$

where $c = \left(\sum_{j=1}^{r} \sum_{k=1}^{n} \alpha_{jk}^2 \right)^{\frac{1}{2}}$. Clearly, matrix multiplication is bounded.

Theorem 14.2.22 (finite dimension)**.** *Every linear operator on finite dimensional normed space is bounded.*

Proof. Let X be normed space with $\dim(X) = n < \infty$. Let $T : X \to X$ be a linear operator on X. Then,

$$\begin{aligned}\|Tx\| &= \left\| T\left(\sum_{j=1}^{n} \xi_j e_j \right) \right\| \\ &= \left\| \sum_{j=1}^{n} \xi_j T(e_j) \right\| \quad \text{since } T \text{ is linear} \\ &\leq \sum_{j=1}^{n} |\xi_j| \ \|T(e_j)\|\end{aligned}$$

Let e_k is an element of the standard basis of X such that $\|T(e_k)\|$ is maximum. Then,

$$\|T(x)\| \leq \|T(e_k)\| \sum_{j=1}^{n} |\xi_j|$$

We have, the standard basis of X, $\{e_1, e_2, \dots, e_n\}$ is a linearly independent set of vectors. We have $x = \sum_{j=1}^{n} \xi_j e_j$. Thus, by linear combination lemma, there exists a positive real number c such that $\sum_{j=1}^{n} |\xi_j| \leq \frac{1}{c} \left\| \sum_{j=1}^{n} \xi_j e_j \right\| = \frac{1}{c}\|x\|$. Thus,

$$\|T(x)\| \leq \frac{\|T(e_k)\|}{c}\|x\|$$

Therefore, linear operators on finite dimensional spaces are bounded. □

Theorem 14.2.23. *Let X, Y be normed spaces and $T : X \to Y$ be a linear operator. Then*

1. *T is continuous if and only if T is bounded.*
2. *T is continuous at a point, then it is continuous.*

Proof. Suppose $T \neq 0$. If $T = 0$, then the results are trivial as zero operator is always continuous and bounded. Since $T \neq 0$, $\|T\| \neq 0$.

Suppose T is bounded. Let $\varepsilon > 0$. Define $\delta = \frac{\varepsilon}{\|T\|}$.
Let $x_0 \in \mathscr{D}(T)$. Let $x \in \mathscr{D}(T)$ such that $\|x - x_0\| < \delta = \frac{\varepsilon}{\|T\|}$. Now, we have

$$\|Tx - Tx_0\| = \|T(x - x_0)\| \leq \|T\| \; \|x - x_0\| \leq \delta\|T\| \leq \varepsilon$$

Therefore, T is continuous.

Let T be continuous at $x_0 \in \mathscr{D}(T)$. Then, for any $\varepsilon > 0$, there exists $\delta > 0$ such that if $\|x - x_0\| \leq \delta$, then $\|T(x) - T(x_0)\| \leq \varepsilon$.

Let $y \neq 0$. Define $x \in X$ such that $x = x_0 + \frac{\delta}{\|y\|}y$. Then $\|x - x_0\| = \delta$. Therefore,

$$\varepsilon \geq \|Tx - Tx_0\| = \|T(x - x_0)\| = \frac{\delta}{\|y\|}\|Ty\|$$

And $\|Ty\| \leq \frac{\varepsilon}{\delta}\|y\|$. Therefore, T is bounded.

Suppose T is continuous at a point x_0. Then T is bounded and therefore T is continuous on $\mathscr{D}(T)$. □

Corollary 14.2.23.1. *Let T be a bounded linear operator.*

1. $x_n \to x \implies Tx_n \to Tx$
2. *Null space $\mathcal{N}(T)$ is closed.*

Proof. Let $T : X \to Y$ be a bounded linear operator. Suppose $x_n \to x$.

$$\lim_{n\to\infty} \|Tx_n - Tx\| = \lim_{n\to\infty} \|T(x_n - x)\| \leq \|T\| \; \|x_n - x\| = 0$$

Clearly, $Tx_n \to Tx$.

Suppose $x \in \overline{\mathcal{N}(T)}$. †[9] Then there exists a sequence $\{x_n\} \in \mathcal{N}(T)$ such that $x_n \to x$. Then, $Tx_n \to Tx$. However, $Tx_n = 0$ since $x_n \in \mathcal{N}(T)$. Therefore, $Tx = 0$ and $x \in \mathcal{N}(T)$. Clearly, $\mathcal{N}(T)$ is closed. □

Remark. Range of a bounded linear operator is not necessarily closed.

Proof. Consider $T : l^\infty \to l^\infty$ defined by $Tx = y$ where $x = (\xi_j)$, $y = (\eta_j)$ and $\eta_j = \xi_j/j$. Since l^∞ is not complete, there exists a sequence $\{x_n\}$ in l^∞ which coverges to an unboudned sequence $x \notin l^\infty$.

For example, consider $x_n = (\xi_{n,j})$ where†[10]

$$\xi_{n,j} = \begin{cases} j - \frac{1}{n} & j \le n \\ 0 & j > n \end{cases}$$

Clearly, $x_n \in l^\infty$ since x_n is bounded by n. And $\{Tx_n\} = (\eta_{n,j})$ defined by $\eta_{n,j} = \frac{\xi_{n,j}}{j}$ is a sequence of bounded functions which converges to $y = (\eta_j)$ where $\eta_j = 1$. However, there doesn't exist a bounded sequence x such that $Tx = y$. Therefore, the range of T is not closed. □

Remark. Let $T_2 : X \to Y$ and $T_1 : Y \to Z$ be bounded linear opertors where X, Y, Z are normed spaces. Then $\|T_1 T_2\| = \|T_1\| \, \|T_2\|$.

Proof. Suppose T_1, T_2 are bounded, linear opertors. Then T_1T_2 is linear since $T_1(T_2(\alpha x + \beta y)) = T_1(\alpha T_2 x + \beta T_2 y) = \alpha T_1 T_2 x + \beta T_1 T_2 y$.

And T_1T_2 is bounde since

$$\begin{aligned} \|T_1T_2(x)\| &\le \|T_1\| \, \|T_2(x)\| \text{ since } T_1 \text{ is bounded} \\ &\le \|T_1\| \, \|T_2\| \, \|x\| \text{ since } T_2 \text{ is bounded} \\ &\le c\|x\| \text{ where } c = \|T_1\| \, \|T_2\| \end{aligned}$$

Now we have,

$$\begin{aligned} \|T_1T_2\| &= \sup_{\substack{x \in \mathscr{D}(T_2) \\ \|x\|=1}} \|T_1T_2x\| \\ &\le \sup_{\substack{x \in \mathscr{D}(T_2) \\ \|x\|=1}} \|T_1\| \, \|T_2\| \, \|x\| \\ &\le \|T_1\| \, \|T_2\| \end{aligned}$$

□

---

[9] To prove X is closed, we often choose $x \in \overline{X}$ which is limit of an arbitrary sequence x_n in X. Then it is enough to show that the limit $x \in X$.

[10] We know that l^∞ is the set of all bounded sequences. We may use simpler sequence $x_n = (\xi_{n,j})$ where $\xi_{n,j} = j$, $j \le n$ and vanishes for $j > n$. That is, $x_1 = 1, 0, 0, \ldots$, $x_2 = 1, 2, 0, 0, \ldots$ and $x_n = 1, 2, \ldots, n, 0, 0, \ldots$. Then $Tx_n = 1, 1, \ldots, 1, 0, 0, \ldots$. Clearly, $x_n \to y$ where $y = 1, 2, 3, \ldots$ is unbounded and $Tx_n \to Ty$ where $Ty = 1, 1, 1, \ldots$ is bounded.

Remark. Let $T : X \to X$ be a bounded, linear operator. Then $\|T^2\| \le \|T\|^2$. And by mathematical induction, $\forall n \in \mathbb{N},\ \|T^n\| \le \|T\|^n$.

Definitions 14.2.11 (equal operator)**.** Two linear operators T_1, T_2 are equal if $\mathscr{D}(T_1) = \mathscr{D}(T_2)$ and $\forall x \in \mathscr{D}(T_1),\ T_1x = T_2x$.

Theorem 14.2.24 (bounded linear extension)**.** *Let $T : X \to Y$ be a bounded, linear operator from a normed space into a Banach space. Then T has a bounded, linear extension $\tilde{T} : \overline{\mathscr{D}(T)} \to Y$ such that $\|\tilde{T}\| = \|T\|$.*

Proof. Suppose $x \in \overline{\mathscr{D}(T)}$. Then there exists a sequence $\{x_n\}$ in $\mathscr{D}(T)$ such that $x_n \to x$. By Cauchy-criterion, there exists $N \in \mathbb{N}$ such that $\forall n, m > N,\ \|x_n - x_m\| \le \varepsilon$. Since T is a bounded linear operator, for every $n, m > N$ we have

$$\|Tx_n - Tx_m\| = \|T(x_n - x_m)\| \le \|T\|\ \|x_n - x_m\| \le \varepsilon\|T\|$$

Therefore, sequence $\{Tx_n\}$ in Y converges to a point, say $y \in Y$ since Y is complete. Define $\tilde{T} : \overline{\mathscr{D}(T)} \to Y$ such that $\tilde{T}(x) = y$.

Now, we show that the function $\tilde{T}$ is defined independent of the choice of the sequence $\{x_n\}$. Let $x_n \to x$ and $z_n \to x$. Consider $\{v_n\} : x_1, z_1, x_2, z_2, \ldots$. Then $v_n \to x$. And $\{Tv_n\}$ converges. Then, the subsequences $\{Tx_n\}$ and $\{Tzn\}$ converges and they must have the same limit. Thus, $\tilde{T}$ is uniquely defined (well defined).

Suppose $x_n \to x$ and $z_n \to z$. Then $\tilde{T}$ is linear since

$$T(\alpha x_n + \beta z_n) = \alpha Tx_n + \beta Tz_n \to \alpha\tilde{T}x + \beta\tilde{T}z = \tilde{T}(\alpha x + \beta z)$$

Since $\tilde{T}$ is an extension of T, $\tilde{T}x = Tx$ in the domain of T. Also we have, $\|Tx_n\| \le \|T\|\ \|x_n\|$. Suppose $Tx_n \to y = \tilde{T}x$. We have, the norm function which maps $x \to \|x\|$ is continuous. As $n \to \infty$ we have, $\|\tilde{T}x\| \le \|T\|\ \|x\|$. Thus, $\tilde{T}$ is bounded and $\|\tilde{T}\| \le \|T\|$. Since $\tilde{T}$ is an extension of T, we also have $\|T\| \le \|\tilde{T}\|$. Therefore, $\|\tilde{T}\| = \|T\|$. □

Challenge 6. Any linear operator on X has a linear extension into $\overline{X}$ preserving norm. Whether this extension is unique on $\overline{X}$?

14.2.8 Linear Functionals

Definitions 14.2.12 (linear functional)**.** Let X be a vector space over field K. Then linear functional on X is a linear operator $f : X \to K$.

Definitions 14.2.13 (bounded linear functional)**.** A linear functional on X is bounded if there exists a real number c such that

$$|f(x)| \le c\|x\|, \quad \forall x \in X$$

Theorem 14.2.25. *Linear functional on a normed space is continuous ifand only if bounded.*

Proof. Let f be a linear functional on a normed space X. Then f can be viewed as a linear operator from normed space X into normed space K. We know that, linear operators are continuous if and only if bounded. Since f is a linear operator, f is continuous if and only if f is bounded. □

Norm

Let X be a normed space. Then norm is a non-linear functional on X.

Let X be a normed space. Then $\|\cdot\| : X \to \mathbb{R}$ is a non-linear functional on X. The functional $\|\cdot\|$ is non-linear due to triangular inequality.

Dot Product

The dot product is a bounded linear functional on any finite dimensional Euclidean space.

Let $a \in \mathbb{R}^3$. Then $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x) = x \cdot a$ is a bounded, linear functional on $\mathbb{R}^3$. We know that dot product is linear.

$$(\alpha x + y) \cdot a = \alpha x \cdot a + y \cdot a$$

By Cauchy-Schwarz's inequality, we have

$$\|x \cdot a\| = \left\|\sum \xi_j a_j\right\| \leq \left\|\left(\sum \xi_j^2\right) \frac{1}{2} \ \left(\sum a_j^2\right)^{\frac{1}{2}}\right\| \leq \|x\| \ \|a\|$$

Thus f is bounded and for any $x \in X$, we have

$$|f(x)| = \|x \cdot a\| \leq \|x\| \ \|a\|$$

This inequality is true for any $x \in X$ and the supremum as well.

$$\|f\| = \sup_{\|x\|=1} |f(x)| \leq \sup_{\|x\|=1} \|x\| \ \|a\| = \|a\|$$

When $x = a$ we have,

$$\|f\| = \sup_{x \neq 0} \frac{|f(x)|}{\|x\|} \geq \frac{|f(a)|}{\|a\|} = \frac{|a \cdot a|}{\|a\|} \leq \frac{\|a\|^2}{\|a\|} = \|a\|$$

Thus norm of the dot product, $\|f\| = \|a\|$ where $f(x) = x \cdot a$.

Definite Integral

Let $X = C[a, b]$, the set of all continuous functions on closed interval $[a, b]$ with norm $\|x\| = \sup_{x \in [a,b]} |x(t)|$. Let $f(x) = \int_a^b x(t) \ dt$.

We know that, f is linear, since

$$\begin{aligned}
f(\alpha x + \beta y) &= \int_a^b (\alpha x + \beta y)(t) \ dt \\
&= \int_a^b (\alpha x)(t) + (\beta y)(t) \ dt \\
&= \int_a^b \alpha x(t) \ dt + \int_a^b \beta y(t) \ dt \\
&= \alpha \int_a^b x(t) \ dt + \beta \int_a^b y(t) \ dt \\
&= \alpha f(x) + \beta f(y)
\end{aligned}$$

And f is bounded.

$$|f(x)| = \left|\int_a^b x(t)\ dt\right| \leq \int_a^b |x(t)|\ dt \leq \|x\| \int_a^b dt \leq (b-a)\|x\|$$

since $\|x\| = \sup|x(t)|$, we have $|x(t)| \leq \|x\|$ for any $t \in [a,b]$.

Thus, $\|f\| \leq (b-a)$. And for constant function x_0 on $[a,b]$ where $x_0(t) = 1$, $\forall t \in [a,b]$, we know that $\|x_0\| = \sup|x_0(t)| = 1$. And $f(x_0) = \int_a^b dt = (b-a)$.

$$\|f\| = \sup_{x \neq 0} \frac{|f(x)|}{\|x\|} \leq \frac{|f(x_0)|}{\|x_0\|} = (b-a)$$

Therefore, $\|f\| = (b-a)$ where $f(x) = \int_a^b f(x)\ dt$.

A linear functional on l^2 space

Let $a \in l^2$. That is, $a = (\alpha_j)$ is a sequence such that $\sum |\alpha_j|^2 < \infty$. Consider $f(x) = \displaystyle\sum_{j=1}^{\infty} \xi_j \alpha_j$ where $x = (\xi_j) \in l^2$.
We have $x, a \in l^2$. Thus, $\sum \xi_j^2 < \infty$ and $\sum \alpha_j^2 < \infty$.

We have, $f(x)$ is absolutely convergent since

$$\sum |\xi_j \alpha_j| \leq \left(\sum \xi_j^2)\right)^{\frac{1}{2}} \ \left(\sum \alpha_j\right)^{\frac{1}{2}} < \infty$$

By Cauchy-Schwarz's inequality,

$$\sum (\xi_j \alpha_j)^2 \leq \sum |\xi_j \alpha_j|^2 \leq \sum \xi_j^2 \ \sum \alpha_j^2 < \infty$$

Therefore, $f(x) \in l^2$. Clearly, $f(x)$ is well-defined.

Let $x, y \in l^2$, $x = (\xi_j)$ and $y = (\eta_j)$. Then, f is linear since,

$$f(\alpha x + \beta y) = \sum_{j=1}^{\infty} (\alpha \xi_j + \beta \eta_j)\alpha_j = \alpha \sum_{j=1}^{\infty} \xi_j \alpha_j + \beta \sum_{j=1}^{\infty} \eta_j \alpha_j = \alpha f(x) + \beta f(y)$$

By Cauchy-Schwarz's inequality, we have

$$|f(x)| = \left|\sum_{j=1}^{\infty} \xi_j \alpha_j\right| \leq \sum_{j=1}^{\infty} |\xi_j \alpha_j| \leq \left(\sum_{j=1}^{\infty} \xi_j\right)^{\frac{1}{2}} \ \left(\sum_{j=1}^{\infty} \alpha_j\right)^{\frac{1}{2}} = \|x\| \ \|a\|$$

Thus,

$$\|f\| = \sup_{\|x\|=1} |f(x)| \leq \sup_{\|x\|=1} \|x\|\|a\| = \|a\|$$

And when $x = a$, we get $\|f\| = \sup_{x \neq 0} |f(x)| \geq |f(a)| = \|a\|$. Therefore, $\|f\| = \|a\|$.

Remark. Let X be a normed space over field K. Then the set X^* of all linear functional on X is a vector space over the same field K.

Proof. Let X^* be the set of all linear functional on X. Then X^* is a vector space with vector addition $f+g$ defined as $(f+g)(x) = f(x)+g(x)$ and scalar multplicaton (αf) defined as $(\alpha f)(x) = \alpha f(x)$.

Let $f, g \in X^*$.

$$\begin{aligned}(f+g)(\alpha x+\beta y) &= f(\alpha x+\beta y)+g(\alpha x+\beta y)\\ &= \alpha f(x)+\beta f(y)+\alpha g(x)+\beta g(y)\\ &= \alpha(f+g)(x)+\beta(f+g)(y)\end{aligned}$$

$$\begin{aligned}(\gamma f)(\alpha x+\beta y) &= \gamma f(\alpha x+\beta y)\\ &= \gamma\alpha f(x)+\gamma\beta f(y)\\ &= \alpha(\gamma f)(x)+\beta(\gamma f)(y)\end{aligned}$$

Then $\alpha f+\beta g \in X^*$. Therefore, vector addition is closed.

Let $f, g \in X^*$. Then $(f+g)(x) = f(x)+g(x) = g(x)+f(x) = (g+f)(x)$. Thus, vector addition is commutative.

Consider $0 : X \to K$ defined by $0(x) = 0$, $\forall x \in X$. Clearly $0 \in X^*$ since $0(\alpha x+\beta y) = 0 = \alpha 0(x)+\beta 0(x)$. And $(f+0)(x) = f(x)+0(x) = f(x)$. Thus, $0 \in X^*$ is the additive identity.

Let $f \in X^*$. Consider $-f : X \to K$ defined by $(-f)(x) = -f(x)$, $\forall x \in X$. Clearly, $-f \in X^*$ since $-f(\alpha x+\beta y) = \alpha(-f(x))+\beta(-f(y))$. Then, $(f+-f)(x) = f(x)+(-f)(x) = 0 = 0(x)$. Thus, every linear functional f in X^* has a unique additive inverse in X^*.

$\vdots$

$\square$

Definitions 14.2.14 (Algebraic Dual Space)**.** Let X be a normed space. The set X^* of all linear functionals on X is the algebraic dual space of X.

Definitions 14.2.15 (Second Algebraic Dual Space)**.** Let X be a normed space and X^* be its algebraic dual space of all linear functionals on X. Then the algebraic dual of X^*, $(X^*)^*$ is the second algebraic dual X^{**} of X.

Remark. Let $g \in X^{**}$ and $f \in X^*$. Then $g(f) = f(x)$ for some fixed $x \in X$.

Remark. Let $g_x : X^* \to K$ defined by $g_x(f) = f(x)$. Then g_x is linear since

$$\begin{aligned}g_x(\alpha f_1+\beta f_2) &= (\alpha f_1+\beta f_2)(x)\\ &= (\alpha f_1)(x)+(\beta f_2)(x)\\ &= \alpha f_1(x)+\beta f_2(x)\\ &= \alpha g_x(f_1)+\beta g_x(f_2)\end{aligned}$$

Definitions 14.2.16 (canonical mapping)**.** Let $C : X \to X^{**}$ defined by $x \to g_x$ where $C(x) = g_x$ and $g_x(f) = f(x),\ \forall f \in X^*$. Then C is a canonical mapping on X.

Remark. Canonical mapping on X is an isomorphism of X into a subset of X^{**}.

Definitions 14.2.17 (algebraic reflexivity)**.** Let X be a normed space. If the canonical mapping $C : X \to X^{**}$ is surjective, then X is isomorphic to X^{**}. Then, X is **algebraically reflexive** since X^{**} is an algebraic reflection of X.

Definitions 14.2.18 (isomorphism)**.** Isometry/Isomorphism of a metric space is a bijection that preserves metric/distance.

$$T : X \to Y \text{ such that } \forall x, y \in X,\ \tilde{d}(Tx, Ty) = d(x, y)$$

Isomorphism of a vector space is a bijection that preserves both the binary operations, say vector addition and scalar multiplication.

$$T : X \to Y \text{ such that } \forall x, y \in X, \forall \alpha, \beta \in K,\ T(\alpha x + \beta y) = \alpha Tx + \beta Ty$$

Isomorphism of a normed space is a vector space isomorphism that preserves norm.

$$T : X \to Y \text{ such that } \forall x \in X,\ \|Tx\| = \|x\|$$

14.2.9 Exercise §2.8

Definitions 14.2.19 (Null space of a set)**.** Let X be a normed space and X^* be its algebraic dual space. Let $M^* \subset X^*$. Then null space of M^*, $\mathcal{N}(M^*)$ is the set of all $x \in X$ such that $f(x) = 0,\ \forall f \in M^*$.

$$\mathcal{N}(M^*) = \{x \in X : f(x) = 0,\ \forall f \in M^*\}$$

Definitions 14.2.20 (quotient space)**.** Let Y be a subspace of a vector space X. The quotient space X/Y is set $\{x + Y : x \in X\}$ together with vector addition $(x_1 + Y) + (x_2 + Y) = (x_1 + x_2) + Y$ and $\alpha(x + Y) = (\alpha x + Y)$ where $x + Y = \{x + y \in X : y \in Y\}$.

Definitions 14.2.21 (codimension)**.** Let Y be a subspace of vector space X. The codimension of Y is the dimension of the quotient space X/Y.

Definitions 14.2.22 (hyperplane)**.** Let Y be a subspace of vector space X with codimension 1. Then the quotient space is a hyperplane parallel to Y.

Remark ([Kreyszig, 2014, §2.8 Exercise 12]). Let $f \in X^*$ and $f \neq 0$. Then, constant functionals on X forms a hyperplane parallel to the null space of f.

Definitions 14.2.23 (half space)**.** Let $f \in X^*$ and $f \neq 0$ such that $f(x) = c,\ \forall x \in X$. Then the hyperplane divides the vector space X into two half spaces X_1, X_2 such that

$$X_1 = \{x \in X : f(x) \leq c\} \quad \text{and} \quad X_2 = \{x \in X : f(x) \geq c\}$$

14.2.10 Linear Operators and Functionals on finite dimensional spaces

Matrix of Linear Operator

Let X, Y be finite dimensional vector spaces over the same field K. Let $E = \{e_1, e_2, \ldots, e_n\}$ be a basis for X and $B = \{b_1, b_2, \ldots, b_r\}$ be a basis for Y. Let $T : X \to Y$ be a linear operator uniquely defined by $y_k = Te_k$. Then, $y_k = \sum \tau_{j,k} b_k = Te_k$. And the matrix $T_{EB} = (\tau_{j,k})$ is the matrix of the linear operator T with respect to the bases E and B.

For example[11], Let $T : \mathbb{R}^3 \to \mathbb{R}^3$ defined by $(\xi_1, \xi_2, \xi_3) \to (\xi_1, \xi_2, -\xi_1 - \xi_2)$.

$$
\begin{aligned}
y_1 &= Te_1 = T(1,0,0) = (1, 0-1) = 1(1,0,0) + 0(0,1,0) - 1(0,0,1) \\
y_2 &= Te_2 = T(0,1,0) = (0,1,-1) = 0(1,0,0) + 1(0,1,0) - 1(0,0,1) \\
y_3 &= Te_3 = T(0,0,1) = (0,0,0) = 0(1,0,0) + 0(0,1,0) + 0(0,0,1)
\end{aligned}
$$

Thus,

$$
T_{E,B} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{pmatrix}
$$

Linear operator T determines a unique matrix representing T with respect to the given bases. Similarly, any $r \times n$ matrix determines a unique linear operator which it represents.

Definitions 14.2.24 (Dual basis)**.** Let X be a finite dimensional vector space over K with basis $E = \{e_1, e_2, \ldots, e_n\}$. Let $f_1, f_2, \ldots, f_n$ be linear functionals $f_j : X \to K$ such that $f_j(e_k) = \delta_{j,k}$. Then $\{f_1, f_2, \ldots, f_n\}$ is the dual basis of the basis E.

Theorem 14.2.26. *Let X be a finite dimensional vector space. Let $E = \{e_1, e_2, \ldots, e_n\}$ be a basis for X and $F = \{f_1, f_2, \ldots, f_n\}$ be the dual basis of E. Then F is a basis for the algebraic dual of X, X^*. And $\dim X^* = \dim X$.*

Proof. Let $\sum \beta_k f_k = 0$. Then, for each $j = 1, 2, \ldots, n$

$$
0 = 0(e_j) = \left(\sum_{k=1}^{n} \beta_k f_k \right)(e_j) = \sum_{k=1}^{n} \beta_k f_k(e_j) = \sum_{k=1}^{n} \beta_k \delta_{j,k} = \beta_j
$$

Clearly, $\beta_1 f_1 + \beta_2 f_2 + \cdots + \beta_n f_n = 0$ if and only if $\beta_j = 0$ for all j. Therefore, F is linearly independent.

Let $f : X \to K$ such that $f(e_j) = \alpha_j$. Let $x = \sum \xi_j e_j$ and $f_j(x) = \xi_j$. Since f is linear,

$$
f(x) = \sum_{j=1}^{n} \xi_j f(e_j) = \sum_{j=1}^{n} \xi_j \alpha_j = \sum_{j=1}^{n} \alpha_j f_j(x) = \left(\sum_{j=1}^{n} \alpha_j f_j \right)(x)
$$

Therefore, every linear functional on X is a linear combination of F.
Clearly, F is a basis for X^* and $\dim X^* = n = \dim X$ □

[11] [Kreyszig, 2014, §2.10 Exercise 2]

Lemma 14.2.27. *Let X be a finite dimensional vector space. If $x_0 \in X$ such that $f(x_0) = 0$ for any $f \in X^*$, then $x_0 = 0$.*

Proof. Let $x_0 = \sum \xi_j e_j$. Then,

$$f(x_0) = 0 = \sum_{j=1}^{n} \xi_j f(e_j) = \sum_{j=1}^{n} \xi_j \alpha_j$$

Thus, $\sum \xi_j \alpha_j = 0$ for any choice of $\alpha_1, \alpha_2, \ldots, \alpha_n \in K$. Therefore, for each j, we have $\xi_j = 0$. Clearly, $x_0 = 0$. □

Theorem 14.2.28 (algebraic reflexive)**.** *A finite dimensional vector space is algebraically reflexive.*

Proof. Let $C : X \to X^{**}$ be the cannonical mapping such that

$$Cx(f) = g_x(f) = f(x)$$

where $x \in X$, $f \in X^*$ and $g \in X^{**}$.

Isomorphism C is injective, since $C_{x_0} = 0$ only if $x_0 = 0$.

$$\begin{aligned} &C_{x_0}(f) = g_{x_0}(f) = f(x_0) = 0, \ \forall f \in X^* \\ &\implies x_0 = 0 \text{ by lemma} \end{aligned}$$

Thus, there exists an inverse function $C^{-1} : \mathscr{R}(C) \to X$ and $\dim X \leq \dim \mathscr{R}(C)$. Thus, $\dim X^{**} = \dim X^* = \dim X$ and $\dim \mathscr{R}(C) = \dim X^{**}$. Therefore, C is surjective and the vector space X is algebraically reflexive. □

14.2.11 Normed spaces of Operators, Dual space

Definitions 14.2.25 ($B(X,Y)$)**.** Let X, Y be normed spaces. The set $B(X,Y)$ of all bounded, linear functionals from X into Y is a vector space together with vector addition, $T_1 + T_2$ defined by,

$$T_1 + T_2(x) = T_1(x) + T_2(x)$$

and scalar multiplication αT defined by,

$$(\alpha T)(x) = \alpha T(x)$$

In addition, $B(X,Y)$ is a normed space with norm

$$\|T\| = \sup_{\substack{x \in X \\ x \neq 0}} \frac{\|Tx\|}{\|x\|} = \sup_{\substack{x \in X \\ \|x\|=1}} \|Tx\|$$

Proof. Clearly, vector addition is closed, associative, and commutative. The identity vector 0 is defined by $0x = 0$, $\forall x \in X$. And additive inverse $(-T)$ is defined by $(-T)(x) = -T(x)$ for each $T \in B(X,Y)$. Also we have, $1T = T$, $\alpha(\beta T) = \alpha\beta T$, $\alpha(T_1 + T_2) = \alpha T_1 + \alpha T_2$ and $(\alpha_1 + \alpha_2)T = \alpha_1 T + \alpha_2 T$. Thus, $B(X,Y)$ is a vector space. □

Theorem 14.2.29. *If Y is a Banach space, then $B(X,Y)$ is a Banach space.*

Proof. Let $\{T_n\}$ be a Cauchy sequence in $B(X,Y)$. That is, $\forall \varepsilon > 0, \exists N \in \mathbb{N}$ such that $\forall n, m > N$ we have $\|T_n - T_m\| \leq \varepsilon$. Thus, for any $x \in X$ we have

$$\|T_n(x) - T_m(x)\| \leq \|T_n - T_m\| \; \|x\| \leq \varepsilon \|x\|$$

Fix $x \in X$ and $\varepsilon > 0$, then we have $\varepsilon_x > 0$ such that $\varepsilon_x \|x\| < \varepsilon$. Then, there exists $N \in \mathbb{N}$ such that $\forall n, m > N$, $\|T_n(x) - T_m(x)\| < \varepsilon$. Thus, $\{T_n(x)\}$ is Cauchy in Y. Since, Y is complete, $\{T_n(x)\}$ converges, say $T_n(x) \to y$.

Define $T : X \to Y$ such that $Tx = y$ where $T_n(x) \to y$ as $n \to \infty$. Operator T is linear since

$$T(\alpha x + \beta z) = \lim_{n\to\infty} T_n(\alpha x + \beta z) = \lim_{n\to\infty} \alpha T_n(x) + \lim_{n\to\infty} \beta T_n(z) = \alpha T(x) + \beta T(z)$$

We have,

$$\|T_n(x) - T(x)\| = \left\| T_n(x) - \lim_{m\to\infty} T_m(x) \right\| \leq \|T_n - T_m\| \; \|x\| \leq \varepsilon \|x\|$$

Thus, $T_n - T$ is bounded and

$$\sup_{\substack{x\in X \\ \|x\|=1}} \|(T_n - T)x\| = \sup_{\substack{x\in X \\ \|x\|=1}} \|T_n(x) - T(x)\| \leq \varepsilon$$

Thus, $\|T_n - T\| \to 0$ as $n \to \infty$. We have, $T = T_n - (T_n - T)$ and

$$\|T(x)\| \leq \|T_n(x)\| + \|(T_n - T)(x)\| = \varepsilon \|x\| + \varepsilon \|x\|$$

Therefore, T is bounded. □

Definitions 14.2.26 (Dual space X')**.** Let X be a normed space. The set of X' of all bounded, linear functionals $B(X,K)$ is the dual space of X. The dual space X' of X is a normed space with norm

$$\|f\| = \sup_{\substack{x\in X \\ x\neq 0}} \frac{|f(x)|}{\|x\|} = \sup_{\substack{x\in X \\ \|x\|=1}} |f(x)|$$

Theorem 14.2.30. *The dual space X' of any normed space X is a Banach space.*

Proof. We know that, $B(X,Y)$ is Banach space if Y is a Banach space. We have $K = \mathbb{R}$ or $K = \mathbb{C}$. And $\mathbb{R}$ as well as $\mathbb{C}$ are complete normed spaces. Therefore, $X' = B(X,K)$ is a Banach space. □

14.2.12 Dual Spaces

Remark. Dual space of $\mathbb{R}^n$ is $\mathbb{R}^n$

Proof. Let $f \in (\mathbb{R}^n)'$. Let $E = \{e_1, e_2, \ldots, e_n\}$ be a basis for $\mathbb{R}^n$. Then f is of the form, $f = \sum \xi_j \alpha_j$ where $\alpha_j = f(e_j)$.

By Cauchy-Schwarz's inequality, we have

$$|f(x)| \leq \sum |\xi_j \alpha_j| \leq \left(\sum \xi_j^2\right)^{\frac{1}{2}} \left(\sum \alpha_j^2\right)^{\frac{1}{2}} = \|x\| \left(\sum \alpha_j^2\right)^{\frac{1}{2}}$$

Thus,

$$\|f\| = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|=1}} |f(x)| \leq \left(\sum \alpha_j^2\right)^{\frac{1}{2}}$$

Also for $x = (\alpha_j)$ we get,

$$\|f\| = \left(\sum \alpha_j^2\right)^{\frac{1}{2}}$$

And for any $c = (\alpha_1, \alpha_2, \ldots, \alpha_n) \in \mathbb{R}^n$, we have a linear functional f such that $f(x) = \sum \xi_k \alpha_k$ where $x = (\xi_1, \xi_2, \ldots, \xi_n)$. We know that, for finite dimensional normed space X, every linear functional on X is bounded. Therefore, f is bounded and $f \in \mathbb{R}^{n\prime}$. Thus, the map from the dual space of $\mathbb{R}^n$ into $\mathbb{R}^n$ given by $f \to c$ is an isomorphism preserving the Euclidean norm on $\mathbb{R}^n$. Therefore, the dual space of $\mathbb{R}^n$ is $\mathbb{R}^n$ itself. □

Remark. Dual space of l^1 is l^∞

Proof. The Schauder basis for l^1 is (e_k) where $e_k = (\delta_{k,j})$. That is, each sequence e_k has one as its kth term and all other terms are zeroes. For example, $(\frac{1}{2}, \frac{1}{4}, \ldots) = \sum \frac{1}{2^k} e_k$. Then any $x \in l^1$ is of the form

$$x = (\xi_j) = \sum_{k=1}^{\infty} \xi_k \delta_{k,j} = \sum_{k=1}^{\infty} \xi_k e_k$$

Let $f \in l^{1\prime}$. Then,

$$f(x) = f\left(\sum \xi_k e_k\right) = \sum \xi_k f(e_k) = \sum \xi_k \gamma_k$$

where $\gamma_k = f(e_k)$. Since f is bounded, we have $|f(x)| \leq \|f\| \, \|x\|$. Also we have, $\|e_k\| = 1$ since $e_k = (\delta_{k,j})$. Thus,

$$|\gamma_k| = |f(e_k)| \leq \|f\| \, \|e_k\| = \|f\|$$

We have, $|\gamma_k| \leq \|f\|$ for any k. Therefore,

$$\sup_k |\gamma_k| \leq \|f\|$$

For any $b = (\beta_k) \in l^\infty$, there exists a functional $g \in l^{1\prime}$ such that

$$g(x) = \sum \xi_k \beta_k$$

where $x = (\xi_k) \in l^1$. Clearly, g is linear. And,

$$|g(x)| \leq \sum |\xi_k \beta_k| \leq \sup_j |\beta_j| \sum |\xi_j| = \|x\| \sup_j |\beta_j|$$

Thus, g is bounded and $g \in l^{1'}$. And,

$$|f(x)| = \left|\sum \xi_k \gamma_k\right| \leq \sup_j |\gamma_j| \sum |\xi_k| = \|x\| \ \sup_j |\gamma_j|$$

Thus,

$$\|f\| = \sup_{\substack{x \in l^1 \\ \|x\|=1}} |f(x)| \leq \sup_j |\gamma_j|$$

Therefore, $\|f\| = \sup_j |\gamma_j|$.

Clearly, the mapping from the dual space of l^1 into l^∞ given by $f \to c$ where $c = \sum \gamma_j$ is an isomorphism preserving the norm on l^∞. Thus, the dual space of l^1 is isomorphic to l^∞. $\square$

Remark. Dual space of l^p is l^q where p, q are conjugate exponents.

Proof. Sequence (e_k) is a Schauder basis for l^p where $e_k = \delta_{k,j}$. And every $x \in l^p$ is of the form

$$x = (\xi_k) = \sum_{k=1}^{\infty} \xi_k e_k$$

Let $f \in l^{p'}$. Then f is linear and

$$f(x) = f\left(\sum_{k=1}^{\infty} \xi_k e_k\right) = \sum_{k=1}^{\infty} \xi_k f(e_k) = \sum_{k=1}^{\infty} \xi_k \gamma_k$$

where $\gamma_k = f(e_k)$.

Define sequence $\{x_n\}$ in l^p such that $x_n = (\xi_{k,n})$ where

$$\xi_{k,n} = \begin{cases} \frac{|\gamma_k|^q}{\gamma_k} & k \leq n \text{ and } \gamma_k \neq 0 \\ 0 & k > n \text{ or } \gamma_k = 0 \end{cases}$$

Then,

$$f(x_n) = \sum \xi_{k,n} \gamma_k = \sum |\gamma_k|^q$$

Let p, q be conjugate exponents. Then $pq = p + q$. And $|\xi_{k,n}| = |\gamma_k|^{(q-1)}$. Thus,

$$\begin{aligned} |f(x_n)| &\leq \|f\| \ \|x_n\| \\ &= \|f\| \ \left(\sum |\xi_{k,n}|^p\right)^{\frac{1}{p}} \\ &= \|f\| \ \left(\sum |\gamma_k|^{(q-1)p}\right)^{\frac{1}{p}} \\ &= \|f\| \ \left(\sum |\gamma_k|^q\right)^{\frac{1}{p}} \ \text{ since } pq = p + q \end{aligned}$$

Therefore,

$$|f(x_n)| = \sum |\gamma_k|^q \leq \|f\| \ \left(\sum |\gamma_k|^q\right)^{\frac{1}{p}}$$

$$\implies \left(\sum |\gamma_k|^q\right)^{\left(1-\frac{1}{p}\right)} = \left(\sum |\gamma_k|^q\right)^{\frac{1}{q}} \leq \|f\|$$

Let $b = (\beta_k) \in l^q$. And there exists a functional g on l^p such that

$$g(x) = \sum \xi_k \beta_k$$

where $x = (\xi_k) \in l^p$. Clearly, g is linear. And,

$$|g(x)| \leq \left|\sum \xi_k \beta_k\right| \leq \left(\sum |\xi_k|^p\right)^{\frac{1}{p}} \left(\sum |\beta_k|^q\right)^{\frac{1}{q}} = \|x\| \left(\sum |\beta_k|^q\right)^{\frac{1}{q}}$$

Therefore, g is a bounded, linear functional on l^p. That is, $g \in l^{p\prime}$.

And we have,

$$\begin{aligned} |f(x)| &\leq \left|\sum \xi_k \gamma_k\right| \\ &\leq \left(\sum |\xi_k|^p\right)^{\frac{1}{p}} \left(\sum |\gamma_k|^q\right)^{\frac{1}{q}} \\ &\leq \|x\| \left(\sum |\gamma_k|^q\right)^{\frac{1}{q}} \\ \|f\| &\leq \sup_{\substack{x \in l^p \\ \|x\|=1}} \left(\sum |\gamma_k|^q\right)^{\frac{1}{q}} \end{aligned}$$

Then,

$$\|f\| = \left(\sum \gamma_k|^q\right)^{\frac{1}{q}}$$

Clearly, the mapping from the dual space of l^p into l^q given by $f \to c$ is an isomorphism preserving norm. Therefore, the dual space of l^p is isomorphic to l^q where p, q are conjugate exponents. That is, $\frac{1}{p} + \frac{1}{q} = 1$. $\square$

Module 3 - Hilbert Spaces

14.3 Hilbert Spaces

14.3.1 Inner Product space, Hilbert space

Definitions 14.3.1 (inner product)**.** Let X be a vector space over the field K. Inner product is a function $\langle \cdot, \cdot \rangle : X \times X \to K$ which satisfies

1. $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
2. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$
3. $\langle x, y \rangle = \overline{\langle y, x \rangle}$
4. $\langle x, x \rangle \geq 0$ and
$\langle x, x \rangle = 0 \iff x = 0$

Definitions 14.3.2. A vector space X on field K together with an inner product function $\langle \cdot, \cdot \rangle : X \times X \to K$ is an inner product space.

Definitions 14.3.3 (Hilbert)**.** Hilbert spaces are complete, inner product spaces.

Remark. Inner product spaces are normed spaces with norm

$$\|x\| = \sqrt{\langle x, x \rangle}$$

Definitions 14.3.4 (sesquilinear)**.** An operator $T : X \to Y$ is sesquilinear if

$$T(\alpha x + \beta y) = \alpha T(x) + \overline{\beta} T(y)$$

Inner product is a sesquilinear operator.

Remark. Inner product spaces are metric spaces with metric

$$d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$$

Remark. Let X be a real, inner product space. Then,

$$\langle x, y \rangle = \langle y, x \rangle$$

Remark. Let X be a complex, inner product space. Then, $\langle x, y \rangle = \overline{\langle y, x \rangle}$.

Proof. Let $x = a + ib$ and $y = c + id$ where $a, b, c, d \in \mathbb{R}$. Then,

$$\langle x, y \rangle = \langle a + ib, c + id \rangle = \langle a, c \rangle - i \langle a, d \rangle + i \langle b, c \rangle + \langle b, d \rangle$$

$$\begin{aligned}
\langle y, x \rangle &= \langle c + id, a + ib \rangle \\
&= \langle c, a \rangle - i \langle c, b \rangle + i \langle d, a \rangle + \langle d, b \rangle \\
&= \langle a, c \rangle - i \langle b, c \rangle + i \langle a, d \rangle + \langle b, d \rangle \\
&= \overline{\langle x, y \rangle}
\end{aligned}$$

Therefore,

$$\langle x, y \rangle = \overline{\langle y, x \rangle}$$

□

Remark. From the axiom of inner product, we have

1. $\langle \alpha x + \beta y, z\rangle = \alpha \langle x, z\rangle + \beta \langle y, z\rangle$
2. $\langle x, \alpha y\rangle = \overline{\alpha} \langle x, y\rangle$
3. $\langle x, \alpha y + \beta z\rangle = \overline{\alpha} \langle x, y\rangle + \overline{\beta} \langle x, z\rangle$

Lemma 14.3.1 (parallelogram equality)**.** *Let X be an inner product space. Let $\|\cdot\|$ be a the norm induced by the inner product. That is, $\|x\| = \sqrt{\langle x, x\rangle}$. Then,*

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$$

Proof. Let $\|\cdot\|$ be a normed induced by inner product. Then $\|x\|^2 = \langle x, x\rangle$ and $\|y\|^2 = \langle y, y\rangle$.

$$\begin{aligned}
\|x + y\|^2 + \|x - y\|^2 &= \langle x + y, x + y\rangle + \langle x - y, x - y\rangle \\
&= \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle \\
&\quad + \langle x, x\rangle + \langle x, -y\rangle + \langle -y, x\rangle + \langle -y, -y\rangle \\
&= \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle \\
&\quad + \langle x, x\rangle - \langle x, y\rangle - \langle y, x\rangle + \langle y, y\rangle \\
&= 2\langle x, x\rangle + 2\langle y, y\rangle = 2(\|x\|^2 + \|y\|^2)
\end{aligned}$$

□

Not all normed spaces are inner product spaces.
For example, $C[a, b]$, l^p are not inner product spaces.

Definitions 14.3.5 (orthogonality)**.** Let X be an inner product space and $x, y \in X$. Then x, y are orthogonal if $\langle x, y\rangle = 0$, say $x \perp y$.

Let A, B be subsets of X. Then A and B are orthogonal if $\forall a \in A$, $\forall b \in B$, $\langle a, b\rangle = 0$.

$$A \perp B \iff \forall a \in A,\ \forall b \in B,\ a \perp b = 0 \iff \forall a \in A,\ \forall b \in B,\ \langle a, b\rangle = 0$$

Examples

Euclidean Space

Banach space $\mathbb{R}^n$ is a Hilbert space with inner product,

$$\langle x, y\rangle = \sum_{j=1}^{n} \xi_j \eta_j$$

Unitary Space

Banach space $\mathbb{C}^n$ is a Hilbert space with inner product,

$$\langle x, y\rangle = \sum_{j=1}^{n} \xi_j \overline{\eta_j}$$

L^2 Space

Banach space of Lebesgue square integrable functions is a Hilbert space with inner product,

$$\langle x, y\rangle = \left(\int_a^b x(t)\ \overline{y(t)}\ dt\right)^{\frac{1}{2}}$$

L^2 space is the completion of the incomplete normed space of all continuous complex-valued functions on $[a, b]$ with norm $\|x\| = \sqrt{\langle x, x\rangle}$.

Remember that, the space of all continuous functions on $[a, b]$ is incomplete since the Cauchy sequence of continuous functions $x_n : [a, b] \to [0, 1]$ defined by

$$x_n(t) = \begin{cases} 0 & a \le t \le \frac{b+a}{2} \\ \frac{2t-b-a}{2n} & \frac{b+a}{2} < t < \frac{b+a}{2} + \frac{1}{n} \\ 1 & \frac{b+a}{2} + \frac{1}{n} \le t \le b \end{cases}$$

doesn't converge to continuous function.

Also note that, the above sequence is not a Cauchy sequence in $C[a, b]$ with norm $\|x\| = \max |x(t)|$ since

$$\|x_n(t) - x_m(t)\| = 1 - \frac{1}{mn}$$

l^2 Space

Banach space l^2 is a Hilbert space with inner product,

$$\langle x, y\rangle = \sum_{j=1}^{\infty} \xi_j \overline{\eta_j}$$

Counter-examples

Remark. Banach space $C[a, b]$ with norm $\|x\| = \max |x(t)|$ is not an inner product space.

Proof. We have $C[a, b]$, the space of all continuous real-valued function on compact interval $[a, b]$ is a normed space with

$$\|x\| = \max_{t \in [a,b]} |x(t)|$$

Let $x(t) = 1$ and $y(t) = \frac{t-a}{b-a}$. Then $x + y(t) = 1 + \frac{t-a}{b-a}$ and $x - y(t) = \frac{b-t}{b-a}$. And $\|x\| = 1$, $\|y\| = 1$, $\|x + y\| = 2$†[12] and $\|x - y\| = 1$. Clearly, $\|x + y\|^2 + \|x - y\|^2 = 5 \neq 4 = 2(\|x\|^2 + \|y\|^2)$. Therefore, $\|\cdot\|$ can't be induced from any inner product on $C[a, b]$. Thus, $C[a, b]$ (with above norm) is not an inner product space. □

[12] Since $x+y(b) = 2$ is the maximum value of $x+y$ in $[a, b]$, we have $\|x+y\| = \max |x+y| = 2$.

Remark. Banach space l^p is not an inner product space, if $p \neq 2$.

Proof. Suppose $p \neq 2$. We have, $e_j = \delta_{j,k}$ is a Schauder basis for l^p space. Let $x = e_1 + e_2$ and $y = e_1 - e_2$ where $e_j = \delta_{j,k}$. Then $x + y = 2e_1$ and $x - y = 2e_2$. And $\|x\| = 2^{\frac{1}{p}}$, $\|y\| = 2^{\frac{1}{p}}$, $\|x+y\| = 2$ and $\|x - y\| = 2$. Then,

$$\|x+y\|^2 + \|x-y\|^2 = 8 \neq 4 \times 2^{\frac{2}{p}} = 2(\|x\|^2 + \|y\|^2), \qquad \because p \neq 2$$

Thus, the norm on l^p can't be induced from an inner product when $p \neq 2$. Therefore, l^p is not an inner product space, if $p \neq 2$. □

Remark (Hilbert). The sequence space l^2 was introduced by David Hilbert and is known as **Hilbert's sequence space**. It is a separable, complete, inner product space. Later, the condition of separability was dropped by mathematicians as most of the theory didn't make use of it.

Remark. Let X be a real inner product space. Then,

$$\langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$$

Theorem 14.3.2 (Polarization identity)**.** *Let X be a complex inner product space. Then, the real and imaginary parts of $\langle x, y\rangle$ are*

$$\Re(\langle x, y\rangle) = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$$
$$\Im(\langle x, y\rangle) = \frac{1}{4}\left(\|x+iy\|^2 - \|x-iy\|^2\right)$$

Proof. Let X be a comples, inner product space. Let $x, y \in X$. Then,

$$\begin{aligned}\|x+y\|^2 - \|x-y\|^2 &= \langle x+y, x+y\rangle - \langle x-y, x-y\rangle \\ &= \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle \\ &\quad - \langle x, x\rangle - \langle x, -y\rangle - \langle -y, x\rangle - \langle -y, -y\rangle \\ &= \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle \\ &\quad - \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle - \langle y, y\rangle \\ &= 2(\langle x, y\rangle + \langle y, x\rangle)\end{aligned}$$

We know that, $\langle x, y\rangle = \overline{\langle y, x\rangle}$. Thus,

$$\|x+y\|^2 - \|x-y\|^2 = 2(\langle x, y\rangle + \overline{\langle x, y\rangle}) = 4\Re\langle x, y\rangle$$

Therefore,

$$\Re\langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$$

Let $z = iy$.

$$\begin{aligned}\|x+iy\|^2 - \|x-iy\|^2 &= \|x+z\|^2 - \|x-z\|^2 \\ &= 2\langle x, z\rangle + 2\langle z, x\rangle \\ &= 2\langle x, iy\rangle + 2\langle iy, x\rangle) \\ &= -2i\langle x, y\rangle + 2i\langle y, x\rangle \\ &= -2i(\langle x, y\rangle - \langle y, x\rangle) \\ &= -2i(\langle x, y\rangle - \overline{\langle x, y\rangle}) \\ &= 4\Im\langle x, y\rangle\end{aligned}$$

$$\Im \langle x, y \rangle = \frac{1}{4}\left(\|x + iy\|^2 - \|x - iy\|^2\right)$$

□

Exercise §3.1

Theorem 14.3.3 (Pythagorean Identity)**.** *Let X be an inner product space. Let $x, y \in X$. If $x \perp y$, then*

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2$$

Proof. Suppose $x \perp y$. Then $\langle x, y \rangle = \langle y, x \rangle = 0$. And,

$$\begin{aligned}\|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \\ &= \|x\|^2 + \|y\|^2\end{aligned}$$

□

Remark.

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 \not\Rightarrow x \perp y$$

Proof.

$$\begin{aligned}\|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \\ &= \|x\|^2 + \|y\|^2 + \langle x, y \rangle + \langle y, x \rangle\end{aligned}$$

Thus, if $\|x + y\|^2 = \|x\|^2 + \|y\|^2$, then $\langle x, y \rangle + \langle y, x \rangle = 0$.

We know that $\langle x, y \rangle = \overline{\langle y, x \rangle}$. And, $\langle x, y \rangle + \overline{\langle x, y \rangle} = 0$ only if $\langle x, y \rangle$ is purely imaginary. Thus, the Pythagoran theorem is true if $\langle x, y \rangle$ is purely imaginary. □

Theorem 14.3.4 (Apollonius Identity)**.** *Let X be an inner product space. Let $x, y, z \in X$. Then,*

$$\|z - x\|^2 + \|z - y\|^2 = \frac{1}{2}\|x - y\|^2 + 2\|z - \frac{1}{2}(x + y)\|^2$$

Proof. Let $s = z - \frac{1}{2}(x+y)$ and $t = \frac{1}{2}(x-y)$. Then $s+t = z-y$ and $s-t = z-x$.

$$\begin{aligned}\|z - y\|^2 + \|z - x\|^2 &= \|s + t\|^2 + \|s - t\|^2 \\ &= 2(\|s\|^2 + \|t\|^2) \\ &= 2\left(\left\|z - \frac{1}{2}(x + y)\right\|^2 + \left\|\frac{1}{2}(x + y)\right\|^2\right) \\ &= 2\left\|z - \frac{1}{2}(x + y)\right\|^2 + \frac{1}{2}\|x + y\|^2\end{aligned}$$

□

14.3.2 Further properties of inner product spaces

Remark.

$$\|\alpha x\|^2 = \langle \alpha x, \alpha x\rangle = \alpha\overline{\alpha}\,\langle x, x\rangle = |\alpha|^2\|x\|^2$$

Lemma 14.3.5 (Schwarz Inequality)**.** *Let X be an inner product space. Let $x, y \in X$.*

$$|\,\langle x, y\rangle\,| \leq \|x\|\ \|y\|$$

And equality holds if $\{x, y\}$ are linearly independent. Also,†[13]

$$\|x + y\| \leq \|x\| + \|y\|$$

And equality holds if $y = 0$ or $x = cy$.

Proof. If $y = 0$, then $\langle x, y\rangle = 0$ and $\|y\| = 0$. And $|\,\langle x, y\rangle\,| = \|x\|\ \|y\| = 0$. Similarly, if $x = 0$, then $\|\,\langle x, y\rangle\,\| = \|x\|\ \|y\| = 0$.

Suppose $x \neq 0$ and $y \neq 0$. Then

$$\begin{aligned}\|x - \alpha y\|^2 &= \langle x - \alpha y, x - \alpha y\rangle \\ &= \langle x, x\rangle - \overline{\alpha}\,\langle x, y\rangle - \alpha\,\langle y, x\rangle + \alpha\overline{\alpha}\,\langle y, y\rangle\end{aligned}$$

Let $\overline{\alpha} = \frac{\langle y,x\rangle}{\langle y,y\rangle}$. Then

$$\begin{aligned}\|x - \alpha y\|^2 &= \langle x, x\rangle - \frac{\langle y, x\rangle\,\langle x, y\rangle}{\langle y, y\rangle} \\ &= \|x\|^2 - \frac{\|\,\langle x, y\rangle\,\|^2}{\|y\|^2}\end{aligned}$$

We have

$$0 \leq \|x - \alpha y\|^2 \leq \|x\|^2 - \frac{\|\,\langle x, y\rangle\,\|^2}{\|y\|^2}$$

Thus,

$$\|\,\langle x, y\rangle\,\| \leq \|x\|\ \|y\|$$

Clearly, equality holds if and only if $\|x - \alpha y\|^2 = 0$ or $x = 0$ or $y = 0$. In other words, equality holds if and only if $x = \alpha y$ or $x = 0$ or $y = 0$. Therefore, equality holds if and only if x, y are linearly independent.

We have,

$$\begin{aligned}\|x + y\|^2 &= \langle x + y, x + y\rangle \\ &= \|x\|^2 + \langle x, y\rangle + \langle y, x\rangle + \|y\|^2\end{aligned}$$

Applying modulus on either sides, we get

$$\|x + y\|^2 \leq \|x\|^2 + |\,\langle x, y\rangle\,| + |\,\langle y, x\rangle\,| + \|y\|^2$$

[13] Every inner product space is also a normed space. Then the induced norm should possess the triangular inequality property. I think it is too late to prove this !

By Schwarz inequality, we have

$$\|x+y\|^2 \leq \|x\|^2 + 2\|x\|\ \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2$$

Therefore,

$$\|x+y\| \leq \|x\| + \|y\|$$

And equality holds if and only if

$$\langle x, y\rangle + \langle y, x\rangle = 2\|x\|\ \|y\|$$

That is,

$$\Re\langle x, y\rangle = \|x\|\ \|y\| \geq |\langle x, y\rangle|$$

We know that the real part of a complex number can never exceed its absolute value. Therefore, $\|x\|\ \|y\| = |\langle x, y\rangle|$.

From Schwarz inequality, the equality holds if and only if x, y are linearly independent. Therefore, $\|x+y\| = \|x\| + \|y\|$ if and only if x, y are linearly independent. □

Lemma 14.3.6 (Continuity)**.** *Let X be an inner product space. Let $\{x_n\}$, $\{y_n\}$ be sequences in X. Suppose $x_n \to x$ and $y_n \to y$. Then $\langle x_n, y_n\rangle \to \langle x, y\rangle$.*

In other words, Inner product is continuous.

Proof. We have $x_n \to x$ and $y_n \to y$ as $n \to \infty$. Then, $\|x_n - x\| \to 0$ and $\|y_n - y\| \to 0$ as $n \to \infty$.

$$\begin{aligned} |\langle x_n, y_n\rangle - \langle x, y\rangle| &= |\langle x_n, y_n\rangle - \langle x_n, y\rangle + \langle x_n, y\rangle - \langle x, y\rangle| \\ &\leq |\langle x_n, y_n - y\rangle| + |\langle x_n - x, y\rangle| \end{aligned}$$

By Schwarz inequality

$$\leq \|x_n\|\ \|y_n - y\| + \|x_n - x\|\ \|y\|$$

Therefore, $|\langle x_n, y_n\rangle - \langle x, y\rangle| \to 0$ as $n \to \infty$. In other words, $\langle x_n, y_n\rangle \to \langle x, y\rangle$ as $n \to \infty$. □

Definitions 14.3.6 (isomorphism)**.** Let X, Z be inner product spaces. If there exists a bijection $T : X \to Z$ such that $\langle Tx, Ty\rangle = \langle x, y\rangle$ for any $x, y \in X$. Then T is an isomorphism preserving inner product.

Theorem 14.3.7. *Every inner product space X has a unique completion H in which it is dense.*

Proof. Let X be an inner product space. Then X is a normed space with inner product induced norm $\|x\| = \sqrt{\langle x, x\rangle}$. By completion theorem for normed spaces, there exists a unique†[14] Banach space H which contains a dense subset W which is isomorphic to X.

[14]Unique upto isomorphism - If there exists two Banach spaces H_1, H_2 which contain a dense subset which is isomorphic to X, then H_1, H_2 are isomorphic normed spaces.

We have, W is dense in H and H is the set of all equivalent classes of Cauchy sequences on X. And $\hat{x}_n \sim \hat{w}$ if and only if $\|x_n - w\| \to 0$ as $n \to \infty$ for some $w \in W$.

Defining an inner product on H,

$$\langle \hat{x}, \hat{y} \rangle = \lim_{n\to\infty} \langle x_n, y_n \rangle$$

By polarisation identity,

$$\begin{aligned}\langle x_n, y_n \rangle &= \frac{1}{4}\left(\|x_n + y_n\|^2 - \|x_n - y_n\|^2\right) + \frac{i}{4}\left(\|x_n + iy_n\|^2 - \|x_n - iy_n\|^2\right) \\ &= \frac{1}{4}\left(\langle x_n + y_n, x_n + y_n\rangle - \langle x_n - y_n, x_n - y_n\rangle\right) \\ &\quad + \frac{i}{4}\left(\langle x_n + iy_n, x_n + iy_n\rangle - \langle x_n - iy_n, x_n - iy_n\rangle\right)\end{aligned}$$

And,

$$\begin{aligned}\langle Tx, Ty \rangle &= \langle \hat{x}, \hat{y} \rangle \\ &= \lim_{n\to\infty} \langle x_n, y_n \rangle \\ &= \frac{1}{4}\left(\langle x + y, x + y\rangle - \langle x - y, x - y\rangle\right) \\ &\quad + \frac{i}{4}\left(\langle x + iy, x + iy\rangle - \langle x - iy, x - iy\rangle\right) \\ &= \langle x, y \rangle\end{aligned}$$

Thus, $T : X \to W$ is an isomorphism preserving norms.

Let H_1, H_2 be two completions of X. Then, we have normed space isomorphism $T : H_1 \to H_2$ such that $\|Tx\| = \|x\|$. Clearly, $\langle Tx, Ty \rangle = \langle x, y \rangle$. Therefore, completion of X is unique upto an inner product isomorphism. $\square$

Theorem 14.3.8. *Let Y be a subspace of a Hilbert space. Then,*

1. *Y is complete if and only if Y is closed in H.*
2. *If Y is finite dimensional, then Y is complete.*
3. *If H is separable, then Y is separable.*

Proof. Let H is a Hilbert space. Then H is a Banach space with induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. Then, any subset Y of H is complete if and only if$Y$ is closed in H.

Let Y be a finite dimensional subspace of Hilbert space H. Then, H is a Banach space with induced norm and any subspace of H is a normed space. We know that any finite dimensional normed space is complete. Thus, Y is complete.

Let H be a separable, Hilbert space. Then H has a countable, dense subset W. Let Y be a subspace of H. Then $W \cap Y$ is countable, dense subset of Y. Thus, subspace Y is separable. $\square$

Every subset of separable inner product space is separable.

14.3.3 Orthogonal complements and direct sums

Definitions 14.3.7 (distance from a set)**.** Let X be metric space. Let $x \in X$ and $M \subset X$. The distance of x from M is defined as

$$d(x, M) = \inf_{y \in M} d(x, y)$$

If X is normed space, then

$$d(x, M) = \inf_{y \in M} \|x - y\|$$

Definitions 14.3.8 (convex)**.** Let M be a subset of a vector space X. Then M is convex if for any two points $x, y \in M$, the segment $z = \alpha x + (1 - \alpha)x$ is contained in M. That is,

$$\forall x, y \in M, \forall \alpha \in [0, 1], \ \alpha x + (1 - \alpha)y \ \in M$$

Theorem 14.3.9 (minimising vector)**.** *Let X be an inner product space and M be a nonempty, convex, complete subset of X. Then for any $x \in X$, there exists a unique $y \in M$ such that $d(x, M) = d(x, y)$.*

Proof. Let X be an inner product space with metric $d(x, y) = \sqrt{\langle x - y, x - y\rangle}$.

Step 1 : Existence of minimising vector
Let $\delta = d(x, M) = \inf_{y \in M} d(x, y) = \inf_{y \in M} \|x - y\|$. Let $\{\delta_n\}$ be a sequence converging to δ. Then, there exists $y_n \in M$ such that $\delta < d(x, y_n) < \delta_n$.

We have,

$$\|y_n - y_m\|^2 = \|y_n - x - y_m + x\|^2$$

Let $v_n = y_n - x$ and $v_m = y_m - x$.Then

$$\begin{aligned}
&= \|v_n - v_m\|^2 \\
&= 2\left(\|v_n\|^2 + \|v_m\|^2\right) - \|v_n + v_m\|^2 \\
&= 2\left(\|y_n - x\|^2 + \|y_m - x\|^2\right) - \|y_n + y_m - 2x\|^2 \\
&\leq 2(\delta_n^2 + \delta_m^2) - (2\delta)^2
\end{aligned}$$

Thus, the sequence $\{y_n\}$ is Cauchy. Since M is complete and $\{y_n\}$ is Cauchy, $y_n \to y \in M$. Therefore, there do exist $y \in M$ such that $d(x, M) = d(x, y) = \|x - y\| = \delta$.

Step 2 : Uniqueness of minimising vector
Let $y, y_0 \in M$ such that $\|x - y\| = \|x - y_0\| = \delta$.

$$\|y - y_0\|^2 = \|y - x - y_0 + x\|^2$$

Let $v = y - x$ and $w = y_0 - x$

$$\|y - y_0\|^2 = \|v - w\|^2$$

By parallelogram equality, we have

$$\begin{aligned}\|y-y_0\|^2 &= 2\|v\|^2+2\|w\|^2-\|v+w\|^2\\ &= 2\|y-x\|^2+2\|y_0-x\|^2-\|y+y_0-2x\|^2\\ &\le 2\delta^2+2\delta^2-4\delta^2=0 \qquad \text{since } d\left(x,\frac{y+y_0}{2}\right)\ge\delta\end{aligned}$$

Thus, $y=y_0$. Therefore, the (distance) minimising vector (in M) is unique for any $x\in X$. □

In Banach spaces, there may not exist such a point or there may exist many points at minimal distance !

Definitions 14.3.9 (minimising vector)**.** Let X be a Hilbert space. Let $x\in X$ and $M\subset X$ be a complete, convex subset. Then minimising vector of M is a unique vector $y\in M$ such that distance from x to y is minimal among other points in M.

Lemma 14.3.10 (Orgonality of minimising vector)**.** *Let X be a Hilbert space. Let Y be a complete, convex subspace of X. Let $x\in X$ and $y\in Y$ be its minimising vector. Then $z=x-y$ is orthogonal to Y.*

Proof. Suppose $z\perp Y$ is false.
Then there exists $y_1\in Y$ such that $\langle y_1,z\rangle=\beta\neq 0$. Let $\alpha\in K$. Then,

$$\begin{aligned}\|z-\alpha y_1\|^2 &= \langle z-\alpha y_1, z-\alpha y_1\rangle\\ &= \langle z,z\rangle-\overline{\alpha}\langle z,y_1\rangle-\alpha\langle y_1,z\rangle+\alpha\overline{\alpha}\langle y_1,y_1\rangle\end{aligned}$$

When $\alpha=\frac{\beta}{\langle y_1,y_1\rangle}=\frac{\langle z,y_1\rangle}{\langle y_1,y_1\rangle}$. Then $\overline{\alpha}\beta=\frac{\overline{\beta}\beta}{\langle y_1,y_1\rangle}=\frac{|\beta|^2}{\|y_1\|^2}$. Then,

$$\|z-\alpha y_1\|^2=\|z\|^2-\frac{|\beta|^2}{\|y_1\|^2}<\delta^2$$

This is a contradiction, since $\|z-\alpha y_1\|=\|x-y-\alpha y_2\|=\|x-y_2\|\ge\delta$ where $y_2=y+\alpha y_1\in Y$. Therefore, $z\perp Y$ is true. □

Definitions 14.3.10 (direct sum)**.** A vector space X is a direct sum of two subspaces Y,Z if each $x\in X$ has a unique representation of the form $x=y+z$ where $y\in Y$ and $z\in Z$. And we write, $X=Y\oplus Z$.

Definitions 14.3.11 (algebraic complement)**.** Let $X=Y\oplus Z$. Then, Y,Z are algebraic complements.

Definitions 14.3.12 (orthogonal complement)**.** Let $X=Y\oplus Z$ where $Z=Y^\perp=\{x\in X : x\perp Y\}$. Then Y,Z are orthogonal complements.

Theorem 14.3.11 (direct sum)**.** *Let Y be any closed subspace of a Hilbert space H. Then $H=Y\oplus Z$ where $Z=Y^\perp$.*

Proof. Let H be a Hilbert space. Let Y be a closed subspace of H. Then Y is complete. And Y is convex, since Y is a subspace.

Since Y is a complete, convex subset of an inner product space H, for each $x \in H$ there exists a unique minimising vector $y \in Y$ such that $x - y \perp Y$. Define $Z = \{x \in H : (x-y) \perp Y\}$. If $z \in Z$, then there exists $x \in H$ and $y \in Y$ such that $z = x - y \in Y^{\perp}$. Thus each $x \in H$ has a representation, $x = y + z$ where $y \in Y$ and $z = x - y \in Y^{\perp}$.

We know that $Y \cap Y^{\perp} = \{0\}$, the zero subspace of H. Suppose $x = y_1 + z_1$ and $x = y_2 + z_2$ where $y_1, y_2 \in Y$ and $z_1, z_2 \in Y^{\perp}$. Then $y_1 - y_2 = z_2 - z_1$. Since Y is a subspace, $y_1 - y_2 \in Y$ and $z_2 - z_1 \in Y^{\perp}$. Thus, $y_1 - y_2 = z_2 - z_1 \in Y \cap Y^{\perp} = \{0\}$. That is, $y_1 - y_2 = 0$. Therefore, the representation $x = y + z$ is unique. □

Definitions 14.3.13 (Projection)**.** Let H be an inner product space. Let Y be a closed subspace of H. We have, $H = Y \oplus Y^{\perp}$. Then, the function $P : H \to Y$ defined by $P(x) = y$ is an orthogonal projection of H onto Y.

Properties of Projection

1. P is idempotent. That is, $P^2 = P$
   Let $x \in H$ such that $x = y + z$. We have, $P(x) = y$. Then, $P^2(x) = P(y) = P(y + 0) = y = P(x)$.

2. $P|_Y$ is the identity operator on Y.
   Let $y \in Y \subset H$. Then $y = y + 0$ and $P(y) = y$.

Lemma 14.3.12 (Null Space)**.** *Let $H = Y \oplus Y^{\perp}$. Let $P : H \to Y$ be the projection defined by $P(x) = y$. Then $Y^{\perp} = \mathcal{N}(P)$.*

Proof. Let $H = Y \oplus Y^{\perp}$. Let $x \in H$. Then $x = y + z$ where $y \in Y$ and $z \in Y^{\perp}$. Let $P : H \to Y$ be the projection defined by $P(x) = y$.

Let $z \in Y^{\perp}$. Then $z = 0 + z$ and $P(z) = 0$. Thus, $Y^{\perp} \subset \mathcal{N}(P)$.

Let $z \in \mathcal{N}(P)$. Then, $P(z) = 0$. That is, $z = 0 + z$ where $z \in Y^{\perp}$. Thus, $\mathcal{N}(P) \subset Y^{\perp}$. □

Definitions 14.3.14 (Annihilator)**.** Let M be a nonempty subset of an inner product space H. Annihilator of M is given by, $M^{\perp} = \{x \in H : x \perp M\}$.

Remark. Let M be a nonempty subset of an inner product space H. Then, annihilator $M^{\perp}$ is a subspace of H.

Proof. Let $v \in M$. Let $x, y \in M^{\perp}$. Then $\langle x, v \rangle = 0$ and $\langle y, v \rangle = 0$.
Let $\alpha, \beta \in K$. Thus, $\langle \alpha x + \beta y, v \rangle = \alpha \langle x, v \rangle + \beta \langle y, v \rangle = 0$.
Clearly, $\alpha x + \beta y \in M^{\perp}$.

Also, we know that $\langle 0, v \rangle = 0$. Thus, $0 \in M^{\perp}$. Therefore $M^{\perp}$ is a subspace of H. □

Remark. Let M be a nonempty subset of an inner product space H. Then,

$$M \subset M^{\perp\perp}$$

Proof. Let $v \in M$. Then for any $x \in M^{\perp}$, we have $\langle x, v\rangle = 0$.
Thus, $v \in (M^{\perp})^{\perp} = M^{\perp\perp}$.

$$v \in M \implies v \perp M^{\perp} \implies v \in M^{\perp\perp}$$

Therefore, $M \subset M^{\perp\perp}$. □

Lemma 14.3.13 (closed subspace)**.** *Let Y be a closed subspace of a Hilbert space H. Then, $Y = Y^{\perp\perp}$.*

Proof. We know that $Y \subset Y^{\perp\perp}$. It is enough to prove that $Y^{\perp\perp} \subset Y$.
Let $x \in Y^{\perp\perp}$. Then $x \perp Y^{\perp}$. If $x = y + z$ where $y \in Y$ and $z \in Y^{\perp}$. Clearly, $y \in Y^{\perp\perp}$ and $z = x - y \in Y^{\perp\perp}$. Then $z = 0$ since $z \in Y^{\perp} \cap Y^{\perp\perp} = \{0\}$. Therefore, $x = y + 0 = y \in Y$. Thus, $Y = Y^{\perp\perp}$. □

Remark. Let Y be a closed subspace of Hilbert space H. Then, $H = Y \oplus Z$ where $Z = Y^{\perp}$. And, $H = Z \oplus Z^{\perp}$ where $Z^{\perp} = Y^{\perp\perp} = Y$. And $P_Z : H \to Z$ where $P_Z(x) = z$ where $x = y + z$ is another projection from H onto Z.

Lemma 14.3.14 (dense)**.** *Let M be a nonempty subset of a Hilbert space H. The span of M is dense in H if and only if $M^{\perp} = \{0\}$.*

Proof. Let $x \in M^{\perp}$. Suppose *span* $M = V$ is dense in H. Then $x \in V$ and there exists a sequence $\{x_n\}$ in V such that $x_n \to x$. Since $x \in M^{\perp}$ and $x_n \in$ *span* M, we have $\langle x_n, x\rangle = 0$. We know that, inner product is continuous. Thus, $\langle x_n, x\rangle \to \langle x, x\rangle = 0$. Clearly, $x = 0$.

Suppose $M^{\perp} = \{0\}$. If $x \in V^{\perp}$, then $x \perp V$. Thus, $x \in M^{\perp}$ since $V =$ *span* M. Clearly, $x = 0$. Therefore, $V^{\perp} = \{0\}$. Consider the closed subspace $\overline{V}$. We know that $V \subset \overline{V}$. Thus, $\overline{V}^{\perp} \subset V^{\perp} = \{0\}$. Thus, $H = \overline{V} \oplus \{0\} = \overline{V}$. Therefore, V is dense in H. □

14.3.4 Orthonormal sets and sequences

Definitions 14.3.15 (orthogonal)**.** Let H be an inner product space. Subset M of H is an orthogonal set if its elements are pairwise orthogonal.

$$\forall x, y \in M,\ x \perp y \qquad (x \neq y)$$

Definitions 14.3.16 (orthonormal)**.** Let H be an inner product space. Subset M of H is orthonormal set if it is an orthogonal set with elements of unit norm.

$$\langle x, y\rangle = \begin{cases} 0 & x \neq y \\ 1 & x = y \end{cases}$$

Remark. Let $M = \{x_1, x_2, \ldots, x_n\}$ be an orthogonal set.
Then by Pythagorean theorem,

$$\|x_1 + x_2\|^2 = \|x_1\|^2 + \|x_2\|^2 \qquad \text{since } x_1 \perp x_2$$

By finite mathematical induction,

$$\|x_1 + x_2 + \cdots + x_n\|^2 = \|x_1\|^2 + \|x_2\|^2 + \cdots + \|x_n\|^2$$

Lemma 14.3.15. *Orthonormal sets are linearly independent.*

Proof. Let $E = \{e_k : k \in I\}$ be an orthonormal set. Let S be a finite subset of the index set I. Suppose $\sum_{k\in S} \alpha_k e_k = 0$ where $\alpha_k \in K$. Fix $j \in S$. Then,

$$\left\langle \sum_{k\in S} \alpha_k e_k, e_j \right\rangle = \sum_{k\in S} \alpha_k \langle e_k, e_j\rangle$$
$$\langle 0, e_j\rangle = \sum_{k\in S} \alpha_k \delta_{k,j} = \alpha_j$$

Clearly, $\alpha_j = 0, \forall j \in S$. Thus, zero vector can be expressed as a finite linear combination of vectors only if all the scalars are zero. Therefore, M is linearly independent. □

Examples of Orthonormal Sets in different Hilbert Spaces

1. $\mathbb{R}^3,\ M = \{(1,0,0),(0,1,0),(0,0,1)\}$
2. $l^2,\ M = \{e_j\}$ where $e_j = \delta_{n,j}$
3. $C[0,2\pi],\ M = \{\cos nt : n = 0,1,2,\dots\}$
4. $C[0,2\pi],\ M = \{\sin nt : n \in \mathbb{N}\}$

Theorem 14.3.16 (Bessel Inequality)**.** *Let sequence $\{e_k\}$ be an orthonormal sequence in an inner product space X. Let $x \in X$. Then,*

$$\sum_{k=1}^{\infty} |\langle x, e_k\rangle|^2 \le \|x\|^2$$

Proof. Let $\{e_k\}$ be an orthonormal sequence in an inner product space X. Let $x \in X$. Let $Y_n = span\ \{e_1, e_2, \dots, e_n\}$. Then, there exists $y \in Y_n$ such that

$$y = \sum_{k=1}^{n} \langle x, e_k\rangle e_k$$

Define $z = x - y$. Then,

$$\begin{aligned}
\langle z, y\rangle &= \langle x - y, y\rangle \\
&= \langle x, y\rangle - \langle y, y\rangle \\
&= \left\langle x, \sum_{k=1}^{n} \langle x, e_k\rangle e_k \right\rangle - \|y\|^2 \\
&= \sum_{k=1}^{n} \overline{\langle x, e_k\rangle} \langle x, e_k\rangle - \|y\|^2 \\
&= \sum_{k=1}^{n} |\langle x, e_k\rangle|^2 - \sum_{k=1}^{n} |\langle x, e_k\rangle|^2 = 0
\end{aligned}$$

Thus, $z \perp y$. Therefore, we have $x = y + z$ where $y \in Y_n$ and $z \perp y$.

By Pythagorean theorem,

$$\|x\|^2 = \|y\|^2 + \|z\|^2$$

Therefore,

$$0 \leq \|z\|^2 = \|x\|^2 - \|y\|^2 = \|x\|^2 - \sum_{k=1}^{n} |\langle x, e_k\rangle|^2$$

Thus,

$$\sum_{k=1}^{n} |\langle x, e_k\rangle|^2 \leq \|x\|^2$$

The series is monotonic increasing and bounded by $\|x\|^2$. Therefore, converges and the sum is bounded by $\|x\|^2$.

$$\lim_{n\to\infty} \sum_{k=1}^{n} |\langle x, e_k\rangle|^2 = \sum_{k=1}^{\infty} |\langle x, e_k\rangle|^2 \leq \|x\|^2$$

□

Remark (Gram-Schmidt Process). Let X be a finite dimensional inner product space. Let set $\{x_1, x_2, \ldots, x_n\}$ be a set of n linearly independent vectors in X. Then an orthonormal set of vectors can be obtained by the following process.

Define $e_1 = x_1/\|x_1\|$. Then, $x_2 = \langle x_2, e_1\rangle\, e_1 + v_2$. Thus, $v_2 = x_2 - \langle x_2, e_1\rangle\, e_1$ and $v_2 \perp e_1$. Define $e_2 = v_2/\|v_2\|$. Continuing like this, we get

$$v_n = x_n - \sum_{k=1}^{n-1} \langle x_n, e_k\rangle\, e_k$$

such that $v_n \perp \mathit{span}\ \{e_1, e_2, \ldots, e_{n-1}\}$. And $e_n = v_n/\|v_n\|$.

14.3.5 Series related to Orthonormal sequences and sets

Remark. Define $u_n(t) = \cos nt$ and $v_n(t) = \sin nt$. Then

$$\langle u_j, u_k\rangle = \int_0^{2\pi} \cos jt\ \cos kt\ dt = \begin{cases} 0 & j \neq k \\ \pi & j = k \neq 0 \\ 2\pi & j = k = 0 \end{cases}$$

$$\langle v_j, v_k\rangle = \int_0^{2\pi} \sin jt\ \sin kt\ dt = \begin{cases} 0 & j \neq k \\ \pi & j = k \neq 0 \\ 2\pi & j = k = 0 \end{cases}$$

$$\langle u_j, v_k\rangle = \int_0^{2\pi} \cos jt\ \sin kt\ dt = \frac{1}{2}\int_0^{2\pi} \sin(k+j)t + \sin(k-j)t\ dt = 0$$

Clearly, $u_j \perp u_k\ (j \neq k)$, $v_j \perp v_k\ (j \neq k)$ and $\forall j, k,\ u_j \perp v_k$.

Theorem 14.3.17 (Fourier). *Any periodic, (reasonably) continuous function may be expressed as series of sine and cosine terms.*

Remark. We may write $x \in C[0,2\pi]$ as $x = y + z$ where $y, z \in C[0,2\pi]$ such that

$$y(t) = \sum_{k=0}^{\infty} \frac{\langle x, u_k\rangle}{\pi} u_k = \sum_{k=0}^{\infty} a_k \cos kt \text{ where } a_k = \frac{1}{\pi}\int_0^{2\pi} x(t)\cos kt \, dt$$

$$z(t) = \sum_{k=0}^{\infty} \frac{\langle x, v_k\rangle}{\pi} v_k = \sum_{k=0}^{\infty} b_k \sin kt \text{ where } b_k = \frac{1}{\pi}\int_0^{2\pi} x(t)\sin kt \, dt$$

The coefficients a_k, b_k's are the **Fourier Coefficients** of $x(t)$.

Challenge 7. Find the orthonomal series that gives the Fourier representation of $x(t)$? [Apostol, 1973]

Theorem 14.3.18 (convergence)**.** *Let $\{e_k\}$ be an orthonormal sequence in a Hilbert space H. Then,*

1. *The series $\sum \alpha_k e_k$ converges* [15] *if and only if $\sum |\alpha_k|^2$ converges.*
2. *If $\sum \alpha_k e_k$ converges, then coefficients α_k are the Fourier coefficients $\langle x, e_k\rangle$ of the sum x.*
$$x = \sum_{k=1}^{\infty} \langle x, e_k\rangle e_k$$
3. *For any $x \in H$, the series $\sum \langle x, e_k\rangle e_k$ converges.*

Proof. Consider the series $\sum \alpha_k e_k$ and the sequence of partial sums $\{s_n\}$.

$$s_n = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_n e_n$$

Similarly, consider the series $\sum |\alpha_k|^2$ and the sequence of partial sums $\{\sigma_n\}$.

$$\sigma_n = |\alpha_1|^2 + |\alpha_2|^2 + \cdots + |\alpha_n|^2$$

$$\begin{aligned}
\|s_n - s_m\|^2 &= \|\alpha_{m+1}e_{m+1} + \alpha_{m+2}e_{m+2} + \cdots + \alpha_n e_n\| \\
&= |\alpha_{m+1}|^2 + |\alpha_{m+2}|^2 + \cdots + |\alpha_n|^2 \text{ since } \langle e_n, e_m\rangle = \delta_{n,m} \\
&= |\sigma_n - \sigma_m|
\end{aligned}$$

Clearly, $\{s_n\}$ is Cauchy if and only if $\{\sigma_n\}$ is Cauchy. We know that, $\mathbb{R}, H$ are complete spaces and every Cauchy sequence in these spaces converges. Therefore, $\sum \alpha_k e_k$ converges if and only if $\sum |\alpha_k|^2$ converges.

Suppose $\sum \alpha_k e_k$ converges to sum x. Consider the sequence of partial sum $\{s_n\}$. Then, $s_n \to x$. Fix e_k,

$$\langle s_n, e_k\rangle = \left\langle \sum_{j=1}^{n} \alpha_j e_j, e_k \right\rangle = \sum_{j=1}^{n} \langle \alpha_j e_j, e_k\rangle = \sum_{j=1}^{n} \alpha_j \delta_{j,k} = \alpha_k$$

[15] The convergence in inner product space is defined in terms of the norm induced by the inner product. That is, In an inner product space $x_n \to x$ if and only if $\|x_n - x\| \to 0$ where $\|x_n - x\| = (\langle x_n - x, x_n - x\rangle)^{\frac{1}{2}}$.

Since inner product is continuous $\langle s_n, e_j\rangle \to \langle x, e_j\rangle$ as $n \to \infty$. Thus, $\alpha_j = \langle x, e_j\rangle$. Therefore,

$$x = \sum_{k=1}^{\infty} \alpha_k e_k = \sum_{k=1}^{\infty} \langle x, e_k\rangle e_k$$

Let H be a Hilbert space. Let $x \in H$ and $\{e_k\}$ be an orthonormal sequence in H. Define $\alpha_k = \langle x, e_k\rangle$. By Bessel Inequality,

$$\sum_{k=1}^{\infty} |\alpha_k|^2 = \sum_{k=1}^{\infty} |\langle x, e_k\rangle|^2 \leq \|x\|^2$$

Therefore, $\sum |\alpha_k|^2$ converges. We know that, if $\sum |\alpha_k|^2$ converges then $\sum \alpha_k e_k$ converges and the sum is given by

$$x = \sum_{k=1}^{\infty} \alpha_k e_k$$

$\square$

Lemma 14.3.19 (Fourier coefficients)**.** *Any x in an inner product space has at most countably many non-zero Fourier coefficients with respect to an orthonormal family $\{e_k : k \in I\}$ in X.*

Proof. **Step 1 :**
Let $x \in X$. Suppose $x = \sum \alpha_x e_k$. Without loss of generality, consider the rearrange of $\{e_k\}$ such that $x = \sum \langle x, e_k\rangle e_k$. Then,

$$\sum_{k=1}^{n} |\langle x, e_k\rangle|^2 \leq \|x\|^2, \ \forall n$$

Let $m \in \mathbb{N}$. Then the number of e_k's such that $|\langle x, e_k\rangle| > \frac{1}{m}$ is less than $m\|x\|^2$. Thus the number of e_k's such that $|\langle x, e_k\rangle| > 0$ is countable since it is a countable union of finite sets.

Step 2 : Independent of rearrangements
It remains to prove that, The series representation is independent of rearrangemet of elements in the orthonormal sequence $\{e_k\}$.

Let $\{w_k\}$ be a rearrangement of $\{e_k\}$. Then, there exists a bijection $\phi : w \to e$ given by $\phi(w_j) = e_k$ where $j \to k$ is a bijection on $\mathbb{N}$.

Define $\alpha_k = \langle x, e_k\rangle$, $\beta_j = \langle x, w_j\rangle$,

$$x_1 = \sum_{n=1}^{\infty} \alpha_n e_n \quad \text{and} \quad x_2 = \sum_{n=1}^{\infty} \beta_n w_n$$

Then,

$$\alpha_j = \langle x, e_j\rangle = \langle x_1, e_j\rangle \quad \textit{and} \quad \beta_k = \langle x, w_k\rangle = \langle x_2, w_k\rangle$$

$$\begin{aligned}\langle x_1 - x_2, e_j\rangle &= \langle x_1, e_j\rangle - \langle x_2, e_j\rangle \\ &= \langle x, e_j\rangle - \langle x_2, e_j\rangle = 0\end{aligned}$$

Similarly, $\langle x_1 - x_2, w_k\rangle = 0$.

$$\begin{aligned}\|x_1 - x_2\|^2 &= \left\langle x_1 - x_2, \sum \alpha_n e_n - \sum \beta_k w_k \right\rangle \\ &= \sum \overline{\alpha}_j \langle x_1 - x_2, e_j\rangle - \sum \overline{\beta}_k \langle x_1 - x_2, w_k\rangle = 0\end{aligned}$$

Clearly, $x_1 = x_2$. Thus, the sum is the same for representation of x with respect to any rearrangement of $\{e_k\}$. □

14.3.6 Total Orthonormal sets and sequences

Definitions 14.3.17 (total orthonormal)**.** A subset M in a vector X is total if the span of M is dense in X. Total orthonormal is a subset which is total and orthonormal in inner product space X.

Every non-empty Hilbert space has a total orthonormal subset.

Proof. It is included in module IV. □

In a Hilbert space H, total orthonormal subsets have same cardinality.

Proof. It is out of our scope. This cardinality is the Hilbert dimension or orthonormal dimension of H. □

Theorem 14.3.20 (duality)**.** *Let M a subset of an inner product space X. Then*

1. *If M total in X, then there does not exists a non-zero $x \in X$ which is orthogonal to every element of M.*

$$x \perp M \implies x = 0$$

2. *If X is complete and X has no non-zero element orthogonal to M, then M is total in X.*

If M is total in X, then $M^\perp = \{0\}$. Conversely, if X is complete and $M^\perp = \{0\}$, then M is total in X.

Proof. Let H be the completion of inner product space X. Then, X is dense in H. And *span* M is dense in X since M is total. Thus, *span* M is dense in H.

Thus, orthogonal complement of M in H is $\{0\}$. Thus, if $x \in X$ and $x \perp M = 0$. Then, $x \in H$ and $x = y + 0 \in M \oplus M^\perp$ and $x \perp M = y = 0$. Therefore, $x = 0$.

Let X be a Hilbert space. Suppose X has no non-zero element orthogonal to M. Then, $M^\perp = \{0\}$. By lemma, *span* M is dense in X since $M^\perp = \{0\}$. Therefore, M is total in X. □

Theorem 14.3.21. *An orthonormal subset M is total in a Hilbert space if and only if Parsevals relation is true for any $x \in H$.*

$$\sum_k |\langle x, e_k\rangle|^2 = \|x\|^2$$

Proof. Suppose Parsevals relation is true for each $x \in H$. Suppose M is not total in Hilbert space H. Then, there exist $x \in H$ such that $x \perp M$ and $x \neq 0$. Since $x \in M^\perp$, we have $\forall k,\ \langle x, e_k\rangle = 0$, then $\|x\|^2 = 0$ by Parsevals relation. And $x = 0$ since $\|x\| = 0$. This is a contradiction. Therefore, M is total in H.

Suppose M is total in H. Then $M^\perp = \{0\}$. Let $x \in H$. We know that, there are at most countably many non-zero Fourier coefficients for any $x \in H$. Arrange non-zero Fourier coefficients of x in a (finite) sequence $\{\langle x, e_k\rangle\}$.

Define

$$y = \sum_k \langle x, e_k\rangle\, e_k$$

Case 1 : $e_k \in M$ **such that** $\langle x, e_k\rangle \neq 0$
Fix an e_j such that $\langle x, e_j\rangle \neq 0$. Then,

$$\begin{aligned}
\langle x-y, e_j\rangle &= \langle x, e_j\rangle - \langle y, e_j\rangle \\
&= \langle x, e_j\rangle - \left\langle \sum_k \langle x, e_k\rangle\, e_k, e_j \right\rangle \\
&= \langle x, e_j\rangle - \sum_k \langle x, e_k\rangle\ \langle e_k, e_j\rangle \\
&= \langle x, e_j\rangle - \sum_k \langle x, e_k\rangle\, \delta_{j,k} \\
&= \langle x, e_j\rangle - \langle x, e_j\rangle = 0
\end{aligned}$$

Thus, if $\langle x, e_j\rangle \neq 0$, then $x - y \perp e_j$.

Case 2 : $v \in M$ **such that** $\langle x, v\rangle = 0$
Fix $v \in M$ such that $\langle x, v\rangle = 0$. Then,

$$\langle x-y, v\rangle = \langle x, v\rangle - \langle y, v\rangle = \langle x, v\rangle - \sum_k \langle x, e_k\rangle \langle e_k, v\rangle = 0$$

since $\langle x, v\rangle = 0$ and $\forall k,\ \langle e_k, v\rangle = 0$.

Clearly, from cases 1 & 2, we have $x - y \perp M$. Thus, $x - y \in M^\perp = \{0\}$. Therefore, $x = y$.

$$\begin{aligned}
\|x\|^2 &= \left\langle \sum_k \langle x, e_k\rangle\, e_k, \sum_m \langle x, e_m\rangle\, e_m \right\rangle \\
&= \sum_k \langle x, e_k\rangle\, \overline{\langle x, e_k\rangle} \\
&= \sum_k |\langle x, e_k\rangle|^2
\end{aligned}$$

Thus, Parsevals relation is true for any $x \in H$. $\square$

Theorem 14.3.22. *Let H be a Hilbert space. Then,*

1. *If H is separable, then every orthonormal subset in H is countable.*

2. *If H contains an orthonormal sequence which is total in H, then H is separable.*

Proof. Let H be a Hilbert space. Suppose H is separable. Then, H has a dense subset, say B. And H has an orthonormal subset which may be obtained from a Hamel basis by Gram-Schimdt process, say M.

Let $x, y \in M$. Then,

$$\|x - y\|^2 = \langle x - y, x - y\rangle = \langle x, x\rangle + \langle y, y\rangle = 2$$

Thus, $d(x, y) = \sqrt{2}$ for any $x, y \in M$. Thus, x, y have disjoint neighbourhoods N_x, N_y respectively. We know that, the dense set B intersects with any neighbourhood in H. Let $b \in N_x \cap B$ and $\tilde{b} \in N_y \cap B$. Clearly, $b \neq \tilde{b}$ since $N_x \cap N_y = \phi$.

Suppose M is uncountable. Then there exists uncountably many such pair $(b, \tilde{b})$ of elements in B for each pair (x, y) of elements in M. This is a contradiction since B is countable. Therefore, M is countable.

—continue— $\square$

Definitions 14.3.18 (isomorphism)**.** Let $H, \tilde{H}$ be Hilbert spaces. A bijection $T : H \to \tilde{H}$ preserving inner product is an isomorphism. That is,

$$\langle Tx, Ty\rangle = \langle x, y\rangle$$

Theorem 14.3.23. *Two Hilbert space over the same field are isomorphic if and only if they have the same Hilbert dimension.*

Proof. Let $T : H \to \tilde{H}$ be an isomorphism. Let M be a total, orthonormal subset of H. Then $T(M)$ is a total, orthonormal subset of $\tilde{H}$. Thus, these sets have same cardinality. That is, they have same Hilbert dimension.

Suppose $H, \tilde{H}$ are two Hilbert spaces over the same field K. Suppose they have same Hilbert dimension. Then they have total, orthonormal subsets $M_1 = \{e_k\}$ and $M_2 = \{w_j\}$ respectively. Let $x \in H$. Then,

$$x = \sum_{k \in I} \langle x, e_k\rangle\, e_k$$

where $\langle x, e_k\rangle \neq 0$ for at most countably many e_k's. And from Bessel inequality, we know that this series converges. Define $T : H \to \tilde{H}$ such that $e_k \to w_k$ by rearranging $\{w_j\}$ if necessary.†[16] Consider,

$$\begin{aligned} T(x) &= T\left(\sum_{k\in I} \langle x, e_k\rangle\, e_k\right) \\ &= \sum_{k\in I} \langle x, e_k\rangle\; T(e_k) \\ &= \sum_{k\in I} \langle x, e_k\rangle\; w_k \end{aligned}$$

This series converges, thus $T(x) = \tilde{x} \in \tilde{H}$.

And T preserves norms since,

$$\begin{aligned} \|\tilde{x}\|^2 &= \|T(x)\|^2 \\ &= \sum_{k\in I} \langle x, e_k\rangle\, w_k \\ &= \sum_{k\in I} |\langle x, e_k\rangle|^2 = \|x\|^2 \end{aligned}$$

Thus, $\|x - y\| = \|T(x - y)\| = \|Tx - Ty\| = 0$. Therefore, T is injective.

Suppose $\tilde{x} = \sum_{k\in I} \alpha_k w_k$. Then $\sum |\alpha_k|^2 < \infty$ by Bessel inequality. Thus, $x = \sum_{k\in I} \alpha_k e_k \in H$. Therefore, T is surjective.

And T preserves inner product since,

$$\begin{aligned} \langle \tilde{x}, \tilde{y}\rangle &= \langle Tx, Ty\rangle \\ &= \left\langle \sum_{k\in I} \langle x, e_k\rangle\, w_k, \sum_{j\in I} \langle y, e_j\rangle\, w_j \right\rangle \\ &= \sum_{k\in I} < x, e_k > \overline{< y, e_k >} \\ &= \left\langle \sum_{k\in I} \langle x, e_k\rangle\, e_k, \sum_{j\in I} \langle y, e_j\rangle\, e_j \right\rangle \\ &= \langle x, y\rangle \end{aligned}$$

□

14.3.8 Representation of Functionals on Hilbert spaces

Theorem 14.3.24 (Riesz)**.** *Every bounded linear functional f on a Hilbert space H can be represented in terms of inner product,*

$$f(x) = \langle x, z\rangle$$

where z is uniquely determined by f and $\|z\| = \|f\|$.

[16] We know that, for Hilbert spaces the representation exists against any rearrangement of orthonormal sequence.

Proof. Let f be a bounded, linear functional on H. If $f = 0$ or $x = 0$, then the resutl is trivial.

Step 1 : f has a representation
Suppose $f \neq 0$. If $\mathcal{N}(f) = H$. Then $f(x) =< x, z >= 0$ for any $x \in H$ which is a contradiction since $f \neq 0$. Thus, $\mathcal{N}(f) \neq H$ and $\mathcal{N}(f)^{\perp} \neq \phi$.

Let $z_0 \in \mathcal{N}(f)^{\perp}$. Let $x \in H$. Define $v = f(x)z_0 - f(z_0)x$. Then $f(v) = f(x)f(z_0) - f(z_0)f(x) = 0$. Thus, $v \in \mathcal{N}(f)$.

$$\begin{aligned} 0 &= \langle v, z_0 \rangle \\ &= \langle f(x)z_0 - f(z_0)x, z_0 \rangle \\ &= f(x) \langle z_0, z_0 \rangle - f(z_0) \langle x, z_0 \rangle \end{aligned}$$

Then,

$$f(x) = \frac{f(z_0)}{\langle z_0, z_0 \rangle} \langle x, z_0 \rangle = \langle x, z_0 \rangle$$

Therefore, f has a represenation, $f(x) = \langle x, z \rangle$ for any $x \in H$ where z_0 is a non-vanishing point of f and

$$z = \frac{f(z_0)}{\langle z_0, z_0 \rangle} z_0$$

Step 2 : z is unique
Suppose $f(x) = \langle x, z_1 \rangle = \langle x, z_2 \rangle$ for all $x \in H$. Then $\langle x, z_1 - z_2 \rangle = \langle x, z_1 \rangle - \langle x, z_2 \rangle = 0$. Put $x = z_1 - z_2$. Then $\langle z_1 - z_2, z_1 - z_2 \rangle = \|z_1 - z_2\|^2 = 0$. Thus, $z_1 = z_2$. And the representation $f(x) = \langle x, z \rangle$ is unique.

Step 3 : norm of f
Since f is bounded,

$$\|z\|^2 = \langle z, z, = \rangle f(z) \leq \|f\| \ \|z\|$$

Thus, $\|z\| \leq \|f\|$.

Also we have,

$$|f(x)| \leq |\langle x, z \rangle| \leq \|x\| \ \|z\|$$

The inequality if true for any $x \in H$. Thus,

$$\|f\| = \sup_{x \in H} \frac{|f(x)|}{\|x\|} \leq \|z\|$$

Therefore, $\|f\| = \|z\|$ □

Lemma 14.3.25 (equality)**.** *Let $\langle v_1, w \rangle = \langle v_2, w \rangle$ for each w in an inner product space X. Then $v_1 = v_2$. In particular, if $\langle v_1, w \rangle = 0$ for every $w \in X$, then $v_1 = 0$.*

Proof. Let v_1, v_2 be vectors in an inner product space X. Suppose $\langle v_1, w, =\rangle \langle v_2, w\rangle$ for each $w \in X$. Then,

$$\langle v_1 - v_2, w\rangle = \langle v_1, w\rangle - \langle v_2, w\rangle = 0$$

Let $w = v_1 - v_2$. Then $\|v_1 - v_2\|^2 = \langle v_1 - v_2, v_1 - v_2\rangle = 0$. Thus, $v_1 = v_2$.

Suppose $\langle v_1, w\rangle = 0$ for each $w \in X$. Then $\|v_1\|^2 = \langle v_1, v_1\rangle = 0$. Therefore, $v_1 = 0$. □

Definitions 14.3.19 (sesquilinear form)**.** Let X, Y be vector spaces over the same field K. Then a sesequilinear form is a mapping $h : X \times Y \to K$ such that

1. $h(x_1 + x_2, y) = h(x_1, y) + h(x_2, y)$
2. $h(x, y_1, y_2) = h(x, y_1) + h(x, y_2)$
3. $h(\alpha x, y) = \alpha h(x, y)$
4. $h(x, \beta y) = \overline{\beta} h(x, y)$

Definitions 14.3.20 (bilinear form)**.** Let X, Y be vector spaces over the same field K. Then a bilinear form is a mapping $h : X \times Y \to K$ such that

1. $h(x_1 + x_2, y) = h(x_1, y) + h(x_2, y)$
2. $h(x, y_1, y_2) = h(x, y_1) + h(x, y_2)$
3. $h(\alpha x, y) = \alpha h(x, y)$
4. $h(x, \beta y) = \beta h(x, y)$

Definitions 14.3.21. Let X, Y be inner product spaces over the same field K. A function $h : X \times Y \to K$ is bounded if

$$|h(x, y)| \le c\|x\| \ \|y\|$$

Definitions 14.3.22. Suppose $h : X \times Y \to K$ is bounded. Then,

$$\|h\| = \sup_{\substack{x \neq 0 \\ y \neq 0}} \frac{|h(x, y)|}{\|x\| \ \|y\|}$$

Theorem 14.3.26 (Riesz representation)**.** *Let H_1, H_2 be Hilbert spaces over K. And function $h : H_1 \times H_2 \to K$ be a bounded, sesequilinear form. Then h has representation,*

$$h(x, y) = \langle Sx, y\rangle$$

where $S : H_1 \to H_2$ is a bounded, linear operator uniquely determined by h and $\|h\| = \|S\|$.

Proof. **Step 1 : Constructon of S**
Let $h : H_1 \times H_2 \to K$ be a bounded, sesquilinear form. Then $f(y) = \overline{h(x, y)}$ is a bounded, linear functional on H_2. By Riesz's theorem for bounded functionals, f has an inner product representation, $f(y) = \langle y, z\rangle$. Thus, $h(x, y) = \langle z, y\rangle$.

Define an operator $S : H_1 \to H_2$ such that $S(x) = z$. Then $h(x, y) = \langle Sx, y\rangle$.

Step 2 : S is linear
Let $x_1, x_2 \in H_1$, $y \in H_2$ and $\alpha, \beta \in K$. We have $\langle Sx, y\rangle = h(x, y)$. Thus,

$$\begin{aligned}
\langle S(\alpha x_1 + \beta x_2), y\rangle &= h(\alpha x_1 + \beta x_2, y) \\
&= \alpha h(x_1, y) + \beta h(x_2, y) \\
&= \alpha \langle S(x_1), y\rangle + \beta \langle S(x_2), y\rangle \\
&= \langle \alpha S(x_1) + \beta S(x_2), y\rangle
\end{aligned}$$

Clearly, $S(\alpha x_1 + \beta x_2) = \alpha S(x_1) + \beta S(x_2)$. Therefore, S is linear.

Step 3 : S is bounded
Since h is a bounded sesquilinear form. We have,

$$\begin{aligned}
\|h\| &= \sup_{\substack{x\neq 0 \\ y\neq 0}} \frac{|h(x, y)|}{\|x\| \; \|y\|} \\
&= \sup_{\substack{x\neq 0 \\ Sx\neq 0}} \frac{|\langle Sx, y\rangle|}{\|x\| \; \|y\|}
\end{aligned}$$

Suppose $y = Sx$. Then,

$$\begin{aligned}
\|h\| &\geq \sup_{\substack{x\neq 0 \\ Sx\neq 0}} \frac{|\langle Sx, Sx\rangle|}{\|x\| \; \|Sx\|} \\
&\geq \sup_{\substack{x\neq 0 \\ Sx\neq 0}} \frac{\|Sx\|}{\|x\|} = \|S\|
\end{aligned}$$

Thus, S is bounded and $\|S\| \leq \|h\|$.

By Schwarz inequality, $|\langle Sx, y\rangle| \leq \|Sx\| \; \|y\|$. Thus,

$$\|h\| = \sup_{\substack{x\neq 0 \\ y\neq 0}} \frac{|\langle Sx, y\rangle|}{\|x\| \; \|y\|} \leq \sup_{\substack{x\neq 0 \\ y\neq 0}} \frac{\|Sx\| \; \|y\|}{\|x\| \; \|y\|} = \sup_{x\neq 0} \frac{\|Sx\|}{\|x\|} = \|S\|$$

Therefore $\|S\| = \|h\|$.

Suppose $T : H_1 \to H_2$ be a bounded, linear operator. Then

$$h(x, y) = \langle Sx, y\rangle = \langle Tx, y\rangle$$

Clearly, $Sx = Tx$ for each $x \in H_1$. Therefore, S is unique. $\square$

Module 4 - Hilbert-Adjoint Operator & Hahn-Banach Theorem

14.4 Hilbert Spaces

14.4.1 Hilbert-Adjoint Operator

Theorem 14.4.1 (Hilbert-Adjoint operator)**.** *Let H_1, H_2 be Hilbert spaces. Let $T : H_1 \to H_2$ be any bounded, linear operator. Then there exists a unique, bounded, linear operator $T^* : H_2 \to H_1$ such that $\langle Tx, y\rangle = \langle x, T^*y\rangle$ and $\|T^*\| = \|T\|$.*

Proof. Let $T : H_1 \to H_2$ be a bounded, linear operator. Then, we have $h(y, x) = \langle y, Tx\rangle$ is a sesquilinear form on $H_2 \times H_1$, by the linearity of T and conjugate linearity of inner product.

By Schwarz inequality,

$$|h(y, x)| = |\langle y, Tx\rangle| \leq \|y\| \, \|Tx\| \leq \|T\| \, \|x\| \, \|y\|$$

Therefore, h is bounded and $\|h\| \leq \|T\|$.

Also we have,

$$\|h\| = \sup_{\substack{x\neq 0\\ y\neq 0}} \frac{|\langle y, Tx\rangle|}{\|y\| \, \|x\|} \geq \sup_{\substack{x\neq 0\\ Tx\neq 0}} \frac{|\langle Tx, Tx\rangle|}{\|Tx\| \, \|x\|} = \sup_{x\neq 0} \frac{\|Tx\|}{\|x\|} = \|T\|$$

Therefore, $\|h\| = \|T\|$.

By Riesz theorem, h has an inner product representation using a unique linear operator $S = T^*$ given by,

$$h(y, x) = \langle T^*y, x\rangle$$

where $\|T^*\| = \|h\|$. Therefore, $\langle y, Tx\rangle = \langle T^*y, x\rangle$ and $\|T\| = \|T^*\|$. □

Definitions 14.4.1 (Hilbert-Adjoint operator)**.** Let H_1, H_2 be Hilbert spaces. Let $T : H_1 \to H_2$ be a bounded, linear operators. Then Hilbert-Adjoint operator T^* of T is given by

$$T^* : H_2 \to H_1 \text{ such that } \forall x \in H_1, y \in H_2,\ \langle Tx, y\rangle = \langle x, T^*y\rangle$$

Lemma 14.4.2 (zero vector)**.** *Let X, Y be inner product spaces. Let $Q : X \to Y$ be a bounded, linear operator. Then*

1. *$Q = 0$ if and only if $\langle Qx, y\rangle = 0$ for all $x \in X$ and $y \in Y$.*
2. *If $Q : X \to X$, X is complex, and $\forall x \in X$, $\langle Qx, x\rangle = 0$, then $Q = 0$.*

Proof. If $Q = 0$, then $Qx = 0$ and $\langle Qx, y\rangle = \langle 0, y\rangle = 0$.

Suppose $\forall x \in X$, $\forall y \in Y$, $\langle Qx, y\rangle = 0$. Since $\langle y, y\rangle = 0$ if and only if $y = 0$. Thus, zero is the only vector in Y which is orthogonal to entire Y. Then

$\forall x \in X,\ Qx = 0$. That is, $Q = 0$.

Suppose $Q : X \to X$ such that for any $x \in X,\ \langle Qx, x\rangle = 0$. Let $\alpha x + y \in X$. Then

$$\begin{aligned} 0 &= \langle Q(\alpha x + y, \alpha x + y\rangle \\ &= |\alpha|^2 \langle Qx, x\rangle + \langle Qy, y\rangle + \alpha \langle Qx, y\rangle + \overline{\alpha} \langle Qy, x\rangle \end{aligned}$$

Put $\alpha = 1$, we get $\langle Qx, y\rangle + \langle Qy, x\rangle = 0$.
Put $\alpha = i$, we get $\langle Qx, y\rangle - \langle Qy, x\rangle = 0$. Thus, $\langle Qx, y\rangle = 0$ for any $y \in X$. Therefore, $Qx = 0$ for any $x \in X$ and Thus $Q = 0$. ☐

Theorem 14.4.3 (properties of Hilbert-Adjoint operators)**.** *Let H_1, H_2 be Hilbert spaces. Let $\alpha \in K$. Let $S : H_1 \to H_2$ and $T : H_1 \to H_2$ be bounded linear operators. Then,*

1. $\langle T^* y, x\rangle = \langle y, Tx\rangle$
2. $(S + T)^* = S^* + T^*$
3. $(\alpha T)^* = \overline{\alpha} T^*$
4. $(T^*)^* = T$
5. $\|T^*T\| = \|TT^*\| = \|T\|^2$
6. $T^*T = 0 \iff T = 0$
7. $(ST)^* = T^*S^*$

Proof. 1. $\langle T^* y, x\rangle = \langle y, Tx\rangle$

$$\begin{aligned} \langle T^* y, x\rangle &= \overline{\langle x, T^* y\rangle} \\ &= \overline{\langle Tx, y\rangle} \\ &= \langle y, Tx\rangle \end{aligned}$$

---

2. $(S + T)^* = S^* + T^*$

$$\begin{aligned} \langle x, (S + T)^* y\rangle &= \langle (S + T)x, y\rangle \\ &= \langle Sx, y\rangle + \langle Tx, y\rangle \\ &= \langle x, S^* y\rangle + \langle x, T^* y\rangle \\ &= \langle x, (S^* + T^*)y\rangle \end{aligned}$$

---

3. $(\alpha T)^* = \overline{\alpha} T^*$

$$\begin{aligned} \langle x, (\alpha T)^* y\rangle &= \overline{\alpha} \langle x, T^* y\rangle \\ &= \overline{\alpha} \langle Tx, y\rangle \\ &= \left\langle \overline{(\alpha} T)x, y \right\rangle \\ &= \langle x, (\overline{\alpha} T)^* y\rangle \end{aligned}$$

---

4. $(T^*)^* = T$

$$\begin{aligned}\langle x, (T^*)^* y\rangle &= \langle T^* x, y\rangle \\ &= \overline{\langle y, T^* x\rangle} \\ &= \overline{\langle Ty, x\rangle} \\ &= \langle x, Ty\rangle\end{aligned}$$

---

5. $\|T^*T\| = \|TT^*\| = \|T\|^2$

$$\begin{aligned}\|Tx\|^2 &= \langle Tx, Tx\rangle \\ &= \langle Tx, (T^*)^* x\rangle \\ &= \langle T^*Tx, x\rangle\end{aligned}$$

By Schwarz inequality, $\langle T^*Tx, x\rangle \leq \|T^*Tx\| \; \|x\|$.

$$\begin{aligned}\|Tx\|^2 &\leq \|T^*Tx\| \; \|x\| \\ &\leq \|T^*T\| \; \|x\|^2\end{aligned}$$

Taking supremum of $\|x\| = 1$, we get $\|T\|^2 \leq \|T^*T\|$. Since T^*, T are bounded, linear operators, we have $\|T^*T\| \leq \|T^*\| \; \|T\|$. Also we have, $|T^*\| = \|T\|$. Thus,

$$\|T\|^2 \leq \|T^*T\| \leq \|T^*\| \; \|T\| \leq \|T\|^2$$

Therefore, $\|T^*T\| = \|T\|^2$.

Replace T by T^* we get,

$$\|(T^*)^*T^*\| = \|T^*\|^2 = \|T\|^2$$

---

6. $T^*T = 0 \iff T = 0$

$$\|T\|^2 = \|T^*T\| = 0 \iff \|T\| = 0$$

---

7. $(ST)^* = T^*S^*$

$$\begin{aligned}\langle (ST)x, y\rangle &= \langle Tx, S^*y\rangle \\ \implies \langle x, (ST)^*y\rangle &= \langle x, (T^*S^*)y\rangle\end{aligned}$$

□

14.4.2 Self-Adjoint, Unitary and Normal Operators

Definitions 14.4.2. Let H be a Hilbert space. Let $T : H \to H$ be a bounded, linear operator.

- Then T is self-adjoint/Hermitian if $T^* = T$.

- Then T is unitary, if $T^* = T^{-1}$.
- Then T is normal, if $TT^* = T^*T$.

Remark. If T is unitary, then $T^*T = T^{-1}T = TT^{-1} = TT^*$. If T is self-adjoint, then $T^*T = T^2$. Therefore, if T is self-adjoint or unitary, then T is normal. And $2iI$ is a normal operator which is neither self-adjoint or unitary.

Remark. Let $T : \mathbb{C}^n \to \mathbb{C}^n$ be a linear operator. Let A, B be the matrices of transformations T, T^* with respect to the standard basis on $\mathbb{C}^n$. Then

$$\langle Tx, y\rangle = x^t A^t \overline{y} = \langle x, T^* y\rangle = x^t \overline{B}\overline{y}$$

Thus, $B = \overline{A}^t$.

Remark. Suppose T is self-adjoint/Hermitian, then $A = \overline{A}^t$. ie, A is Hermitian. If K is real then A is symmetric since $A = A^t$.

Suppose T is unitary, then $A^{-1} = \overline{A}^t$. ie, A is unitary. If K is real, then A is orthogonal since $A^{-1} = A^t$.

Suppose T is normal, then $BA = AB$ where $B = \overline{A}^t$. ie, A is normal since A commutes with its conjugate transpose.

Theorem 14.4.4 (self-adjoint)**.** *Let T be a bounded, linear opeator on a Hilbert space H.*

1. *If T is self-adjoint, $\langle Tx, x\rangle$ is real for any $x \in H$.*
2. *If H is complex and $\langle Tx, x\rangle$ is real for all $x \in E$. Then T is self-adjoint.*

Proof. Suppose T is self-adjoint. Then $< Tx, y >=< x, Ty >$.

$$\langle Tx, x\rangle = \langle x, Tx\rangle = \overline{\langle Tx, x\rangle}$$

Therefore, $\langle Tx, x\rangle$ is real.

Suppose H is complex, Hilbert space. Suppose $\langle Tx, x\rangle$ is real for any $x \in H$. Then,

$$\langle Tx, x\rangle = \overline{\langle x, Tx\rangle} = \langle x, Tx\rangle$$

And,

$$\langle Tx, x\rangle = \langle x, T^* x\rangle$$

Thus,

$$\langle x, (T - T^*)x\rangle = \langle x, 0\rangle = 0, \ \forall x \in H$$

Thus, $T - T^* = 0$(by zero vector lemma). Therefore, $T = T^*$. ☐

Theorem 14.4.5 (self-adjoint product)**.** *The product of two self-adjoint, bounded operators S, T is self-adjoint if and only if S, T commute.*

Proof. We have, $(ST)^* = T^*S^* = TS$. Thus,

$$(ST)^* = ST \iff TS = ST$$

☐

Theorem 14.4.6. *Let T_n be a sequence of self-adjoint, bounded linear operators $T_n : H \to H$. Suppose $\{T_n\}$ converges to T with respect to the norm on $B(H, H)$. Then the limit operator T is a self-adjoint, bounded, linear operator on H.*

Proof. Let $T_n : H \to H$ be self-adjoint, bounded, linear operator for each $n \in \mathbb{N}$. Suppose $T_n \to T$. Since T_n are bounded and linear, T is also bounded and linear.

It remains to prove that T is self-adjoint. Let $(T_n - T)^*$ be the adjoint operator of $T_n - T$. Then $\|(T_n - T)^*\| = \|T_n - T\|$. Therefore,

$$\|T_n^* - T^*\| = \|(T_n - T)^*\| = \|T_n - T\|$$

We have, $T_n^* = T_n$ and $\|T_n^* - T_n\| = 0$. Thus,

$$\begin{aligned}\|T - T^*\| &\leq \|T - T_n\| + \|T_n - T_n^*\| + \|T_n^* - T^*\| \\ &\leq 2\|T_n - T\|\end{aligned}$$

As $n \to \infty$, $\|T_n - T\| \to 0$ since $T_n \to T$. Clearly the left side is independent of the limit, and $\|T - T^*\| = 0$. Thus, $T - T^* = 0$. In other words, $T = T^*$. ☐

Theorem 14.4.7. *Let H be a Hilbert space. Let $U : H \to H$ and $V : H \to H$ are unitary. Then*

1. *U is isometric. ie, $\forall x \in H$, $\|Ux\| = \|x\|$*
2. *$\|U\| = 1$*
3. *U^{-1} is unitary*
4. *UV is unitary*
5. *U is normal*
6. *A bounded, linear operator T on a complex, Hilbert space H is unitary if and only if T is isometric and surjective.*

Proof. Let H be a Hilbert space. Let U, V be unitary operators. Then $U^{-1} = U^*$ and $V^{-1} = V^*$.

1. U is isometric.

$$\|Ux\|^2 = \langle Ux, Ux\rangle = \langle x, U^*Ux\rangle = \langle x, Ix\rangle = \|x\|^2$$

Therefore, $\|Ux\| = \|x\|$. In other words, U is isometric.

---

2. $\|U\| = 1$
We have, U is isometric. That is, $\|Ux\| = \|x\|$. Thus,

$$\|U\| = \sup_{\substack{x\in H\\ x\neq 0}} \frac{\|Ux\|}{\|x\|} = 1$$

---

3. U^{-1} is unitary
$$(U^{-1})^* = (U^*)^* = U = (U^{-1})^{-1}$$
If $U^{-1} = V$, then $V^* = V^{-1}$. Therefore, U^{-1} is unitary.

---

4. UV is unitary since
$$(UV)^* = V^*U^* = V^{-1}U^{-1} = (UV)^{-1}$$

---

5. U is normal since
$$UU^* = UU^{-1} = I = U^{-1}U = U^*U$$

---

6. T unitary iff T is isometric and surjective. (H complex)

Suppose T is isometric and surjective. Since T is isometric, we have $\|Tx\| = \|x\|$. Suppose $Tx = Ty$. Then,
$$\|Tx - Ty\| = \|T(x-y)\| = \|x-y\| = 0$$
Thus, $Tx = Ty \implies x = y$ Therefore, T is injective. Thus,T^{-1} exists.
$$\langle T^*Tx, x\rangle = \langle Tx, Tx\rangle = \|Tx\|^2 = \|x\|^2 = \langle x, x\rangle = \left\langle T^{-1}Tx, x\right\rangle$$
Therefore, $T^* = T^{-1}$. In other words, T is unitary.

Suppose T is unitary. Then T is isometric by part (a). We have, $T^* = T^{-1}$. Thus T is bijective, otherwise T^{-1} won't even exist.

□

Exercise §3.10

Definitions 14.4.3 (unitarily equivalence)**.** Let H be a Hilbert space. Let $S: H \to H$ and $T: H \to H$ be linear operators. Then S is unitarily equivalent to T if there exists a unitary operator $U: H \to H$ such that $S = UTU^{-1}$.

14.5 Fundamental Theorems for Banach Spaces

14.5.1 Zorn's Lemma

Definitions 14.5.1 (POSET)**.** A set M is a POSET if it satisfyies

1. $a \le a, \ \forall a \in M$
2. $a \le b, b \le a \implies a = b$
3. $a \le b, b \le c \implies a \le c$

Lemma 14.5.1 (Zorn)**.** *LEt M be a non-empty POSET. Suppose every chain in M has an upper bound. Then M has a maximal element.*

Applications : Zorn's Lemma

Remark. Every non-trivial vector space has a Hamel basis.

Proof. Let X be a nontrivial vector space. Let M be the set of all linearly independent subsets of X. Since X has nonzero elemnts, M is nonempty. We know that, M together with set inclusion in a POSET since subsets of linearly independent sets are also linearly independent.

Let C be a chain in M. Then the union of all subsets in C is a bound for C. Thus, every chain in M is bounded above. By Zorn's lemma, M has a maximal element, say B.

Let $Y = span\ B$. Suppose $Y \neq X$. Then there exists $z \in X$ such that $z \in Y$. Then $B \cup \{z\}$ is a linearly independent subset of X. This is a contradiction since B is maximal. Therefore, $Y = X$. Thus, any element $x \in X$ belongs to $span\ M$. In other words, B is a Hamel basis for X. □

Remark. Every nontrivial Hilbert space has a total, orthonormal set.

Proof. Let M be the set of all orthonormal subset of Hilbert space H. Set M is nonempty since H is nontrivial and for any $x \in H$, we have $\frac{x}{\|x\|} \in M$. Set M together with set inclustion is a POSET. And any chain C in M is bounded above by the union of all subsets in C. By Zorn's lemma, M has a maximal element, say F.

Suppose F is not total in H. Then there exists $z \in H$ such that $z \perp F$. Define $e = \frac{z}{\|z\|}$. Then $F \cup e$ is an orthonormal subset of H contradicting maximality of F. Therefore, F is total. Clearly, every Hilbert space has a total, orthonormal subset. □

14.5.2 Hahn-Banach Theorem

Definitions 14.5.2 (extension problem)**.** Let X be any set. Let Z be any subset of X and G be an entity defined on Z. The extension problem is the question whether there exists an extension of G from Z into X such that basic properties of G are preserved.

Definitions 14.5.3 (sublinear)**.** A functional p on a vector space X is sublinear if $p(x+y) \leq p(x) + p(y),\ \forall x, y \in X$.

Definitions 14.5.4 (positive homogenous)**.** A function p on a vector space X is positive homogenous if $p(\alpha x) = \alpha p(x),\ \forall \alpha \geq 0,\ \alpha \in \mathbb{R}$.

For example, Let X be a normed space. Then the norm on X is a positive homogenous, sublinear functional on X.

Theorem 14.5.2 (Hahn-Banach - Linear Functional)**.** *Let X be a real vector space. Let p be a sublinear functional on X. Let f be a linear functional defined on a subspace Z of X such that $f(x) \leq p(x),\ \forall x \in Z$. Then, f has a linear extension $\tilde{f}$ from Z into X satisfying $\tilde{f}(x) \leq p(x),\ \forall x \in X$.*

Proof. Let X be a real vector space and Z be a subspace of X. Let p be a sublinear functional on X. Let f be linear functional on Z such that $f(x) \leq p(x)\ \forall x \in Z$.

Step 1 : Construction of $\tilde{f}$, by Zorn's lemma
Let E be the set all linear extensions g of f such that $g(x) = f(x),\ \forall x \in Z$ and $g(x) \leq p(x),\ \forall x \in \mathscr{D}(g)$. Then E is nonempty, since $f \in E$. Define a partial order $\leq$ on E such that $g \leq h$ if $\mathscr{D}(g) \subset \mathscr{D}(h)$ and $g(x) = h(x),\ \forall x \in \mathscr{D}(g)$. Clearly E is a POSET.

Let C be a chain in E. Now we can construct, an upper bound for any such chain in E by considering the extension on the union of domains of all extension in C. Define $\hat{Y} = \bigcup\limits_{g\in C} \mathscr{D}(g)$. Define $\hat{g} : \hat{Y} \to \mathbb{R}$ such that $\hat{g}(x) = g(x),\ x \in \mathscr{D}(g)$. Clearly, every chain C in E is bounded above by such an extension $\hat{g}$. By Zorn's lemma, E has a maximal element, say $\tilde{f}$. Since $\tilde{f} \in E$, we have $f(x) \leq p(x),\ \forall x \in \mathscr{D}(f)$.

Step 2 : $\mathscr{D}(\tilde{f}) = X$
Suppose $\mathscr{D}(\tilde{f}) \neq X$. Then there exists $y_1 \in X$ such that $y_1 \notin \mathscr{D}(\tilde{f})$. Consider the subspace Y_1 of X spanned by $\mathscr{D}(\tilde{f}) \cup \{y_1\}$. Then any $x \in Y_1$ is of the form $x = y + \alpha y_1$ where $y \in \mathscr{D}(\tilde{f})$.

This representation is unique since $x = y + \alpha y_1 = \tilde{y} + \beta y_1 \implies y - \tilde{y} = (\beta - \alpha) y_1 \in \mathscr{D}(\tilde{f})$. We know that $y_1 \notin \mathscr{D}(\tilde{f})$, thus $\alpha = \beta$.

Define $g_1 : Y_1 \to \mathbb{R}$ such that $g_1(x) = \tilde{f}(x) + \alpha c$ where $c \in \mathbb{R}$ is fixed. Clearly g_1 is linear since $\tilde{f}$ is linear and

$$g_1(x+\tilde{x}) = g_1(y+\alpha y_1+\tilde{y}+\beta y_1) = \tilde{f}(y+\tilde{y})+c(\alpha+\beta) = g_1(x+\alpha y_1)+g_1(\tilde{x}+\beta y_1)$$

And $g_1 \leq \tilde{f}$ since $\mathscr{D}(\tilde{f}) \subset \mathscr{D}(g_1)$ and $g_1(x) = \tilde{f}(x),\ \forall x \in \mathscr{D}(\tilde{f})$. However, it remains to prove that $g_1(x) \leq p(x)$. We claim that $g_1(x) \leq p(x),\ \forall x \in \mathscr{D}(g_1)$.

Step 3 : Claim $g_1(x) \leq p(x),\ \forall x \in \mathscr{D}(g_1)$
Suppose $y, z \in \mathscr{D}(\tilde{f})$.

$$\begin{aligned} \tilde{f}(y) - \tilde{f}(z) = \tilde{f}(y-z) &\leq p(y-z) \\ &\leq p(y + y_1 - z - y_1) \\ &\leq p(y + y_1) + p(-z - y_1) \end{aligned}$$

Rearranging terms, we get

$$-\tilde{f}(z) - p(-z - y_1) \leq p(y + y_1) - \tilde{f}(y)$$

Applying supremum over z on the left and infimum over y on the right, we get $m_0 \leq m_1$ where $m_0 = \sup\limits_{z\in\mathscr{D}(\tilde{f})} -\tilde{f}(z) - p(-z - y_1)$ and
$m_1 = \inf\limits_{y\in\mathscr{D}(\tilde{f})} p(y + y_1) - \tilde{f}(y)$.

Let $m_0 \leq c \leq m_1$. Then,

$$-p(-z-y_1) - \tilde{f}(z) \leq c \leq p(y+y_1) - \tilde{f}(y), \quad \forall y, z \in \mathscr{D}(\tilde{f})$$

Case 1 : $\alpha < 0$
Consider $z = \alpha^{-1}y$,

$$-p(-y_1 - \frac{y}{\alpha}) - \tilde{f}(\frac{y}{\alpha}) \leq m_0 \leq c$$

Multiplying with $-\alpha$

$$\begin{aligned} \alpha p(-y_1 - \frac{y}{\alpha}) + \tilde{f}(y) &\leq -\alpha c \\ g_1(x) = \tilde{f}(y) + \alpha c &\leq -\alpha p(-y_1 - \frac{y}{\alpha}) \\ &\leq p(\alpha y_1 + y) = p(x) \end{aligned}$$

Case 2 : $\alpha > 0$
Consider $y = \alpha^{-1}y$,

$$c \leq m_1 \leq p(\frac{y}{\alpha} + y_1) - \tilde{f}(\frac{y}{\alpha})$$

Multiplying with α

$$\begin{aligned} \alpha c &\leq \alpha p(\frac{y}{\alpha} + y_1) - \alpha \tilde{f}(\frac{y}{\alpha}) \\ &\leq p(y + \alpha y_1) - \tilde{f}(y) \\ g_1(x) = \tilde{f}(y) + \alpha c &\leq p(y + \alpha y_1) = p(x) \end{aligned}$$

Case 3 : $\alpha = 0$
Suppose $\alpha = 0$, then $x = y + 0y_1 \in \mathscr{D}(g_1)$. Thus, $g_1(x) = \tilde{f}(x) \leq p(x)$.

Thus, the claim is true. And we have a proper extension g_1 of $\tilde{f}$ which is a contradiction as $\tilde{f}$ is maximal in E. Thus, there does not exist such a nonzero vector $y_1 \in X - \mathscr{D}(f)$. Therefore, $\mathscr{D}(\tilde{f}) = X$. □

14.5.3 Hahn-Banach Theorem for complex vector spaces and normed spaces

Theorem 14.5.3 (Hahn-Banach : generalised)**.** *Let X be a real/complex vector space. Let p be a subadditive, real-valued functional on X and $p(\alpha x) = |\alpha| p(x)$, $\forall \alpha \in K$. Let f be a linear functional defined on a subspace Z of X such that $|f(x)| \leq p(x)$. Then, f has a linear extension $\tilde{f}$ from Z into X satisfying $|\tilde{f}(x)| \leq p(x)$, $\forall x \in X$.*

Proof. **Case 1 : Real Vector Space**
By Hahn-Banach theorem, we know that there exists an extension $\tilde{f}$ such that $\tilde{f} \leq p$, in the case of real vector space X.

Case 2 : Complex Vector Space
Let X be a complex vector space. Then any functional f on Z is of the form $f = f_1 + if_2$ where both f_1, f_2 are real valued functionals . And f_1, f_2 are

linear functionals since f is a linear functional on X. Clearly, $f_1(x) \leq p(x)$ and $f_2(x) \leq p(x)$ since the real/imaginary parts are bounded above by the absolute value of f and f itself if bounded. Therefore, f_1, f_2 are bounded funtionals.

By Hahn-Banach theorem $f_1(x)$ has an extension $\tilde{f}_1(x)$. We have,

$$f(ix) = f_1(ix) + if_2(ix) = -f_2(x) + if_1(x)$$

Thus, $f_2(x) = -f_1(ix)$. Define

$$\tilde{f}(x) = \tilde{f}_1(x) - i\tilde{f}_1(ix)$$

Then, $\tilde{f}$ is an extension of f. It remains to prove that $\tilde{f}$ is linear and $\tilde{f}(x) \leq p(x),\ \forall x \in X$.

Clearly, $\tilde{f}$ is linear, since

$$\begin{aligned}
\tilde{f}((a+ib)x) &= \tilde{f}_1(ax+ibx) - i\tilde{f}_1(iax - bx) \\
&= a\tilde{f}_1(x) + b\tilde{f}_1(ix) - ia\tilde{f}_1(ix) + ib\tilde{f}_1(x) \\
&= a\left(\tilde{f}_1(x) - i\tilde{f}_1(ix)\right) + ib\left(\tilde{f}_1(x) - i\tilde{f}_1(ix)\right) \\
&= (a+ib)\left(\tilde{f}_1(x) - \tilde{f}_1(ix)\right) \\
&= (a+ib)\tilde{f}(x)
\end{aligned}$$

Let $x \in X$ such that $\tilde{f}(x) = 0$. Suppose $p(x) < 0$, then $p(-x) < 0$ since $p(-x) = |-1|p(x) = p(x)$. We know that p is subadditive. That is, $p(x+y) \leq p(x) + p(y)$. Thus, $p(0) = p(x-x) \leq p(x) + p(-x) = 2p(x)$. Therefore,

$$p(x) = p(x+0) \leq p(x) + p(0) \leq p(x) + 2p(x) = 3p(x)$$

which is a contradiction thus $p(x) \geq 0$. Therefore, if $\tilde{f}(x) = 0$, then $\tilde{f}(x) \leq p(x)$.

Let $x \in X$ such that $\tilde{f}(x) \neq 0$. Then $\tilde{f}(x) = |\tilde{f}(x)|e^{i\theta}$. Thus, $|\tilde{f}(x)| = e^{-i\theta}\tilde{f}(x) = \tilde{f}(xe^{-i\theta})$. Since $|\tilde{f}(x)|$ is real, $\tilde{f}(xe^{-i\theta})$ is also real. By Hahn-Banach theorem on linear functionals on real vector spaces, $\tilde{f}_1(x) \leq p(x),\ \forall x \in X$. Therefore,

$$|\tilde{f}(x)| = \tilde{f}(xe^{-i\theta}) = \tilde{f}_1\left(xe^{-i\theta}\right) \leq p\left(xe^{-i\theta}\right) = |e^{-i\theta}|p(x) = p(x)$$

Thus, if $\tilde{f}(x) \neq 0$, then $\tilde{f}(x) \leq p(x)$. Therefore, $\tilde{f}$ is a linear extension of f such that $\tilde{f}(x) \leq p(x),\ \forall x \in X$. □

Theorem 14.5.4 (Hahn-Banach : Normed space)**.** *Let f be a bounded, linear functional on a subspace Z of normed space X. Then there exists a bounded, linear functional $\tilde{f}$ on X which is an extension to f on Z and has the same norm, $\|\tilde{f}\| = \|f\|$ where*

$$\|\tilde{f}\| = \sup_{\substack{x \in X \\ \|x\|=1}} |\tilde{f}(x)| \qquad \|f\| = \sup_{\substack{x \in Z \\ \|x\|=1}} |f(x)|$$

Proof. Without loss of generality suppose that Z is non-trivial. If $Z = \{0\}$, then $f(x) = f(0) = 0$ since f is linear. Thus, f has an extension $\tilde{f}$ defined by $\tilde{f}(x) = 0, \ \forall x \in X$ and $\|f\| = \|\tilde{f}\| = 0$.

Suppose Z is non-trivial. Then $|f(x)| \leq \|f\| \ \|x\|$ since f is bounded†[17]. Define functional p where $p(x) = \|f\| \ \|x\|, \ \forall x \in X$. Then,

$$p(x+y) = \|f\| \ \|x+y\| \leq \|f\|(\|x\| + \|y\|) = p(x) + p(y)$$

And,

$$p(\alpha x) = \|f\| \ \|\alpha x\| = |\alpha| \ \|f\| \ \|x\| = |\alpha| p(x)$$

Clearly, p is a sublinear functional on X such that $f(x) \leq p(x), \ \forall x \in Z$.

Given any functional f on Z, there exists a sublinear functional p such that $f(x) \leq p(x)$. Then, by Hahn-Banach theorem, there exists a bounded, linear functional $\tilde{f} : X \to K$ such that $\tilde{f}(x) \leq p(x), \ \forall x \in X$ and $\tilde{f}$ is an extension of f. That is, $\tilde{f}(x) = f(x), \ \forall x \in Z$.

Since $\tilde{f}(x) \leq p(x), \ \forall x \in X$. We have,

$$\|\tilde{f}\| = \sup_{x \neq 0} \frac{|\tilde{f}(x)|}{\|x\|} = \sup_{x \neq 0} \frac{|p(x)|}{\|x\|} = \|f\|$$

Therefore, any bounded, linear functional on any subspace of any normed space has a bounded, linear extension preserving norm. ☐

Theorem 14.5.5. *Let X be a normed space. Let $x_0 \in X$. Then there exists a bounded, linear functional $\tilde{f}$ on X such that $\|\tilde{f}\| = 1$ and $\tilde{f}(x_0) = \|x_0\|$.*

Proof. Let $x_0 \in X$ and $x_0 \neq 0$. Let Z be the subspace spanned by x_0. That is, $Z = \{\alpha x_0 : \alpha \in K\}$. Define $f : Z \to K$ where $f(x) = f(\alpha x_0) = \alpha \|x_0\|$. Then $|f(x)| = |\alpha| \ \|x_0\| = \|x\|$. Clearly, f is bounded, linear functional with $\|f\| = 1$. By Hahn-Banach therorem for normed spaces f has an extension $\tilde{f}$ on X such that $\|\tilde{f}\| = \|f\|$ and $\tilde{f}(x_0) = f(x_0) = \|x_0\|$. ☐

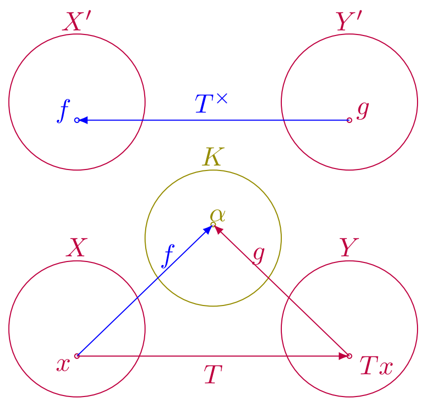
Corollary 14.5.5.1. *Let X be a normed space. Let $x \in X$.*

$$\|x\| = \sup_{\substack{f \in X' \\ f \neq 0}} \frac{|f(x)|}{\|f\|}$$

If $f(x_0) = 0, \ \forall f \in X'$, then $x_0 = 0$.

Proof. Suppose $x \in X$ such that $f(x) = 0$ for every bounded, linear functional f on X. Considering each f as bounded, linear functional on the subspaces spanned by x, we get bounded linear extensions $\tilde{f}$ on X by Hahn-Banach theorem for normed spaces.

[17] Author uses $\|f\|_Z$ to emphasis that the functional is defined only in Z. Proper notation would be $\|f\|_{B(Z,K)}$ and $\|\tilde{f}\|_{B(X,K)}$ instead of $\|f\|_Z$ and $\|\tilde{f}\|_X$



$\forall g \in Y', \ \exists f \in X'$, such that $g(Tx) = (T^{\times}g)(x) = f(x) = \alpha \in K$

Figure 14.2: Adjoint-Operator

Then,

$$\sup_{f \neq 0} \frac{|f(x)|}{\|f\|} \geq \sup_{f \neq 0} \frac{|\tilde{f}(x)|}{\|\tilde{f}\|} = \|x\|$$

Also,

$$\sup_{f \neq 0} \frac{|f(x)|}{\|f\|} \leq \sup_{f \neq 0} \frac{\|f\| \ \|x\|}{\|f\|} = \|x\|$$

Therefore,

$$\sup_{f \neq 0} \frac{|f(x)|}{\|f\|} = \|x\|$$

□

14.5.5 Adjoint Operator

Definitions 14.5.5 (adjoint-operator)**.** Let X, Y be normed spaces over the same field K. Let X', Y' be the dual spaces of X, Y. Let $T : X \to Y$ be a bounded,linear operator. Then the adjoint-operator is $T^{\times} : Y' \to X'$ defined by $f(x) = (T^{\times}g)(x) = g(Tx)$.

Theorem 14.5.6 (norm)**.** *Let $T^{\times}$ be the adjoint operator of a bounded, linear operator T. Then $T^{\times}$ is a bounded, linear operator. And $\|T^{\times}\| = \|T\|$.*

Proof. Let $g_1, g_2 \in Y'$ and $\alpha, \beta \in K$. Then,

$$\begin{aligned}
\left(T^{\times}(\alpha g_1 + \beta g_2)\right)(x) &= (\alpha g_1 + \beta g_2)Tx \\
&= \alpha g_1(Tx) + \beta g_2(Tx) \\
&= \alpha (T^{\times} g_1)(x) + \beta (T^{\times} g_2)(x)
\end{aligned}$$

Clearly, $T^{\times}$ is linear.

We have, $f = T^{\times} g$.

$$\begin{aligned} f(\alpha x_1 + \beta x_2) &= (T^{\times} g)(\alpha x_1 + \beta x_2) \\ &= g(T(\alpha x_1 + \beta x_2)) \\ &= g(\alpha T x_1 + \beta T x_2) \\ &= \alpha g(T x_1) + \beta g(T x_2) \\ &= \alpha (T \times g) x_1 + \beta (T^{\times} g) x_2 \\ &= \alpha f(x_1) + \beta f(x_2) \end{aligned}$$

Thus, f is linear. We have $|g(Tx)| \leq \|g\| \ \|T\| \ \|x\|$, since functional g and operator T are bounded. Thus,

$$\|T^{\times} g\| = \|f\| = \sup_{x \neq 0} \frac{|f(x)|}{\|x\|} = \sup_{x \neq 0} \frac{|g(Tx)|}{\|x\|} \leq \sup_{x \neq 0} \frac{\|g\| \ \|T\| \ \|x\|}{\|x\|} \leq \|g\| \ \|T\|$$

Taking supermum over all $g \in Y'$, we get

$$\|T^{\times}\| \leq \sup_{g \neq 0} \frac{\|T^{\times} g\|}{\|g\|} = \sup_{g \neq 0} \frac{\|g\| \ \|T\|}{\|g\|} = \|T\|$$

We know that for any non-zero $x_0 \in X$ there exists a $g_0 \in Y'$ such that $\|g_0\| = 1$ and $g_0(Tx_0) = \|Tx_0\|$. Let $f_0 = T^{\times} g_0$. Then,

$$\|Tx_0\| = g_0(Tx_0) = f_0(x_0) \leq \|f_0\| \ \|x_0\| = \|T^{\times} g_0\| \ \|x_0\| \leq \|T^{\times}\| \ \|g_0\| \ \|x_0\|$$

Thus, $\|Tx_0\| \leq \|T^{\times}\| \ \|x_0\|$. Taking supremum over x_0 we get, $\|T\| \leq \|T^{\times}\|$.

We have $\|Tx_0\| \leq \|T\| \ \|x_0\|$. And $\|T\| = c$ is the smallest real number satisfying $\|Tx_0\| \leq c\|x_0\|$, $\forall x_0 \in X$. Thus, $\|T^{\times}\| \geq c$. Therefore,

$$\|T^{\times}\| \geq \|T\|$$

Therefore, $\|T^{\times}\| = \|T\|$. □

Remark (Properties of adjoint-operator). Let $S : Y \to Z$ and $T : X \to Y$ be bounded, linear operators. Let $\alpha \in K$. Then,

1. If T_E is the matrix representation of T, then transpose of T_E is the matrix representation of $T^{\times}$.

   Proof. Let T_E be the matrix representation of linear operator $T : \mathbb{R}^n \to \mathbb{R}^n$ with respect to standard basis $E = \{e_1, \ e_2, \ \ldots, \ e_n\}$. Let $x = (\xi_j)$ and $y = (\eta_j)$. Then $y = Tx$ implies $\eta_j = \sum_{k=1}^{n} \tau_{j,k} \xi_k$. Define $T_E = (\tau_{j,k})$. Then $y = T_E x$. We know that $(\mathbb{R}^n)' = \mathbb{R}^n$. Let g be a bounded, linear operator on $\mathbb{R}^n$. Then $g \in \mathbb{R}^n$ and $g = \sum_{k=1}^{n} \alpha_k f_k$ where $F = \{f_1, \ f_2, \ \ldots, \ f_n\}$ is the dual basis of E. That is, $f_k(e_j) = \delta_{k,j}$.

   $$g(y) = g(T_E x) = \sum_{j=1}^{n} \alpha_j \eta_j = \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \tau_{j,k} \xi_k$$

Thus,

$$g(T_E x) = \sum_{k=1}^{n} \beta_k \xi_k \text{ where } \beta_k = \sum_{j=1}^{n} \tau_{j,k}\alpha_j$$

Define $f : X \to K$ where $f(x) = g(Tx)$. Then,

$$f(x) = g(T_E x) = \sum_{k=1}^{n} \beta_k \xi_k$$

Suppose $f = T_E^{\times} g$. Then matrix of $T_E^{\times}$ is given by $(\tau_{j,k})$ since $\beta_k = \sum_{j=1}^{n} \tau_{j,k}\alpha_j$. Thus, matrix of $T^{\times}$ is the transpose of that of T_E. □

2. $(S+T)^{\times} = S^{\times} + T^{\times}$

   Proof.

$$\begin{aligned}((S+T)^{\times} g)(x) &= g((S+T)x) \\ &= g(Sx + Tx) \\ &= g(Sx) + g(Tx) \\ &= (S^{\times} g)(x) + (T^{\times} g)(x)\end{aligned}$$

□

3. $(\alpha T)^{\times} = \alpha T^{\times}$

   Proof.

$$\begin{aligned}((\alpha T)^{\times} g)(x) &= g((\alpha T)x) \\ &= g(\alpha T x) \\ &= \alpha g(Tx) \\ &= \alpha (T^{\times} g)(x)\end{aligned}$$

□

4. $(T^{\times})^{-1} = (T^{-1})^{\times}$
   Suppose $T : X \to Y$ is invertible and $T^{-1} : Y \to X$. Then $(T^{\times})^{-1}$ exists and $(T^{\times})^{-1} = (T^{-1})^{\times}$.

   Proof. We have $T \in B(X, Y)$ and $T^{\times} \in B(Y', X')$. Define $S = T^{-1}$. Then $S \in B(Y, X)$. Thus $S^{\times}$ exists, since S is a bounded, linear operator. And $S^{\times} \in B(X', Y')$.

   Let $g \in Y'$. Then,

$$g(y) = g(TSy) = (T^{\times} g)(Sy) = (S^{\times} T^{\times} g)(y), \ \forall g \in Y'$$

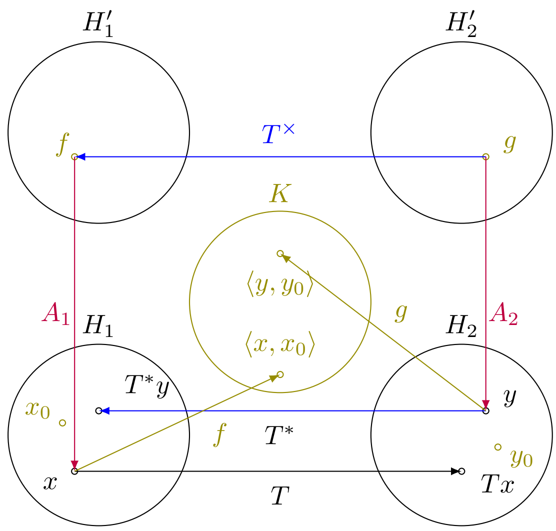


Figure 14.3: Adjoint-Operator vs Hilbert-Adjoint-Operator

Clearly, $T^\times S^\times$ is an identity operator on Y'.

Let $f \in X'$. Then,

$$f(x) = f(STx) = (S^\times f)(Tx) = (T^\times S^\times f)(x), \ \forall f \in X'$$

Clearly, $S^\times T^\times$ is an identity operator on X'.

Therefore, $S^\times = (T^{-1})^\times = (T^\times)^{-1}$. $\square$

Remark ($T^\times$ vs T^*). Let $T : H_1 \to H_2$. Let $T^\times$ be the adjoint-operator of T. That is, $T^\times : H'_2 \to H'_1$. Let $T^* : H_2 \to H_1$ be the Hilbert-adjoint operator of T. That is, $\langle Tx, y \rangle = \langle x, T^*y \rangle$. Then,

$$T^* = A_1 T^\times A_2^{-1}$$

where $A_1 : H'_1 \to H_1$ defined by $A_1(f) = x_0$ such that $f(x) = \langle x, x_0 \rangle$ and $A_2 : H'_2 \to H_2$ defined by $A_2(g) = y_0$ such that $g(y) = \langle y, y_0 \rangle$.

Proof. Let H_1, H_2 be normed spaces. Let $T : H_1 \to H_2$ be a bounded, linear operator. Then there exists unique adjoint operator $T^\times : H'_2 \to H'_1$ defined by $(T^\times g)(x) = g(Tx)$, $\forall g \in Y'$, $\forall x \in H_1$. And there exists unique Hilbert adjoint operator $T^* : H_2 \to H_1$ defined by $\langle Tx, y \rangle = \langle x, T^*y \rangle$, $\forall x \in H_1$, $\forall y \in H_2$.

Let $f \in H'_1$. By Riesz theorem, $\forall f \in H'_1$ there exists unique $x_0 \in H_1$ such that $f(x) = \langle x, x_0 \rangle$. Define $A_1 : H'_1 \to H_1$ defined by $A_1(f) = x_0$ where $f(x) = \langle x, x_0 \rangle$, $\forall x \in H_1$. Clearly, A_1 is bijective by its construction. Let $g \in H'_2$. By Riesz theorem, $\forall g \in H'_2$, there exists unique $y_0 \in H_2$ such that $g(y) = \langle y, y_0 \rangle$. Define $A_2 : H'_2 \to H_2$ defined by $A_2(g) = y_0$ where

$g(y) = \langle y, y_0 \rangle$, $\forall y \in H_2$. Clearly, A_2 is bijective by its construction.

Let $A_1(f_1) = x_1$ and $A_1(f_2) = x_2$. Then, we have $f_1(x) = \langle x, x_1 \rangle$ and $f_2(x) = \langle x, x_2 \rangle$. Thus,

$$\begin{aligned}
(\alpha f_1 + \beta f_2))(x) &= \alpha f_1(x) + \beta f_2(x) \\
&= \alpha \langle x, x_1 \rangle + \beta \langle x, x_2 \rangle \\
&= \langle x, \overline{\alpha} x_1 \rangle + \langle x, \overline{\beta} x_2 \rangle \\
&= \langle x, \overline{\alpha} x_1 + \overline{\beta} x_2 \rangle \\
A_1(\alpha f_1 + \beta f_2) &= \overline{\alpha} x_1 + \overline{\beta} x_2 \\
&= \overline{\alpha} A_1(f_1) + \overline{\beta} A_1(f_2)
\end{aligned}$$

Thus, A_1 is conjugate linear. Similarly, A_2 is also conjugate linear.

Let $S = A_1 T^\times A_2^{-1}$. Let $A_1(f) = x_0$ and $A_2(g) = y_0$. Then $S : H_2 \to H_1$ defined by $S(y_0) = (A_1 T^\times A_2^{-1})(y_0) = (A_1 T^\times)(g) = A_1(f) = x_0$. And S is linear since A_1 and A_2^{-1} are conjugate linear. Then,

$$\langle Tx, y_0 \rangle = g(Tx) = f(x) = \langle x, x_0 \rangle = \langle x, Sy_0 \rangle$$

But, we have, $\langle Tx, y_0 \rangle = \langle x, T^* y_0 \rangle$. And we know that, $w_1 = w_2$ if $\langle v, w_1 \rangle = \langle v, w_2 \rangle , \forall v$. Thus, $S = T^* = A_1 T^\times A_2^{-1}$. □

Remark. Comparison

1. Matrix of adjoint operator is transpose. But, matrix of Hilbert adjoint operator is complex conjugate transpose.
2. Adjoint operator is linear, but Hilbert-adjoint operator is conjugate linear.

14.5.6 Exercises

[Kreyszig, 2014, §4.5 Exercise 6]

Remark.

$$(T^\times)^n = (T^n)^\times$$

Proof. Let $T : X \to X$ be a bounded, linear operator.

$$g(T^2 x) = g(T(Tx)) = (T^\times g)(Tx) = T^\times T^\times g(x) = (T^\times)^2 g(x)$$

Suppose $(T^\times)^{k-1} = (T^{k-1})^\times$ for some $k \in \mathbb{N}$. Then,

$$g(T^k x) = g(T(T^{k-1}x)) = (T^\times g)(T^{k-1}x) = T^\times (T^\times)^{k-1} g(x) = (T^\times)^k g(x)$$

Therefore, by finite mathematical induction the result is true. □

Definitions 14.5.6 (Annihilator in a normed space). Let X be a normed space and M be a subset of X. Then the annihilator of M is given by,

$$M^a = \{f \in X' : f(M) = \{0\}\}$$

Definitions 14.5.7 (Annihilator in a dual space)**.** Let X be a normed space and X' be its dual space. Let B be a subset of X'. Then the annihilator of B is given by,

$$^{a}B = \{x \in X : f(x) = 0, \ \forall f \in B\}$$

Remark.

$$\mathscr{R}(T) \subset \ ^{a}\mathcal{N}(T^{\times})$$

Proof. Let $T : X \to Y$ be a bounded, linear operator. Let $f \in \mathcal{N}(T^{\times})$. Then $T^{\times} f = 0$. Let $y = Tx \in \mathscr{R}(T)$. Then,

$$f(y) = f(Tx) = (T^{\times} f)(x) = 0, \ \forall y \in \mathscr{R}(T)$$

Thus, $y \in \mathscr{R}(T) \implies y \in \ ^{a}\mathcal{N}(T^{\times})$. □

Subject 15

ME010305 Optimization Technique

Semester IV

Subject 16

ME010401 Spectral Theory

Subject 17

ME010402 Analytic Number Theory

Subject 18

ME800401 Differential Geometry

18.1 Graphs and Level Set

Definitions 18.1.1. Let function $f : U \to \mathbb{R}$ where $U \subset \mathbb{R}^{n+1}$. Let c be a real number. Then the **Level set** of f at height c is the set of all points in U with image c.

$$f^{-1}(c) = \{(x_1, x_2, \ldots, x_{n+1}) \in U : f(x_1, x_2, \ldots, x_{n+1}) = c\} \tag{18.1}$$

Definitions 18.1.2. Let function $f : U \to \mathbb{R}$ where $U \subset \mathbb{R}^{n+1}$. Then,

$$graph(f) = \{(x_1, x_2, \ldots, x_{n+2}) \in \mathbb{R}^{n+2} : f(x_1, x_2, \ldots, x_{n+1}) = x_{n+2}\} \tag{18.2}$$

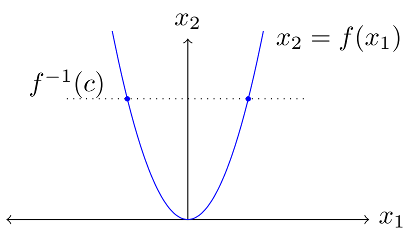


Figure 18.1: Graph of $f(x_1) = x_1^2$ and Level set $f^{-1}(c)$

18.2 Vector Fields

Definitions 18.2.1. A vector $\mathbf{v}$ at a point $p \in \mathbb{R}^{n+1}$ is a pair $\mathbf{v} = (p, v)$ where $v \in \mathbb{R}^{n+1}$.

vector addition $\mathbf{v} + \mathbf{w} = (p, v) + (p, w) = (p, v + w)$.

scalar multiplication Let $c \in \mathbb{R}$, then $c\mathbf{v} = c(p, v) = (p, cv)$.

dot product $\mathbf{v} \cdot \mathbf{w} = (p, v) \cdot (p, w) = v \cdot w$

cross product $\mathbf{v} \times \mathbf{w} = (p, v) \times (p, w) = (p, v \times w)$

Remark. Angle θ between $\mathbf{v}$ and $\mathbf{w}$ is given by,

$$\cos\theta = \mathbf{v} \cdot \mathbf{w} = (p, v) \cdot (p, w) = v.w \tag{18.3}$$

And the length of a vector $\mathbf{v}$ is given by,

$$\|\mathbf{v}\| = \mathbf{v} \cdot \mathbf{v} = (p, v) \cdot (p, v) = v \cdot v = \|v\| \tag{18.4}$$

Remark. Let $c \in \mathbb{R}$ and $p \in \mathbb{R}^{n+1}$. Let $\mathbf{v}, \mathbf{w}$ be two vectors at p. That is, $\mathbf{v} = (p, v)$ and $\mathbf{w} = (p, w)$ for some $v, w \in \mathbb{R}^{n+1}$. Then the set of all vectors at p is a vector space with vector addition $\mathbf{v}+\mathbf{w} = (p, v+w)$ and scalar multiplication $c\mathbf{v} = (p, cv)$. This vector space is denoted by $\mathbb{R}^{n+1}_p$.

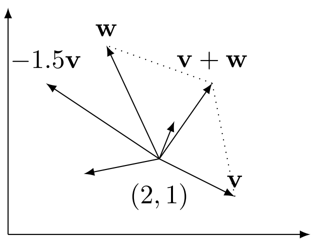


Figure 18.2: The vector space of all vectors at $(2, 1)$, $\mathbb{R}^2_{(2,1)}$

Definitions 18.2.2. The vector field $\mathbf{X}$ on $\mathbb{R}^{n+1}$ is a function which assigns to each point of $\mathbb{R}^{n+1}$ a vector at that point. That is, $\mathbf{X}(p) = (p, X(p))$.

For example, $\mathbf{X}(p) = (p, X(p))$ where the associated function of the vector field, $X : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $X(p) = (1, 2)$ assigns a constant vector $(1, 2)$ at every vector in $\mathbb{R}^2$.

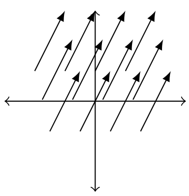


Figure 18.3: Vector field with associated function $X(p) = (1, 2)$

Definitions 18.2.3 (smooth)**.** A function $f : \mathbb{R} \to \mathbb{R}$ is smooth if its partial derivatives of all orders exists and are continuous. A function $f : \mathbb{R}^{n+1} \to \mathbb{R}$ is smooth if its component functions $f = (f_1, f_2, \ldots, f_{n+1})$ are smooth. A vector field $\mathbf{X}$ is smooth if the associated function $X(p)$ is smooth.

Definitions 18.2.4. Let $f : \mathbb{R}^{n+1} \to \mathbb{R}$. Then the gradient of f at p is,

$$\nabla f(p) = \left(p, \frac{\partial f}{\partial x_1}(p), \frac{\partial f}{\partial x_2}(p), \ldots, \frac{\partial f}{\partial x_{n+1}}(p)\right) \tag{18.5}$$

Remark. If f is a smooth function, then the gradient of f at p is a smooth vector field.

For example, $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x_1, x_2) = 2x_1x_2$ is a smooth function. We have, $\frac{\partial f}{\partial x_1} = 2x_2$ and $\frac{\partial f}{\partial x_2} = 2x_1$. And gradient of f at (x_1, x_2) is $(x_1, x_2, 2x_2, 2x_1)$. That is, $(2x_2, 2x_1)$ at (x_1, x_2).

Calculations :

p	(x_1, x_2)	$(0,0)$	$(1,0)$	$(0,1)$	$(-1,0)$	$(0,-1)$
$X(p)$	$(2x_2, 2x_1)$	$(0,0)$	$(0,2)$	$(2,0)$	$(0,-2)$	$(-2,0)$

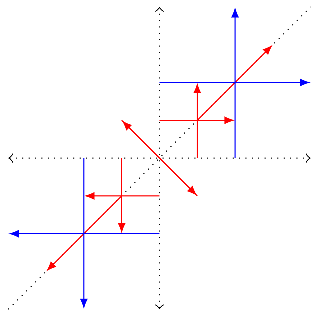


Figure 18.4: The gradient of $f(x_1, x_2) = 2x_1x_2$

Definitions 18.2.5. A parameterised curve is a function, $\alpha : I \to \mathbb{R}^{n+1}$ where I is some open interval in $\mathbb{R}$. The velocity vector of a parameterised curve $\alpha : I \to \mathbb{R}^{n+1}$ at a point $\alpha(t)$ is the tangent to the curve at that point.

$$\dot{\alpha}(t) = \left(\alpha(t), \frac{d\alpha}{dt}(t)\right) \tag{18.6}$$

For example, $\alpha : I \to \mathbb{R}^2$ defined by $\alpha(t) = (2t, t^2)$ is a parameterised curve. We have, $\frac{d\alpha}{dt} = (\frac{dx_1}{dt}(t), \frac{dx_2}{dt}(t)) = (2, 2t)$ where $\alpha(t) = (x_1(t), x_2(t))$. The velocity vector at $t = 3$ is $\dot{\alpha}(3) = (\alpha(t), \frac{d\alpha}{dt}) = (6, 9, 2, 6)$.

Definitions 18.2.6. Let $\mathbf{X}$ be a vector field and let U be an open subet of $\mathbb{R}^{n+1}$. An integral curve α on U is a parameterised curve, $\alpha : I \to \mathbb{R}^{n+1}$ such that for each $\alpha(t) = p \in U$, the velocity vector $\dot{\alpha}(t)$ is the associated vector $\mathbf{X}(p)$ of the vector field $\mathbf{X}$ at that point. Thus, for each $t \in I$, $\dot{\alpha}(t) = \mathbf{X}(\alpha(t))$.

$$\left(\alpha(t), \frac{d\alpha}{dt}(t)\right) = (\alpha(t), X(\alpha(t))) \tag{18.7}$$

Let $X(p) = (X_1(p), X_2(p), \ldots, X_{n+1}(p))$ and $\alpha(t) = (x_1(t), x_2(t), \ldots, x_{n+1}(t))$. Then, comparing components of the vector at $\alpha(t)$ we get the following system of equations,

$$\frac{dx_j}{dt}(t) = X_j(\alpha(t)),\ j = 1, 2, \ldots, (n+1) \tag{18.8}$$

For example, Consider $\alpha : (2, 3) \to \mathbb{R}^2$ defined by $\alpha(t) = (t, t^2)$. Then α is a parameterised curve in vector field, $\mathbf{X}$ which has the associated function $X(x_1, x_2) = (1, 2x_1)$. Then, $\mathbf{X}(x_1, x_2) = (x_1, x_2, 1, 2x_1)$. And

$$\dot{\alpha}(t) = \left(\alpha(t), \frac{d\alpha}{dt}(t)\right) = \left(x_1(t), x_2(t), \frac{dx_1}{dt}(t), \frac{dx_2}{dt}(t)\right) = (t, t^2, 1, 2t)$$

Clearly, α is an integral curve of $\mathbf{X}$ as $\dot{\alpha}(t) = X(\alpha(t))$ for every $t \in (2,3)$.

Calculations:

p	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$	$(-1,0)$	$(0,-1)$	$(-1,1)$	$(1,-1)$	$(-1,-1)$
$X(p)$	$(1,0)$	$(2,2)$	$(1,1)$	$(2,3)$	$(0,-2)$	$(1,1)$	$(0,-1)$	$(2,1)$	$(0,-3)$

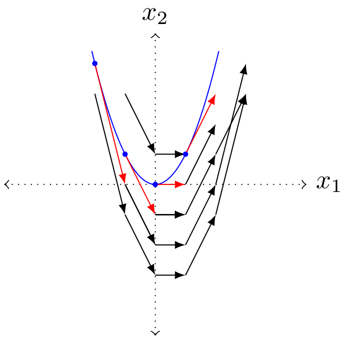


Figure 18.5: Integral Curve $\alpha(t) = (t, t^2)$ in $\mathbf{X}$ with $X(x_1, x_2) = (1, 2x_1)$

Theorem 18.2.1. *Let* $\mathbf{X}$ *be a smooth vector field on an open set* $U \subset \mathbb{R}^{n+1}$ *and let* $p \in U$*. Then there exists an open interval* I *containing* 0 *and an integral curve* $\alpha : I \to U$ *such that*

1. $\alpha(0) = p$
2. *If* $\beta : \tilde{I} \to U$ *is any other integral curve with* $\beta(0) = p$*, then* $\tilde{I} \subset I$ *and* $\beta(t) = \alpha(t)$*, for all* $t \in \tilde{I}$*.*

Proof. Let $\mathbf{X}$ be a smooth vector field. Suppose α be an integral curve in $\mathbf{X}$. Then, $\dot{\alpha}(t) = \mathbf{X}(\alpha(t))$. Let $x_j(t)$ be the components of $\alpha(t)$ and $X_j(p)$ be the components of $X(p)$.

$$\begin{aligned} \dot{\alpha}(t) &= \left(\alpha(t), \frac{d\alpha}{dt}(t)\right) \\ &= \left(x_1(t), \ldots, x_{n+1}(t), \frac{dx_1}{dt}(t), \ldots, \frac{dx_{n+1}}{dt}(t)\right) \\ \mathbf{X}(\alpha(t)) &= (\alpha(t), X(\alpha(t))) \\ &= (x_1(t), \ldots, x_{n+1}(t), X_1(\alpha(t)), \ldots, X_{n+1}(\alpha(t))) \end{aligned}$$

Thus, we a system of $n+1$ first order differential equations in $n+1$ unknowns satisfying the initial condition $\alpha(0) = p$.

$$\begin{aligned} \frac{dx_1}{dt}(t) &= X_1(\alpha(t)) \\ \frac{dx_2}{dt}(t) &= X_2(\alpha(t)) \\ &\vdots \\ \frac{dx_{n+1}}{dt}(t) &= X_{n+1}(\alpha(t)) \end{aligned}$$

$\square$

By the theorem on solution of systems of first order ordinary differential equations, there exists an interval I containing 0 and a solution — a family of functions $\{x_1(t), x_2(t), \dots, x_{n+1}(t)\}$ satisfying the above system of equations satisfying the initial condition $\alpha(0) = p$.

Define $\alpha : I \to U$ using the component functions of α as x_js in the above solution. Then, we have a integral curve of the vector field $\mathbf{X}$ satifying the initial condition $\alpha(0) = p$.

Let $\beta : \tilde{I} \to U$ be another integral curve with $\beta(0) = p$. Then by the uniqueness of the solution for the system of first order ordinary differential equations with an initial condition, $\beta(t) = \alpha(t)$ for every $t \in I \cup \tilde{I}$.

Let $\{\beta_1, \beta_2, \dots\}$ be the family of integral curves with $\beta_j : I_j \to U$ satisfying $\beta_j(0) = p$. Consider $I = \bigcup\limits_{j \in \mathbb{N}} I_j$.

Define $\alpha : I \to U$ by $\alpha(t) = \beta_j(t)$ where $t \in I_j$ for some $j \in \mathbb{N}$. Then α is well-defined and is a maximal integral curve in $\mathbf{X}$ such that $\alpha(0) = p$.

Definitions 18.2.7. A smooth vector field $\mathbf{X}$ on $U \subset \mathbb{R}^{n+1}$ is **complete** if for every $p \in U$, the maximal integral curve through p has domain equal to $\mathbb{R}$.

Definitions 18.2.8. The **divergence** of a smooth vector field $\mathbf{X}$ on $U \subset \mathbb{R}^{n+1}$ is the function $div\ \mathbf{X} : U \to \mathbb{R}$ defined by

$$div\ X(x_1, x_2, \dots, x_{n+1}) = \sum_{i=1}^{n+1} \frac{\partial X_i}{\partial x_i}$$

where X_i are the component function of the associated function X of the vector field $\mathbf{X}$.

For example, Consider $\mathbf{X}$ with associated function $X : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $X(x_1, x_2) = (2x_1, x_1x_2)$. Then $div\ \mathbf{X}(x_1, x_2) = \frac{\partial X_1}{\partial x_1} + \frac{\partial X_2}{\partial x_2} = 2 + x_1$.

18.3 The Tangent Space

18.4 Surfaces

18.5 Vector Fields on Surfaces; Orientation

18.6 The Gauss Map

Suppose S is an n-surface. From the definition of an n-surface, there exists a smooth function $f : U \to \mathbb{R}$ where U is an open subset of $\mathbb{R}^{n+1}$ such that $S = f^{-1}(c)$ for some real value $c \in \mathbb{R}$ and every point on S is a regular point of f. That is $\nabla f(p) \neq \mathbf{0}$ for every point p on the surface S.

We have proved that every n-surface has exactly two orientations $\mathbf{N}_1$ and $\mathbf{N}_2$. These orientations are $\frac{\nabla f}{\|\nabla f\|}$ and $\frac{-\nabla f}{\|\nabla f\|}$. Given an orientation $\mathbf{N}$ (either $\mathbf{N}_1$ or $\mathbf{N}_2$), the surface together with that orientation is collectively referred as an oriented n-surface.

Since orientation N is a smooth, unit normal vector field. The vector field N has an associated function $N : U \to \mathbb{R}^{n+1}$. That is $\mathbf{N}(p) = (p, N(p))$ where $N : U \to \mathbb{R}^{n+1}$. And we already have, $\mathbf{N}(p) = (p, N(p)) = (p, \frac{\pm\nabla f}{\|\nabla f\|})$. This associated function restricted to the n-surface S is the Gauss Map. That is, $N : S \to \mathbb{R}^{n+1}$.

From the definition of orientation, we know that this function is actually assigning direction to each point on that surface S. If you don't remember, the directions are vector in $\mathbb{R}^{n+1}$ of unit length. That is $\|v\| = 1$. Thus, the range of Gauss Map is a subset of the set of all directions. And unit sphere S^n is $\mathbb{R}^{n+1}$ is the set of all directions in $\mathbb{R}^{n+1}$.

Thus, we may write Gauss Map, $N : S \to S^n$

18.6.1 Spherical Image

We already saw that, the Gauss Map $N : S \to S^n$ is a function which maps directions/unit vectors to each point on that oriented surface S.

Do we need an oriented surface ? Yes. The Gauss Map is defined by this orientation. If we are provided with an oriented n Surface S, then we have a unit vector/orientation assigned to each point p on that surface. And Gauss Map assigns this unit vector to the point p on surface S.

We already saw that the range of the Gauss Map is a subset of the unit n Sphere S^n. In other words, the Gauss Map assigns each point on the oriented n-surface S into a subset of the unit n sphere S^n. Thus, **range of the Gauss Map is referred as the spherical image of the oriented n-surface** S.

$$N(S) = \{q \in S^n : q = N(p),\ p \in S\} \tag{18.9}$$

18.6.2 Compact, connected, oriented n Surface

Suppose we have a compact, connected, oriented n-surface S. The compact subsets in Euclidean spaces are closed and bounded subsets. And connected subsets in Euclidean Spaces are path connected.

Theorem 18.6.1 (Spherical Image of Compact, Connected, Oriented Surface)**.** *The Gauss map of a compact, connected, oriented n-surface is surjective.*

Proof. Let $v \in S^n$ be a direction in $\mathbb{R}^{n+1}$. Let S be a compact, connected, oriented n-surface with orientation N such that $S = f^{-1}(c)$ and every point $p \in S$ are regulr points of the smooth function $f : U \to \mathbb{R}$ where U is an open subset of $\mathbb{R}^{n+1}$. Thus, we have the Gauss Map $N : S \to S^n$ defined by the orientation on S.

Since v is arbitary, it is enough to prove that $v \in N(S)$. Suppose there exists $v \in S^n$ such that $v \notin N(S)$, then the Gauss Map is not surjective. In other words, N is surjective if for every $v \in S^n,\ v \in N(S)$ OR for every $v \in S^n$, there exists $p \in S$ such that $v = N(p)$

Let $g : \mathbb{R}^{n+1} \to \mathbb{R}$ defined by $g(p) = p \cdot v$. Then g is a smooth function since first order partial derivatives are constant functions and all other partial derivatives of higher orders vanishes.

Since S is compact, g restricted to S is a continuous function defined on a compact interval. And thus it attains maximum and minimum values, say p and q. The maximum and minimum values of the dot product $p \cdot v$ are $\pm v$.

By Lagrange's multiplier theorem, $\nabla g(p) = \lambda \nabla f(p)$ and $\nabla g(q) = \lambda \nabla f(q)$. From the definition of the Gauss Map, we have $\nabla g(p) = \lambda \nabla f(p) = \lambda \|\nabla f(p)\| \mathbf{N}(p) = \lambda \|\nabla f(p)\|(p, v)$. Thus, v and $N(p)$ are multiples of one another. Similarly, $\nabla g(q) = \lambda \|\nabla f(q)\| \mathbf{N}(q)$. Therefore $N(p) = \pm v$ and $N(q) = \pm v$.

It remains to show that $N(p) \neq N(q)$. Suppose $N(p) = N(q)$. If there exists continuous function α such that $\alpha : [0, 1] \to \mathbb{R}^{n+1}$, $\alpha(0) = p$, $\alpha(1) = q$, $\dot\alpha(0) = (p, v)$ and $\dot\alpha(1) = (q, v)$. And α maps the interior, the open interval $(0, 1)$ outside the surface S. Then by intermediate value theorem, we arrive at a contradiction. And thus $N(p) \neq N(q)$.

Let $\alpha_1 : [0, x] \to \mathbb{R}^{n+1}$ defined by $\alpha_1(t) = p + tv$. Let $\alpha_2 : [y, 1] \to \mathbb{R}^{n+1}$ defined by $\alpha_2(t) = q + (t-1)v$. Let $\alpha_3 : [x, y] \to S_1$ where S_1 is an n sphere properly containing S. Such an n sphere exists, since S is compact (bounded). And there exists such a function α_3, the image of which is a compact subset of S_1.

Now consider $\alpha : [0, 1] \to \mathbb{R}^{n+1}$ defined by

$$\alpha(t) = \begin{cases} \alpha_1(t) & t \in [0, x) \\ \alpha_3(t) & t \in [x, y] \\ \alpha_2(t) & t \in (y, 1] \end{cases} \tag{18.10}$$

Clearly, α is a smooth function with $\alpha(t) \notin S,\ t \in (0, 1)$ and

$$\begin{aligned}
\alpha(0) &= \alpha_1(0) = p + 0v = p \\
\alpha(1) &= \alpha_2(1) = q + (1-1)v = q \\
\dot\alpha(0) &= \frac{d\alpha_1}{dt}(0) = \frac{d(p + tv)}{dt}(0) = v \\
\dot\alpha(1) &= \frac{d\alpha_2}{dt}(1) = \frac{d(q + (t-1)v}{dt} = v
\end{aligned}$$

We have, $f(\alpha(0)) = f(0) = c$ since $p \in S = f^{-1}(c)$. Similarly, $f(\alpha(1)) = c$. And $(f \circ \alpha)'(0) = \nabla f \circ \alpha(0)\dot\alpha(0) = \nabla f(p) \cdot \dot\alpha(0) = \|\nabla f(p)\| N(p) \cdot v$. Similarly, $(f \circ \alpha)'(1) = \|\nabla f(q)\| N(q) \cdot v$. We have assumed that $N(p) = N(q)$. Then, $f \circ \alpha$ is either increasing at both 0 and 1 OR decreasing at both 0 and 1.

Without Loss of Generality, Suppose that $f \circ \alpha$ is increasing at either points. Then, there exists a sufficiently small $\epsilon > 0$ such that $f(\alpha(\epsilon)) > c$ and $f(\alpha(1 - \epsilon)) < c$. Since, $f \circ \alpha$ must have a value greater than c immediately after 0

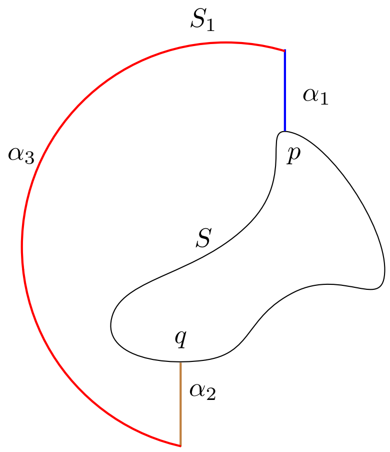


Figure 18.6: Construction of α

and should have a value less than c just before reaching 1 as the function is increasing at either points (and in some small neighbourhood of those points).

By Intermediate Value theorem, the exists $t \in (0,1)$ such that $f \circ \alpha(t) = c$ since the composition of smooth functions f and α is also smooth. But, it is clear from the construction that $\alpha(t)$ doesn't belong to the surface S. And therefore, $\alpha(t) \neq c \implies f(\alpha(t)) \neq c$ for any $t \in (0,1)$. Thus by contradiction, $N(p) \neq N(q)$. And if N achieves v at p. Then it achieves $-v$ at q. And since $v \in S^n$ is arbitrary, $N(S) = S^n$ and the spherical image is the entire n sphere OR the Gauss map is surjective. □

Given a compact, connected oriented, n-surface S, the Gauss Map on S is surjective. In other words, the spherical image of such a surface is the unit n sphere S^n itself.

Connectedness is not that critical(in my opinion). For a compact, orientated n-surface with multiple components, the above observation is valid for each connected component. And thus for any compact, oriented surface. Again, n-surfaces are always closed. Thus, the restriction practically reduces to boundedness of the n-surface.

18.7 Geodesics

We already know that our earth is not flat. Still, we feel like we move in straight lines. And our 'straight lines' are curved for an observer who is not on earth. Geodesics are straightlines on an n-surface S.

vector field along α is function which assigns $X(t)$ at $\alpha(t)$ for each $t \in I$. The defintion of vector field doesn't allow you to assign multiple vectors at a point. But, vector field along α allows you to assign vectors to points on a parametrised curve depending on the value of parameter t.

function along α is function with the same domain I as α.

derivative of vector field X along α is a vector field along α given by $\dot{\mathbf{X}}(t) = \left(\alpha(t), \dfrac{dX}{dt}(t)\right)$ where $\mathbf{X}(t) = (\alpha(t), X(t))$.

velocity of α is a vector field along α defined by $\dot{\boldsymbol{\alpha}}(t) = \left(\alpha(t), \dfrac{d\alpha}{dt}(t)\right)$.

Suppose $\alpha : I \to \mathbb{R}^2$ is defined by $\alpha(t) = (3t, t^2)$.

Then velocity of α is $\dot{\boldsymbol{\alpha}}(t) = \left(\alpha(t), \dfrac{d\alpha}{dt}(t)\right) = (3t, t^2, 3, 2t)$.

speed of α is $\|\dot{\boldsymbol{\alpha}}(t)\|$.

Speed of α is $\|\dot{\boldsymbol{\alpha}}(t)\| = \sqrt{9 + 4t^2}$.

acceleration α is a vector field along α defined by $\ddot{\boldsymbol{\alpha}}(t) = \left(\alpha(t), \dfrac{d^2\alpha}{dt^2}(t)\right)$.

Acceleration of α is $\ddot{\boldsymbol{\alpha}}(t) = (3t, t^2, 0, 2)$

18.7.1 Properties of differentiation

Let $\mathbf{X}, \mathbf{Y}$ be smooth vector fields along parametrised curve $\alpha : I \to \mathbb{R}^{n+1}$.

- $(\mathbf{X} \dot{+} \mathbf{Y}) = \dot{\mathbf{X}} + \dot{\mathbf{Y}}$

$$\begin{aligned}
(\mathbf{X} + \mathbf{Y})(t) =& (\alpha(t), X(t)) + (\alpha(t), Y(t)) = (\alpha(t), X(t) + Y(t)) \\
(\mathbf{X} \dot{+} \mathbf{Y})(t) =& \left(\alpha(t), \frac{d}{dt}X(t) + Y(t)\right) \\
=& \left(\alpha(t), \frac{d}{dt}X(t)\right) + \left(\alpha(t), \frac{d}{dt}Y(t)\right) \\
=& \dot{\mathbf{X}}(t) + \dot{\mathbf{Y}}(t)
\end{aligned}$$

- $(f\dot{\mathbf{X}}) = f'\mathbf{X} + f\dot{\mathbf{X}}$

$$\begin{aligned}
f\mathbf{X}(t) =& f(t)(\alpha(t), X(t)) = (\alpha(t), f(t)X(t)) \\
(f\dot{\mathbf{X}})(t) =& \left(\alpha(t), \frac{d}{dt}f(t)X(t)\right) \\
=& \left(\alpha(t), f'(t)X(t) + f(t)\frac{d}{dt}X(t)\right) \\
=& (\alpha(t), f'(t)X(t)) + \left(\alpha(t), f(t)\frac{d}{dt}X(t)\right) \\
=& f'(t)\,(\alpha(t), X(t)) + f(t)\left(\alpha(t), \frac{d}{dt}X(t)\right) \\
=& f'\mathbf{X}(t) + f\dot{\mathbf{X}}(t)
\end{aligned}$$

- $(\mathbf{X}\cdot\mathbf{Y})' = \dot{\mathbf{X}}\cdot\mathbf{Y} + \mathbf{X}\cdot\dot{\mathbf{Y}}$

$$\begin{aligned}(\mathbf{X}\cdot\mathbf{Y}) =&(\alpha(t), X(t))\cdot(\alpha(t), Y(t)) = \sum_{k=1}^{n+1} X_k(t)Y_k(t)\\ (\mathbf{X}\cdot\mathbf{Y})' =&\frac{d}{dt}\sum_{k=1}^{n+1} X_k(t)Y_k(t)\\ =&\sum_{k=1}^{n+1}\frac{d}{dt}X_k(t)Y_k(t)\\ =&\sum_{k=1}^{n+1} X_k'(t)Y_k(t) + X_k(t)Y_k'(t)\end{aligned}$$

$$\dot{\mathbf{X}}(t)\cdot\mathbf{Y}(t) = \left(\alpha(t), \frac{d}{dt}X(t)\right)\cdot(\alpha(t), Y(t)) = \sum_{k=1}^{n+1} X_k'(t)Y_k(t)$$

$$\mathbf{X}(t)\cdot\dot{\mathbf{Y}}(t) = (\alpha(t), X(t))\cdot\left(\alpha(t), \frac{d}{dt}Y(t)\right) = \sum_{k=1}^{n+1} X_k(t)Y_k'(t)$$

Definitions 18.7.1 (geodesic)**.** Let S be an n-surface. A Geodesic on S is a parametrised curve $\alpha : I \to S$ whose acceleration is orthogonal to S everywhere.

$$\ddot{\alpha}(t) \in S_{\alpha(t)}^{\perp} \tag{18.11}$$

18.7.2 An illustrative example

We know that S^1 given by $x_1^2+x_2^2 = 1$ is a 1-surface in $\mathbb{R}^2$. Consider the cylinder C over S^1, $x_1^2 + x_2^2 = 1$. Clearly, C is a 2 surface in $\mathbb{R}^3$. Also, $f : \mathbb{R}^3 \to \mathbb{R}$ defined by $f(x_1, x_2, x_3) = x_1^2 + x_2^2$ is a smooth function such that $C = f^{-1}(1)$ and $\nabla f(x_1, x_2, x_3) = (x_1, x_2, x_3, 2x_1, 2x_2, 0) \neq \mathbf{0}$.

Clearly, every vector orthogonal to the surface in a scalar multiple of ∇f at that point. Therefore, vectors orthogonal to the surface C at $\alpha(t)$ is of the form $(x_1, x_2, x_3, kx_1, kx_2, 0)$ where $k \in \mathbb{R}$.

If there exists a geodesic α in C, then $\ddot{\alpha}(t) \in S_{\alpha(t)}^{\perp}$. That is, $\ddot{\alpha}(t) = (x_1, x_2, x_3, kx_1, kx_2, 0)$. Thus, we need component functions $x_1(t), x_2(t), x_3(t)$ satisfying

$$\frac{d^2}{dt^2}x_1(t) =kx_1(t) \tag{18.12}$$

$$\frac{d^2}{dt^2}x_2(t) =kx_2(t) \tag{18.13}$$

$$\frac{d^2}{dt^2}x_3(t) =0 \tag{18.14}$$

We have, $\dfrac{d^2}{dt}\cos t = -\cos t$ and $\dfrac{d^2}{dt^2}\sin t = -\sin t$. Thus, the parametrised curve $\alpha : I \to \mathbb{R}^3$ defined by $\alpha(t) = (\cos t, \sin t, t)$ is a geodesic in C since $\ddot{\alpha}(t) = (\cos t, \sin t, t, -\cos t, -\sin t, 0)$.

18.7.3 Maximal Geodesic

The conditions $\alpha(0) = p$, $\dot{\alpha}(0) = \mathbf{v}$ says that parametric curves are unique except for linear transformations on the parameter. That is, Suppose there exists another parametrised curve β in S through p with initial velocity $\mathbf{v}$ with $\beta(t_0) = p$ and $\dot{\beta}(t_0) = \mathbf{v}$. Then, there exists a real number κ such that $\alpha(t) = \beta(\kappa t + t_0)$.

In other words, both α and β passes through the same points and the vectors assigned at each point is the same. And the difference between such two parametrised curves doesn't have any impact on the properties we are interested in.

The condition, if $\beta : \tilde{I} \to S$ is any other geodesic in S with $\beta(0) = p$ and $\dot{\beta}(0) = \mathbf{v}$, then $\tilde{I} \subset I$ and $\beta(t) = \alpha(t)$, $\forall t \in \tilde{I}$. is another way of saying that α is maximal and uniquely defined.

In essence, the following theorem says that if you are standing on an n-surface S at a point, say p. You can move on that surface in straightline from p, with any initial velocity, $\mathbf{v}$. Note that the velocity allows you to choose both direction and speed.

Theorem 18.7.1 (maximal geodesic)**.** *Let S be an n-surface in $\mathbb{R}^{n+1}$. Let $p \in S$ and $\mathbf{v} \in S_p$. Then there exists a unique, maximal geodesic $\alpha : I \to S$ in S through p with initial velocity $\mathbf{v}$.*

Proof. Let S be an n-surface, then there exists a smooth function $f : U \to \mathbb{R}$ where U is an open subset of $\mathbb{R}^{n+1}$. And every points on that surface is regular with respect to f. That is, $\nabla f(p) \neq 0$, $\forall p \in S$. WLOG assume that every points in U is regular.

Define $N = \frac{\nabla f}{\|\nabla f\|}$. A parametrised curve α in S is a geodesic if it satisfies $\ddot{\alpha}(t) \in S^{\perp}_{\alpha(t)}$. Therefore,

$$\ddot{\alpha}(t) = g(t)\mathbf{N}(\alpha(t)) \tag{18.15}$$

Why don't we write $\ddot{\boldsymbol{\alpha}}(t) = \kappa\mathbf{N}(\alpha(t))$? The vectors $\kappa\mathbf{N}(\alpha(t))$ are orthogonal to S. However, the coverse is not true. For a geodesic it is not necessary that $\ddot{\boldsymbol{\alpha}}(t) = \kappa\mathbf{N}(\alpha(t))$. The acceleration could be any vector in $S^{\perp}_{\alpha(t)}$. Since $S^{\perp}_{\alpha(t)}$ is spanned by $\mathbf{N}(\alpha(t))$, at each point on $\alpha(t)$ the scalars may be different. We overcome this with the help of a real valued function $g : I \to \mathbb{R}$ along α.

We have $\ddot{\boldsymbol{\alpha}} = g(\mathbf{N} \circ \alpha)$. Thus $\ddot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha) = g\|\mathbf{N} \circ \alpha\|^2 = g$ since orientation $\mathbf{N}$ assigns directions (vectors of unit length) to each point of that surface.

$[\dot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha)]' = \ddot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha) + \dot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \dot{\circ} \alpha)$ since $(\mathbf{X} \cdot \mathbf{Y})' = \dot{\mathbf{X}} \cdot \mathbf{Y} + \mathbf{X} \cdot \dot{\mathbf{Y}}$.

Thus, $\ddot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha) = [\dot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha)]' - \dot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \dot{\circ} \alpha) = -\dot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \dot{\circ} \alpha)$ since $\dot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha) = 0$ as $\dot{\boldsymbol{\alpha}} \in S_{\alpha(t)}$, $\mathbf{N} \circ \alpha \in S^{\perp}_{\alpha(t)}$ and $\dot{\boldsymbol{\alpha}} \perp (\mathbf{N} \circ \alpha)$. That is, velocity vectors always belongs to the tangent space and $(\mathbf{N} \circ \alpha)$ is an orientation which is orthogonal to all the tangent vectors.

Substituting the value of g in equation(18.15). We get

$$\ddot{\boldsymbol{\alpha}} + [\dot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \dot{\circ} \alpha)] (\mathbf{N} \circ \alpha) = \mathbf{0} \tag{18.16}$$

We have,

$$\frac{dN_1}{dt} = \frac{\partial N_1}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial N_1}{\partial x_2}\frac{dx_2}{dt} + \cdots + \frac{\partial N_1}{\partial x_{n+1}}\frac{dx_{n+1}}{dt} = \sum_{k=1}^{n+1}\frac{\partial N_1}{\partial x_k}\frac{dx_k}{dt} \tag{18.17}$$

Thus,

$$(\mathbf{N}\dot{\circ}\alpha) = \left(\alpha(t), \sum_{k=1}^{n+1}\frac{\partial N_1}{\partial x_k}\frac{dx_k}{dt}, \sum_{k=1}^{n+1}\frac{\partial N_2}{\partial x_k}\frac{dx_k}{dt}, \ldots, \sum_{k=1}^{n+1}\frac{\partial N_{n+1}}{\partial x_k}\frac{dx_k}{dt}\right) \tag{18.18}$$

Equating the components on either sides of the equation(18.16), we get a system of $n+1$ second order differential equations,

$$\frac{d^2}{dt^2}x_i(t) + \left[\sum_{j,k=1}^{n+1}\frac{\partial N_j}{\partial x_k}\frac{dx_k}{dt}\frac{dx_j}{dt}\right]N_i \circ \alpha = 0 \tag{18.19}$$

By the existence theorem for solution of such system of second order differential equations, there exists an open interval I containing 0 and $(n+1)$ functions $x_k : I \to \mathbb{R}$ satisfying the above system. Then $\beta : I \to S$ defined by $\beta(t) = (x_1(t), x_2(t), \ldots, x_{n+1}(t))$

By the uniqueness theorem for solution of such system of second order differential equations, if there exists another open interval $\tilde{I}$ containing 0 and $\beta_1 : I_1 \to U$. Then $\beta(t) = \beta(t)$ for every $t \in I \cap I_1$.

Let $\beta_1, \beta_2, \ldots$ be geodesics through p with inintial velocity $\mathbf{v}$ with parameter domain $I_1, I_2, \ldots$. Let $I = \cup_k I_k$ and $\alpha : I \to U$ defined by $\alpha(t) = \beta_k(t),\ t \in I_k$. Clearly, α is unique and maximal.

Now it is enough to prove that α is parametrised curve in S. We have $(f \circ \alpha)'(t) = \nabla f(\alpha(t)) \cdot \dot{\boldsymbol{\alpha}}(t) = \|\nabla f(\alpha(t))\| \mathbf{N}(\alpha(t)) \cdot \dot{\boldsymbol{\alpha}}(t) = 0$ since $\dot{\boldsymbol{\alpha}} \perp (\mathbf{N} \circ \alpha)$. Therefore, $f \circ \alpha : I \to \mathbf{R}$ is constant. Also, $f(\alpha(0)) = f(p) = c$ since $p \in S$ and $S = f^{-1}(c)$. Thus, $f \circ \alpha = c \implies \alpha \subset f^{-1}(c)$. Thus, α is a parametrised curve in n-surface S. □

18.7.4 Properites of Geodesics

1. $\dot{\boldsymbol{\alpha}} \perp \ddot{\boldsymbol{\alpha}}$ since $\dot{\boldsymbol{\alpha}} \in S_{\alpha(t)}$ and $\ddot{\boldsymbol{\alpha}} \in S_{\alpha(t)}^{\perp}$
   In other words, the acceleration is orthogonal to the velocity vector.

2. Constant Speed, $\|\mathbf{v}\|$
   We have, $(\dot{\boldsymbol{\alpha}} \cdot \dot{\boldsymbol{\alpha}})' = \ddot{\boldsymbol{\alpha}} \cdot \dot{\boldsymbol{\alpha}} + \dot{\boldsymbol{\alpha}} \cdot \ddot{\boldsymbol{\alpha}} = 2\dot{\boldsymbol{\alpha}} \cdot \ddot{\boldsymbol{\alpha}} = 0$
   Clearly, $\frac{d}{dt}\|\dot{\boldsymbol{\alpha}}\|^2 = (\dot{\boldsymbol{\alpha}} \cdot \dot{\boldsymbol{\alpha}})' = 0$. Therefore, α has constant speed.

Remark : Geodesics in cylinder over S^1 are horizontal circles, vertical lines, helix or a constant.

18.8 Parallel Transport

covariant derivative Covariant derivative of $\mathbf{X}$ is the vector field $\mathbf{X}'$ tangent to S along α given by $\mathbf{X}'(t) = \dot{\mathbf{X}}(t) - [\dot{\mathbf{X}}(t) \cdot \mathbf{N}(\alpha(t))]\mathbf{N}(\alpha(t))$. And it is independent of the choie of the orientation.

covariant acceleration Let $\alpha : I \to S$ be a parametrised curve in S. Then covariant acceleration of α is $(\dot{\boldsymbol{\alpha}})' = \ddot{\boldsymbol{\alpha}} - [\ddot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha)]\,(\mathbf{N} \circ \alpha)$ along α.

18.8.1 Properties of Covariant Derivative

Let $\mathbf{X}, \mathbf{Y}$ be smooth vector fields tangent to S along α. Then, $\mathbf{X} \cdot (\mathbf{N} \circ \alpha) = \mathbf{0}$.

1. $(\mathbf{X} + \mathbf{Y})' = \mathbf{X}' + \mathbf{Y}'$

$$\begin{aligned}(\mathbf{X} + \mathbf{Y})' =& (\mathbf{X} \dot{+} \mathbf{Y}) - [(\mathbf{X} \dot{+} \mathbf{Y}) \cdot (\mathbf{N} \circ \alpha)](\mathbf{N} \circ \alpha) \\ =& (\dot{\mathbf{X}} + \dot{\mathbf{Y}}) - [(\dot{\mathbf{X}} + \dot{\mathbf{Y}}) \cdot (\mathbf{N} \circ \alpha)](\mathbf{N} \circ \alpha) \\ =& \left(\dot{\mathbf{X}} - \left[\dot{\mathbf{X}} \cdot (\mathbf{N} \circ \alpha)\right](\mathbf{N} \circ \alpha)\right) + \left(\dot{\mathbf{Y}} - \left[\dot{\mathbf{Y}} \cdot (\mathbf{N} \circ \alpha)\right](\mathbf{N} \circ \alpha)\right) \\ =& \mathbf{X}' + \mathbf{Y}'\end{aligned}$$

2. $(f\mathbf{X})' = f'\mathbf{X} + f\mathbf{X}'$

$$\begin{aligned}(f\mathbf{X})' =& (\dot{f\mathbf{X}}) - \left[(\dot{f\mathbf{X}}) \cdot (\mathbf{N} \circ \alpha)\right](\mathbf{N} \circ \alpha) \\ =& \left(f'\mathbf{X} + f\dot{\mathbf{X}}\right) - \left[(f'\mathbf{X} + f\dot{\mathbf{X}}) \cdot (\mathbf{N} \circ \alpha)\right](\mathbf{N} \circ \alpha) \\ =& f'\left(\mathbf{X} - [\mathbf{X} \cdot (\mathbf{N} \circ \alpha)]\,(\mathbf{N} \circ \alpha)\right) + f\left(\dot{\mathbf{X}} - \left[\dot{\mathbf{X}} \cdot (\mathbf{N} \circ \alpha)\right](\mathbf{N} \circ \alpha)\right) \\ =& f'\mathbf{X} + f\dot{\mathbf{X}} \text{ since } \mathbf{X} \cdot (\mathbf{N} \circ \alpha) = \mathbf{0}\end{aligned}$$

3. $(\mathbf{X} \cdot \mathbf{Y})' = \mathbf{X}' \cdot \mathbf{Y} + \mathbf{X} \cdot \mathbf{Y}'$

$$\begin{aligned}(\mathbf{X} \cdot \mathbf{Y})' =& \dot{\mathbf{X}} \cdot \mathbf{Y} + \mathbf{X} \cdot \dot{\mathbf{Y}} \\ =& \dot{\mathbf{X}} \cdot \mathbf{Y} - \left[\dot{\mathbf{X}} \cdot (\mathbf{N} \circ \alpha)\right]\mathbf{0} + \mathbf{X} \cdot \dot{\mathbf{Y}} - \left[\dot{\mathbf{Y}} \cdot (\mathbf{N} \circ \alpha)\right]\mathbf{0}\end{aligned}$$

Substituting $\mathbf{X} \cdot (\mathbf{N} \circ \alpha) = \mathbf{0}$ and $\mathbf{Y} \cdot (\mathbf{N} \circ \alpha) = \mathbf{0}$, we get

$$\begin{aligned}=& \dot{\mathbf{X}} \cdot \mathbf{Y} - \left[\dot{\mathbf{X}} \cdot (\mathbf{N} \circ \alpha)\right][(\mathbf{N} \circ \alpha) \cdot \mathbf{Y}] + \mathbf{X} \cdot \dot{\mathbf{Y}} - \left[\dot{\mathbf{Y}} \cdot (\mathbf{N} \circ \alpha)\right][\mathbf{X} \cdot (\mathbf{N} \circ \alpha)] \\ =& \left(\dot{\mathbf{X}} - \left[\dot{\mathbf{X}} \cdot (\mathbf{N} \circ \alpha)\right](\mathbf{N} \circ \alpha)\right) \cdot \mathbf{Y} + \mathbf{X} \cdot \left(\dot{\mathbf{Y}} - \left[\dot{\mathbf{Y}} \cdot (\mathbf{N} \circ \alpha)\right](\mathbf{N} \circ \alpha)\right) \\ =& \mathbf{X}' \cdot \mathbf{Y} + \mathbf{X} \cdot \mathbf{Y}'\end{aligned}$$

18.8.2 Parallelism

We have seen earlier that, lines that look straight on a surface (geodesics) is not necessarily straight from outside. The same way, two vectors that look *parallel on a surface* (Levi-Civita parallel) is not necessarily parallel.

Euclidean parallel (p, v) and (q, w) are Euclidean parallel if $v = w$.
Let $\alpha : I \to \mathbb{R}^{n+1}$ be a parametrised curve. A vector field $\mathbf{X}$ is Euclidean parallel along α if the associated function X is constant say, v. Then, $\dot{\mathbf{X}} = \left(\alpha(t), \dfrac{d}{dt}X(t)\right) = \left(\alpha(t), \dfrac{dv}{dt}\right) = (\alpha(t), 0) = \mathbf{0}$

Levi-Civita parallel A vector field $\mathbf{X}$ tangent to the surface S along α is (Levi-Civita) parallel if $\mathbf{X}$ is a constant vector field along α with respect to the surface S. That is, $\mathbf{X}' = \mathbf{0}$.

Properties of Levi-Civita parallel

Applying the properties of covariant derivative, we get

1. If $\mathbf{X}$ is parallel along α, then $\mathbf{X}$ has constant length. That is, $\|\mathbf{X}\|' = 0$.
$$\frac{d}{dt}\|\mathbf{X}\|^2 = \frac{d}{dt}\mathbf{X}\cdot\mathbf{X} = 2\mathbf{X}'\cdot\mathbf{X} = 2(\mathbf{0}\cdot\mathbf{X}) = 0$$
2. If $\mathbf{X}, \mathbf{Y}$ are parallel along α, then $\mathbf{X}\cdot\mathbf{Y}$ is constant along α.
$$(\mathbf{X}\cdot\mathbf{Y})' = \mathbf{X}'\cdot\mathbf{Y} + \mathbf{X}\cdot\mathbf{Y}' = \mathbf{0}\cdot\mathbf{Y} + \mathbf{X}\cdot\mathbf{0} = 0$$
3. If $\mathbf{X}, \mathbf{Y}$ are parallel along α, then angle between them is constant along α.
$$\theta = \cos^{-1}\frac{\mathbf{X}\cdot\mathbf{Y}}{\|\mathbf{X}\|\ \|\mathbf{Y}\|} = \cos^{-1}\kappa,\ \text{since } \|\mathbf{X}\|, \|\mathbf{Y}\|, \mathbf{X}\cdot\mathbf{Y} \text{ are constant}$$
4. If $\mathbf{X}, \mathbf{Y}$ are parallel along α, then $\mathbf{X}+\mathbf{Y}, c\mathbf{X}$ are parallel along α.
$$(\mathbf{X}+\mathbf{Y})' = \mathbf{X}' + \mathbf{Y}' = \mathbf{0}+\mathbf{0} = \mathbf{0}$$
$$(c\mathbf{X})' = c\mathbf{X}' = c\mathbf{0} = \mathbf{0}$$
5. The velocity vector field along a parametrised curve α in S is parallel if and only if α is a geodesic.
$$(\dot{\boldsymbol{\alpha}})' = \ddot{\boldsymbol{\alpha}} - [\ddot{\boldsymbol{\alpha}}\cdot(\mathbf{N}\circ\alpha)]\,(\mathbf{N}\circ\alpha) = \mathbf{0} \iff \ddot{\boldsymbol{\alpha}} \in S^{\perp}_{\alpha(t)}$$
Parametrised curve α is geodesic in S if and only if covariant acceleration $(\dot{\boldsymbol{\alpha}})'$ is zero along α since $\ddot{\boldsymbol{\alpha}} \in S^{\perp}_{\alpha(t)}$ and $[\ddot{\boldsymbol{\alpha}}\cdot(\mathbf{N}\circ\alpha)]\,(\mathbf{N}\circ\alpha) = \ddot{\boldsymbol{\alpha}}$.

We have, the notation f' and $\mathbf{X}'$. You should keep a note of the fundamental differences. f' refers to the derivative of a function along α with respect to the parameter t. And $\mathbf{X}'$ refers to the covariant derivative of a vector field tangent to S along α. You should always check whether it is real value OR vector at a point to understand what they really mean.

For example, $\|\mathbf{X}\|'$ is a derivative of a real valued function (derivative of the length of the vectors assigned at different points of a parametrised curve). And it has nothing to do with the covariant derivative $\mathbf{X}'$.

Theorem 18.8.1. *Let S be an n-surface and $\alpha : I \to S$ be a parametrised curve in S. Let $t_0 \in I$ and $\mathbf{v} \in S_{\alpha(t_0)}$. Then there exists a unique vector field $\mathbf{V}$ tangent to S along α which is parallel and has $\mathbf{V}(t_0) = \mathbf{v}$.*

Proof. Let S be an n-surface with orientation $\mathbf{N}$. Suppose $\mathbf{V}$ is a vector field tangent to S along α and is parallel. That is,$\mathbf{V}(t) \in S_{\alpha(t)}$ and $\mathbf{V}' = \mathbf{0}$.

$$\begin{aligned}
\mathbf{V}' =& \dot{\mathbf{V}} - \left[\dot{\mathbf{V}} \cdot (\mathbf{N} \circ \alpha\right] (\mathbf{N} \circ \alpha) \\
=& \dot{\mathbf{V}} - \left[(\mathbf{V} \cdot (\mathbf{N} \circ \alpha))' - \mathbf{V} \cdot (\mathbf{N} \dot{\circ} \alpha)\right] (\mathbf{N} \circ \alpha) \\
=& \dot{\mathbf{V}} + \left[\mathbf{V} \cdot (\mathbf{N} \dot{\circ} \alpha)\right] (\mathbf{N} \circ \alpha) \text{ since } \mathbf{V} \perp \mathbf{N}
\end{aligned}$$

$$\dot{\mathbf{V}} + [\mathbf{V} \cdot (\mathbf{N} \dot{\circ} \alpha)] (\mathbf{N} \circ \alpha) = \mathbf{0} \tag{18.20}$$

Equating the components on either sides of the equation(18.20), we get the following system of $n+1$ first order differential equations,

$$\frac{dV_i}{dt} + \sum_{j=1}^{n+1} [V_j (\mathbf{N}_j \circ \alpha)'] (\mathbf{N}_i \circ \alpha) = 0, \ \forall i \tag{18.21}$$

By the existence and uniqueness theorem for first order differential equations, there exists $V_1(t), V_2(t), \ldots, V_{n+1}(t)$ satisfying the sytem of equations with initial condition $\mathbf{V}(t_0) = (\alpha(t_0), V_1(t_0), V_2(t_0), \ldots, V_{n+1}(t_0)) = \mathbf{v}$.

It remains to prove that $\mathbf{V}$ is tangent to S along α. By taking dot product with $\mathbf{N} \circ \alpha$ on either sides of equation(18.20), we get

$$(\mathbf{V} \cdot \mathbf{N} \circ \alpha)' = \dot{\mathbf{V}} \cdot (\mathbf{N} \circ \alpha) + \mathbf{V} \cdot (\mathbf{N} \dot{\circ} \alpha) = \mathbf{0} \tag{18.22}$$

Thus, $\mathbf{V} \cdot (\mathbf{N} \circ \alpha) = \kappa$ is constant along α. However, $\mathbf{V}(t_0) \cdot (\mathbf{N} \circ \alpha)(t_0) = \mathbf{v} \cdot N(\alpha(t_0)) = 0$ since $\mathbf{v} \in S_{\alpha(t)}$ is a tangent vector and $\mathbf{v} \perp \mathbf{N}$. Therefore, $\mathbf{V} \cdot (\mathbf{N} \circ \alpha) = 0$. And since $\mathbf{V}$ satisfies equation(18.20), this vector field is tangent to S along α and is parallel. ☐

Corollary 18.8.1.1. *Let S be a 2-surface in $\mathbb{R}^3$ and let $\alpha : I \to S$ be a parametrised curve in S with $\dot{\boldsymbol{\alpha}} \neq \mathbf{0}$. Then the vector field $\mathbf{X}$ tangent to S along α is parallel along α if and only if both $\|\mathbf{X}\|$ and the angle between $\mathbf{X}$ and $\dot{\boldsymbol{\alpha}}$ are constant along α.*

Proof. **Sufficient Part :** Let $\mathbf{X}$ be a tangent vector field tangent to S along α. And α is a geodesic in S. Then $\mathbf{X}$ is parallel along α. Also, we have $\dot{\boldsymbol{\alpha}}$ is also parallel along α since $(\dot{\boldsymbol{\alpha}})' = \mathbf{0}$. Thus $\|\mathbf{X}\|$ is constant and the angle between $\mathbf{X}$ and $\dot{\boldsymbol{\alpha}}$ is constant by properties Levi-Civita parallelism.

Necessary Part : Let $\mathbf{X}$ be a vector field tangent to S along α and α is a geodesic in S. Then $\|\dot{\boldsymbol{\alpha}}\|$ is constant. Suppose $\|\mathbf{X}\|$ and the angle θ between $\mathbf{X}$ and $\dot{\boldsymbol{\alpha}}$ are constant. Since S is a 2-surface, the tangent space $S_{\alpha(t)}$ is spanned by two tangent vectors $\mathbf{v}$ and $\dot{\boldsymbol{\alpha}}(t)$ such that $\mathbf{v} \perp \dot{\boldsymbol{\alpha}}(t)$ and $\|\mathbf{v}\| = 1$.

Let $\mathbf{V}$ be the unique vector field tangent to S along α which is parallel, $\mathbf{V} \cdot \dot{\boldsymbol{\alpha}} = 0$ and $\|V\| = 1$. Then, any vector field $\mathbf{X}$ tangent to S along α can be written as linear combination of $\mathbf{V}$ and $\dot{\boldsymbol{\alpha}}$. That is, $\mathbf{X} = f\dot{\boldsymbol{\alpha}} + g\mathbf{V}$

$$\cos\theta = \frac{\mathbf{X} \cdot \dot{\boldsymbol{\alpha}}}{\|\mathbf{X}\| \, \|\dot{\boldsymbol{\alpha}}\|} = \frac{(f\dot{\boldsymbol{\alpha}} + g\mathbf{V}) \cdot \dot{\boldsymbol{\alpha}}}{\|\mathbf{X}\| \, \|\dot{\boldsymbol{\alpha}}\|} = \frac{f\|\dot{\boldsymbol{\alpha}}\|}{\|\mathbf{X}\|} \implies f \text{ is constant}$$

$$\|\mathbf{X}\|^2 = (f\dot{\boldsymbol{\alpha}} + g\mathbf{V}) \cdot (f\dot{\boldsymbol{\alpha}} + g\mathbf{V}) = f^2\|\dot{\boldsymbol{\alpha}}\| + g^2 \implies g \text{ is constant}$$

Since $\mathbf{V}$ and $\dot{\boldsymbol{\alpha}}$ are parallel along α, their linear combination $\mathbf{X}$ is also parallel along α by linearity property (#4) of Levi-Civita parallelism. □

18.8.3 Transporting tangent vectors using parallelism

Suppose you are standing on an n-surface with you hand stretched out as in a pledge. Then you can move on that surface along any smooth road (without sharp turns, bumps or potholes) from one point to another keeping your hand in the same position. This is what a parallel transport does to tangent vectors. And we can calculate the direction you will be pointing, at each point of your journey.

Definitions 18.8.1 (Parallel transport)**.** Let S be an n-surface. Let $p, q \in S$, and let $\alpha : [a, b] \to S$ be a (smooth) parametrised curve in S from p to q. **Parallel transport** is the function $P_\alpha : S_p \to S_q$ defined by $P_\alpha(\mathbf{v}) = \mathbf{V}(b)$ where $\mathbf{v} \in S_p$ and $\mathbf{V}$ is the unique vector field tangent to S along α which is parallel and $\mathbf{V}(a) = \mathbf{v}$.

The following theorem says that, *Parallel transports along piecewise smooth parametrised curves are vector space isomorphisms preserving dot products.*

Theorem 18.8.2. *Let S be an n-surface in $\mathbb{R}^{n+1}$. Let $p, q \in S$ and let α be a piecewise smooth parametrised curve from p to q. Then parallel transport $P_\alpha : S_p \to S_q$ along α is a vector space isomorphism which preserves dot products.*

Proof. Let S be an n-surface and $p, q \in S$. Let $\alpha : [a, b] \to S$ be a piecewise smooth parametrised curve from p to q. Let $P_\alpha : S_p \to S_q$ be a parallel transport and $\mathbf{v}, \mathbf{w} \in S_p$. Clearly, $\mathbf{v} + \mathbf{w}, c\mathbf{v} \in S_p$

To prove that P_α is a vector space isomorphism preserving dot products, we need to prove the following

1. P_α is linear, $P_\alpha(\mathbf{v} + \mathbf{w}) = P_\alpha(\mathbf{v}) + P_\alpha(\mathbf{w})$ and $P_\alpha(c\mathbf{v}) = cP_\alpha(\mathbf{v})$

$$P_\alpha(\mathbf{v} + \mathbf{w}) = (\mathbf{V} + \mathbf{W})(b) = \mathbf{V}(b) + \mathbf{W}(b) = P_\alpha(\mathbf{v}) + P_\alpha(\mathbf{w})$$

$$P_\alpha(c\mathbf{v}) = (c\mathbf{V})(b) = c\mathbf{V}(b) = cP_\alpha(\mathbf{v})$$

2. P_α is bijective

$$\|P_\alpha(\mathbf{v})\| = \|\mathbf{V}(b)\| = 0 \implies \|\mathbf{v}\| = 0 \implies \ker(P_\alpha) = \{\mathbf{0}\}$$

Thus P_α is a linear map from n-dimensional vector space S_p into another n-dimensional vector space S_q. Thus, P_α is onto.

3. P_α preserves dot products, $P_\alpha(\mathbf{v}) \cdot P_\alpha(\mathbf{w}) = \mathbf{v} \cdot \mathbf{w},\ \forall \mathbf{v}, \mathbf{w} \in S_p$
   We have, $\mathbf{V}$ and $\mathbf{W}$ are parallel along α. Then $\mathbf{V} \cdot \mathbf{W}$ is constant.
   That is, $\mathbf{V}(t) \cdot \mathbf{W}(t) = \kappa$ for every $t \in [a, b]$.

$$P_\alpha(\mathbf{v}) \cdot P_\alpha(\mathbf{w}) = \mathbf{V}(b) \cdot \mathbf{W}(b) = \kappa = \mathbf{V}(a) \cdot \mathbf{W}(a) = \mathbf{v} \cdot \mathbf{w}$$

$\square$

18.9 The Weingarten Map

The Weingarten map L_p gives information about the shape of a surface S at a point $p \in S$.

$\nabla_v f$ is the derivative of a function $f : U \to \mathbb{R}$ with respect to a vector tangent to S, say $\mathbf{v} = (p, v)$ where U is an open subset of $\mathbb{R}^{n+1}$, $p \in U$ and $\alpha : I \to S$ is any parametrised curve in S with $\dot{\boldsymbol{\alpha}}(t_0) = \mathbf{v}$ and $\alpha(t_0) = p$. And $\nabla_v f$ is given by,

$$\nabla_v f = (f \circ \alpha)'(t_0) = \nabla f(\alpha(t_0)) \cdot \dot{\boldsymbol{\alpha}}(t_0) = \nabla f(p) \cdot \mathbf{v}$$

For example : if $f(x_1, x_2) = x_1^2 - x_2^2$ and $\mathbf{v} = (1, 1, \cos\theta, \sin\theta)$ Then, $p = (1, 1)$, $v = (\cos\theta, \sin\theta)$ and $\nabla f = (x_1, x_2, 2x_1, -2x_2)$. Therefore, $\nabla_v f = \nabla f(p) \cdot \mathbf{v} = (1, 1, 2, -2) \cdot (1, 1, \cos\theta, \sin\theta) = 2\cos\theta - 2\sin\theta$.

Note : $\nabla_v f$ is independent of the choice of α.

$\nabla_v \mathbf{X}$ is the derivative of a smooth vector field $\mathbf{X}$ on open subset U of $\mathbb{R}^{n+1}$ with respect to $\mathbf{v} \in S_{\alpha(t)}$. And $\nabla_v \mathbf{X}$ is given by

$$\nabla_v \mathbf{X} = (\alpha(t), \nabla_v X_1, \nabla_v X_2, \ldots, \nabla_v X_{n+1})$$

For example, if $\mathbf{X}(x_1, x_2) = (x_1, x_2, x_1 x_2, x_2^2)$ and $\mathbf{v} = (1, 0, 0, 1)$. Then $p = (1, 0)$ and the component functions of the associated function X are $X_1(x_1, x_2) = x_1 x_2$ and $X_2(x_1, x_2) = x_2^2$.

We have, $\nabla X_1(x_1, x_2) = (x_1, x_2, x_2, x_1)$, $\nabla X_2(x_1, x_2) = (x_1, x_2, 0, 2x_2)$. Thus, $\nabla_v X_1 = \nabla X_1(1, 0) \cdot \mathbf{v} = (1, 0, 0, 1) \cdot (1, 0, 0, 1) = 1$ and
$\nabla_v X_2 = \nabla X_2(1, 0) \cdot \mathbf{v} = (1, 0, 0, 0) \cdot (1, 0, 0, 1) = 0$. Therefore,
$\nabla_v \mathbf{X} = (1, 0, \nabla_v X_1, \nabla_v X_2) = (1, 0, 1, 0)$

$D_v \mathbf{X}$ is the covariant derivative of the smooth vector field $\mathbf{X}$ with respect to $\mathbf{v} \in S_{\alpha(t)}$, where $\mathbf{X}$ is tangent to S along α. And $D_v \mathbf{X}$ is given by,

$$D_v \mathbf{X} = \nabla_v \mathbf{X} - [\nabla_v \mathbf{X} \cdot (\mathbf{N} \circ \alpha)]\,(\mathbf{N} \circ \alpha)$$

It is the component of the derivative $\nabla_v \mathbf{X}$ in the tangent space S_p.

18.9.1 Properties of differentiation, $\nabla_v f$

We have, $\nabla_v f : \mathbb{R}^{n+1} \to \mathbb{R}$ is a linear map.

1. $\nabla_{v+w} f = \nabla_v f + \nabla_w f$

$$\nabla_{v+w} f = \nabla f(p) \cdot (\mathbf{v} + \mathbf{w}) = \nabla f(p) \cdot \mathbf{v} + \nabla f(p) \cdot \mathbf{w}$$

2. $\nabla_{cv} f = c \nabla_v f$

$$\nabla_{cv} f = \nabla f(p) \cdot c\mathbf{v} = c \left(\nabla f(p) \cdot \mathbf{v} \right) = c \nabla_v f$$

Clearly, we have $\nabla_v (f+g) = \nabla_v f + \nabla_v g$ and $\nabla_v cf = c \nabla_v f$.

18.9.2 Properties of differentiation, $\nabla_v \mathbf{X}$

1. $\nabla_v (\mathbf{X} + \mathbf{Y}) = \nabla_v \mathbf{X} + \nabla_v \mathbf{Y}$

$$\begin{aligned}
\nabla_v (\mathbf{X} + \mathbf{Y}) &= (\alpha(t), \nabla_v (X_1 + Y_1), \nabla_v (X_2 + Y_2), \ldots, \nabla_v (X_{n+1} + Y_{n+1})) \\
&= (\alpha(t), \nabla_v X_1 + \nabla_v Y_1, \nabla_v X_1 + \nabla_v Y_2, \ldots, \nabla_v X_{n+1} + \nabla_v Y_{n+1}) \\
&= (\alpha(t), \nabla_v X_1, \nabla_v X_2, \ldots, \nabla_v X_{n+1}) + (\alpha(t), \nabla_v Y_1, \nabla_v Y_2, \ldots, \nabla_v Y_{n+1}) \\
&= \nabla_v \mathbf{X} + \nabla_v \mathbf{Y}
\end{aligned}$$

2. $\nabla_v (f\mathbf{X}) = (\nabla_v f) \mathbf{X}(p) + f(p) \nabla_v \mathbf{X}$

$$\begin{aligned}
\nabla_v (f\mathbf{X}) &= (\alpha(t), \nabla_v f X_1, \nabla_v f X_2, \ldots, \nabla_v f X_{n+1}) \\
&= (\alpha(t), (\nabla_v f) X_1 + f \nabla_v X_1, (\nabla_v f) X_2 + f \nabla_v X_2, \ldots, (\nabla_v f) X_{n+1} + f \nabla_v X_{n+1}) \\
&= (\alpha(t), (\nabla_v f) X_1, (\nabla_v f) X_2, \ldots, (\nabla_v f) X_{n+1}) \\
&\quad + (\alpha(t), f \nabla_v X_1, f \nabla_v X_2, \ldots, f \nabla_v X_{n+1}) \\
&= \nabla_v f (\alpha(t), X_1, X_2, \ldots, X_{n+1}) + f (\alpha(t), \nabla_v X_1, \nabla_v X_2, \ldots, \nabla_v X_{n+1}) \\
&= (\nabla_v f) \mathbf{X} + f \nabla_v \mathbf{X}
\end{aligned}$$

3. $\nabla_v (\mathbf{X} \cdot \mathbf{Y}) = \nabla_v \mathbf{X} \cdot \mathbf{Y}(p) + \mathbf{X}(p) \cdot \nabla_v \mathbf{Y}$

$$\begin{aligned}
\nabla_v (\mathbf{X} \cdot \mathbf{Y}) &= \nabla_v (X_1 Y_1 + X_2 Y_2 + \cdots + X_{n+1} Y_{n+1}) \\
&= (\nabla_v X_1) Y_1 + X_1 \nabla_v Y_1 + (\nabla_v X_2) Y_2 + X_2 \nabla_v Y_2 \\
&\quad + \cdots + (\nabla_v X_{n+1}) Y_{n+1} + X_{n+1} \nabla_v Y_{n+1} \\
&= (\nabla_v X_1) Y_1 + (\nabla_v X_2) Y_2 + \cdots + (\nabla_v X_{n+1}) Y_{n+1} \\
&\quad + X_1 \nabla_v Y_1 + X_2 \nabla_v Y_2 + \cdots + X_{n+1} \nabla_v Y_{n+1} \\
&= (\alpha(t), \nabla_v X_1, \nabla_v X_2, \ldots, \nabla_v X_{n+1}) \cdot (\alpha(t), Y_1, Y_2, \ldots, Y_{n+1}) \\
&\quad + (\alpha(t), X_1, X_2, \ldots, X_{n+1}) \cdot (\alpha(t), \nabla_v Y_1, \nabla_v Y_2, \ldots, \nabla_v Y_{n+1}) \\
&= \nabla_v \mathbf{X} \cdot \mathbf{Y} + \mathbf{X} \cdot \nabla_v \mathbf{Y}
\end{aligned}$$

18.9.3 Properties of covariant differentiation, $D_v \mathbf{X}$

1. $D_v (\mathbf{X} + \mathbf{Y}) = D_v \mathbf{X} + D_v \mathbf{Y}$

$$\begin{aligned}
D_v (\mathbf{X} + \mathbf{Y}) &= \nabla_v (\mathbf{X} + \mathbf{Y}) - \left[\nabla_v (\mathbf{X} + \mathbf{Y}) \cdot (\mathbf{N} \circ \alpha) \right] (\mathbf{N} \circ \alpha) \\
&= \nabla_v \mathbf{X} + \nabla_v \mathbf{Y} - \left[(\nabla_v \mathbf{X} + \nabla_v \mathbf{Y}) \cdot (\mathbf{N} \circ \alpha) \right] (\mathbf{N} \circ \alpha) \\
&= \nabla_v \mathbf{X} - \left[\nabla_v \mathbf{X} \cdot (\mathbf{N} \circ \alpha) \right] (\mathbf{N} \circ \alpha) + \nabla_v \mathbf{Y} - \left[\nabla_v \mathbf{Y} \cdot (\mathbf{N} \circ \alpha) \right] (\mathbf{N} \circ \alpha) \\
&= D_v \mathbf{X} + D_v \mathbf{Y}
\end{aligned}$$

2. $D_v(f\mathbf{X}) = (\nabla_v f)\mathbf{X}(p) + f(p)D_v\mathbf{X}$

$$\begin{aligned}D_v(f\mathbf{X}) =& \nabla_v(f\mathbf{X}) - [\nabla_v(f\mathbf{X}) \cdot (\mathbf{N} \circ \alpha)]\,(\mathbf{N} \circ \alpha)\\ =&(\nabla_v f)\mathbf{X} + f\nabla_v\mathbf{X} - [((\nabla_v f)\mathbf{X} + f\nabla_v\mathbf{X}) \cdot \mathbf{N} \circ \alpha)]\,(\mathbf{N} \circ \alpha)\\ =&(\nabla_v f)\mathbf{X} - [(\nabla_v f)\mathbf{X} \cdot (\mathbf{N} \circ \alpha)]\,(\mathbf{N} \circ \alpha) + f\nabla_v\mathbf{X} - [f\nabla_v\mathbf{X} \cdot (\mathbf{N} \circ \alpha)]\,(\mathbf{N} \circ \alpha)\\ =&\nabla_v f\mathbf{X} + fD_v\mathbf{X}\end{aligned}$$

3. $\nabla_v(\mathbf{X} \cdot \mathbf{Y}) = D_v\mathbf{X} \cdot \mathbf{Y}(p) + \mathbf{X}(p) \cdot D_v\mathbf{Y}$

$$\begin{aligned}\nabla_v(\mathbf{X} \cdot \mathbf{Y}) =& \nabla_v(X_1Y_1) + \nabla_v(X_2Y_2) + \cdots + \nabla_v(X_{n+1}Y_{n+1})\\ =&(\nabla_v X_1)Y_1 + X_1\nabla_v Y_1 + (\nabla_v X_2)Y_2 + X_2\nabla_v Y_2\\ &+ \cdots + (\nabla_v X_{n+1})Y_{n+1} + X_{n+1}\nabla_v Y_{n+1}\\ =&(\nabla_v X_1)Y_1 + (\nabla_v X_2)Y_2 + \cdots + (\nabla_v X_{n+1})Y_{n+1}\\ &+ X_1\nabla_v Y_1 + X_2\nabla_v Y_2 + \cdots + X_{n+1}\nabla_v Y_{n+1}\\ =&(\nabla_v\mathbf{X}) \cdot \mathbf{Y} + \mathbf{X} \cdot \nabla_v\mathbf{Y}\\ =&D_v\mathbf{X} \cdot \mathbf{Y} + \mathbf{X} \cdot D_v\mathbf{Y} \text{ since } \mathbf{X}, \mathbf{Y} \perp \mathbf{N}\end{aligned}$$

18.9.4 Weingarten Map, L_p

In the case of covariant derivatives along α, we came across a linear map – the parallel transport P_α. Here in the case of covariant derivative with respect to a vector, we have another linear map – the Weingarten map L_p given by

$$L_p(\mathbf{v}) = -\nabla_v\mathbf{N} \tag{18.23}$$

Weingarten map tells us how much the tangent space turns/tilts as we move through p along α on an n-surface S. It gives the information about the shape of S. And is the **shape operator** of S at p.

18.9.5 Properties of Weingarten Map, L_p

1. The normal component of acceleration is same for all parametrised curves in S passing through p with velocity $\mathbf{v}$, $\ddot{\boldsymbol{\alpha}}(t_0) \cdot \mathbf{N}(p) = L_p(\mathbf{v}) \cdot \mathbf{v}$

$$\begin{aligned}0 =& \dot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha) \text{ since } \dot{\boldsymbol{\alpha}} \perp \mathbf{N}\\ 0 =& [\dot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha)]'\\ =&\ddot{\boldsymbol{\alpha}} \cdot (\mathbf{N} \circ \alpha) + \dot{\boldsymbol{\alpha}} \cdot (\mathbf{N}\dot{\circ}\alpha)\end{aligned}$$

Rearranging the terms and evaluating at $t = t_0$, we get

$$\begin{aligned}\ddot{\boldsymbol{\alpha}}(t_0) \cdot (\mathbf{N} \circ \alpha(t_0)) =& -\dot{\boldsymbol{\alpha}}(t_0) \cdot (\mathbf{N}\dot{\circ}\alpha)(t_0)\\ \ddot{\alpha}(t_0) \cdot \mathbf{N}(p) =& \mathbf{v} \cdot L_p(\mathbf{v}) \text{ since } (\mathbf{N}\dot{\circ}\alpha)(t_0) = \nabla_v\mathbf{N} = -L_p(\mathbf{v})\end{aligned}$$

2. Weingarten map is self adjoint, $L_p(\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot L_p(\mathbf{w})$

Clearly, we have $\mathbf{v}, \mathbf{w} \in S_{\alpha(t)}$ and $\nabla f \in S_{\alpha(t)}^{\perp}$. Also note that

$$\begin{aligned}
\nabla f(p) &= \left(p, \frac{\partial f}{\partial x_1}(p), \frac{\partial f}{\partial x_2}(p), \ldots, \frac{\partial f}{\partial x_{n+1}}(p)\right) \\
\nabla_v(\nabla f(p)) &= \left(p, \nabla_v \frac{\partial f}{\partial x_1}(p), \nabla_v \frac{\partial f}{\partial x_2}(p), \ldots, \nabla_v \frac{\partial f}{\partial x_{n+1}}(p)\right) \\
&= \left(p, \nabla\left(\frac{\partial f}{\partial x_1}\right)(p)\cdot \mathbf{v}, \nabla\left(\frac{\partial f}{\partial x_2}\right)(p)\cdot \mathbf{v}, \ldots, \nabla\left(\frac{\partial f}{\partial x_{n+1}}\right)(p)\cdot \mathbf{v}\right) \\
&= \left(p, \sum_{j=1}^{n+1} \frac{\partial^2 f}{\partial x_j \partial x_1}(p) v_j, \sum_{j=1}^{n+1} \frac{\partial^2 f}{\partial x_j \partial x_2}(p) v_j, \ldots, \sum_{j=1}^{n+1} \frac{\partial^2 f}{\partial x_j \partial x_{n+1}}(p) v_j\right)
\end{aligned}$$

$$\begin{aligned}
L_p(\mathbf{v}) \cdot \mathbf{w} &= -\nabla_v \mathbf{N} \cdot \mathbf{w} \\
&= -\nabla_v\left(\frac{\nabla f}{\|\nabla f\|}\right) \cdot \mathbf{w} \\
&= \left[\left(-\nabla_v \frac{1}{\|\nabla f\|}\right)\nabla f - \frac{1}{\|\nabla f\|}\nabla_v(\nabla f)\right] \cdot \mathbf{w} \\
&= \left[-\frac{1}{\|\nabla f\|}\nabla_v(\nabla f)\right] \cdot \mathbf{w} \text{ since } \nabla f \perp \mathbf{w} \\
&= -\frac{1}{\|\nabla f\|}\left[\sum_{k=1}^{n+1}\left(\sum_{j=1}^{n+1} \frac{\partial^2 f}{\partial x_j \partial x_k}(p) v_j\right) w_k\right] \\
&= -\frac{1}{\|\nabla f\|}\sum_{j,k=1}^{n+1} \frac{\partial^2 f}{\partial x_j \partial x_k}(p) v_j w_k
\end{aligned}$$

Similarly,

$$L_p(\mathbf{w}) \cdot \mathbf{v} = -\frac{1}{\|\nabla f\|}\sum_{j,k=1}^{n+1} \frac{\partial^2 f}{\partial x_j \partial x_k}(p) w_j v_k$$

Since $\dfrac{\partial^2 f}{\partial x_j \partial x_k}(p) = \dfrac{\partial^2 f}{\partial x_k \partial x_j}(p)$, both the sums are equal

$$L_p(\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot L_p(\mathbf{w})$$

18.10 The Curvature of Plane Curves

curvature Let C be plane curve in $\mathbb{R}^2$. The curvature of C is the function $\kappa : \mathbb{R}^2 \to \mathbb{R}$ defined by $\kappa(p) = L_p(\mathbf{v}) \cdot \mathbf{v}/\|\mathbf{v}\|^2$

For a parametrised curve α in C with $\dot{\boldsymbol{\alpha}} \neq \mathbf{0}$,

$$\kappa(\alpha(t)) = \frac{\ddot{\boldsymbol{\alpha}}(t) \cdot (\mathbf{N} \circ \alpha(t))}{\|\dot{\boldsymbol{\alpha}}(t)\|^2}$$

For unit speed parametrised curve passing through p, curvature at p is the normal component of acceleration at that point.

local parametrisation of plane curve Let C be a plane curve in $\mathbb{R}^2$. Then C is an oriented 1-surface. Parametrisation of a segment of C containing p is a function $\alpha : I \to C$ such that

1. α is regular. That is, $\dot{\boldsymbol{\alpha}}(t) \neq \mathbf{0},\ \forall t \in I$
2. α is consistently oriented with C. That is, the orientation at $\alpha(t)$ should be the same as that of C at that point.
3. $p \in \alpha(I)$

circle of curvature Let C be a plane curve in $\mathbb{R}^2$ with orientation $\mathbf{N}$. The circle of curvature of C at a point p is the unique oriented circle O with orientation $\mathbf{N}_1$ such that

1. O is tangent to C at p, $C_p = O_p$.
2. O is oriented consistently with C, $\mathbf{N}(p) = \mathbf{N}_1(p)$ and
3. Its normal turns at the same rate at p as the normal of the plane curve C, $\nabla_v \mathbf{N} = \nabla_v \mathbf{N}_1$

This circle, O is the circle which huges the curve C closest among all circles containing p.

α is regular. That is, $\dot{\boldsymbol{\alpha}}(t) \neq \mathbf{0},\ \forall t \in I$

radius of curvature The radius of the circle of curvature of C at point p. And radius of curvature, $r = 1/|\kappa(p)|$.

center of curvature The center of the circle of curvature of C at point p.

Meaning of Sign of Curvature If curvature at p is positive, $\kappa(p) > 0$ then the curve at p is turning towards $\mathbf{N}(p)$. If curvature at p is negative, $\kappa(p) < 0$ then the curve at p is turning away from $\mathbf{N}(p)$.

Construction of a local parametrisation We have, $C = f^{-1}(c)$ where $\nabla f(q) \neq \mathbf{0}$. Consider $\mathbf{X}(q) = \left(q, \frac{\partial f}{\partial x_2}(q), -\frac{\partial f}{\partial x_1}(q)\right)$. The maximal integral curve α through p in $\mathbf{X}$ is a local parametrisation of a segment of C containing p.

For example, (refer : Exercise 10.3c) $f(x_1, x_2) = x_2 - ax_1^2$. Then $C = f^{-1}(c) = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 - ax_1^2 = c,\ a \neq 0\}$. The global parametrisation of C with orientation $\nabla f / \|\nabla f\|$ can be obtained by constructing the maximal integral curve for each segment of a piecewise smooth plane curve. Since the given curve is smooth, it is sufficient to construct one integral curve through any point on that curve.

We have, $\frac{dx_1}{dt}(t) = -2ax_1(t)$ and $\frac{dx_2}{dt} = 1$. Thus, $x_1(t) = \cos 2at + c_1$ and $x_2(t) = t + c_2$. We have $(0, c) \in C$ since $c - a0^2 = c$. Therefore, $x_1(0) = 0 = c_1$ and $x_2(0) = c = c_2$. Thus, we have $\alpha : I \to \mathbb{R}^2$ given by $\alpha(t) = (\cos 2at, t + c)$ as a global parametrisation of C.

18.10.1 Unit speed local parametrisation of C is unique.

Theorem 18.10.1. *Local parametrisations of plane curves are unique upto reparametrisation.*

Proof. Let α be the unit speed local parametrisation which is the maximal integral curve through p. Let $\beta : \tilde{I} \to C$ be any parametrisation of a segment of C containing p. Then it is enough to prove that there exists a smooth function $h : \tilde{I} \to \mathbb{R}$ such that $h'(t) > 0$ and $\beta(t) = \alpha(h(t))$ for every $t \in \tilde{I}$.

Define $h : \tilde{I} \to \mathbb{R}$ defined by

$$h(t) = \int_{t_0}^{t} \|\dot{\boldsymbol{\beta}}(u)\| du$$

Clearly, h is monotone, $h(t_0) = 0$ and $h'(t) = \|\dot{\boldsymbol{\beta}}(t)\|$. Since h is monotone and strictly increasing, $h(h^{-1}(t)) = t$. Thus, $(h \circ h^{-1})'(t) = h'(h^{-1}(t))(h^{-1})'(t)$. Therefore, $(h^{-1})'(t) = 1/h'(h^{-1}(t))$.

Now, $\beta \circ h^{-1}$ is a reparametrised curve with velocity, $(\boldsymbol{\beta} \dot{\circ} h^{-1})(t)$ given by,

$$\begin{aligned}(\boldsymbol{\beta} \dot{\circ} h^{-1})(t) =& \dot{\boldsymbol{\beta}}(h^{-1}(t)) \ (h^{-1})'(t) \text{ by chain rule} \\ =& \dot{\boldsymbol{\beta}}(h^{-1}(t))/h'(h^{-1}(t)) \\ =& \dot{\boldsymbol{\beta}}(h^{-1}(t))/\|\dot{\boldsymbol{\beta}}(h^{-1}(t))\| \text{ since } h'(t) = \|\dot{\boldsymbol{\beta}}(t)\|\end{aligned}$$

We know that, $\mathbf{X}(\beta(h^{-1}(t)))$ spans the tangent space $S_{\beta(h^{-1}(t))}$.

$$=\mathbf{X}(\beta(h^{-1}(t)) \text{ since } \beta \dot{\circ} h^{-1}(t)/\|\beta \dot{\circ} h^{-1}(t)\| = \mathbf{X}(\beta(h^{-1}(t)))$$

Thus, $\beta \circ h^{-1}$ is an integral curve through p in $\mathbf{X}$. By uniqueness of integral curves, $\beta \circ h^{-1}(t) = \alpha(t)$ for all $t \in \tilde{I}$. ☐

18.11 Arc Length and Line Integrals

length of arc The length of parametrised arc α is the integral of its speed. Let $\alpha : [a, b] \to \mathbb{R}^{n+1}$. Length of α is

$$l(\alpha) = \int_{a}^{b} \|\dot{\boldsymbol{\alpha}}(t)\| dt \tag{18.24}$$

For Example (refer : Exercise 11.4) : Given $\alpha : [0, 2\pi] \to \mathbb{R}^4$, $\alpha(t) = (\cos t, \sin t, \cos t, \sin t)$. Then, $\dot{\boldsymbol{\alpha}}(t) = (\alpha(t), -\sin t, \cos t, -\sin t, \cos t)$. We have, $\|\dot{\boldsymbol{\alpha}}(t)\| = \sqrt{\sin^2 t + \cos^2 t + \sin^2 t + \cos^2 t} = \sqrt{2}$. Therefore, $l(\alpha) = \int_0^{2\pi} \sqrt{2} dt = 2\sqrt{2}\pi$.

Note : Length of arc is the total distance travelled.
For example : $\alpha_1 : [0, 2\pi] \to \mathbb{R}^2$ defined by $\alpha_1(t) = (\cos t, \sin t)$ has its length 2π, the perimeter of unit circle. But, for $\alpha_2 : [0, 4\pi] \to \mathbb{R}^2$ defined by $\alpha_2(t) = (\cos t, \sin t)$ has its length 4π. And $\alpha : \mathbb{R} \to \mathbb{R}^2$ defined by $\alpha(t) = (\cos t, \sin t)$ has its length $+\infty$.

18.11.1 Properties of Arc Length

1. Arc length is preserved under reparametrisation.
   Let $\alpha : [a,b] \to \mathbb{R}^{n+1}$ be a parametrised curve. And let $\beta : [c,d] \to \mathbb{R}^{n+1}$ be a reparametrisation of α defined by $\beta(t) = \alpha(h(t))$ where $h : [a,b] \to [c,d]$. Then length of β is given by,
$$\begin{aligned} l(\beta) &= \int_c^d \|\dot{\beta}(t)\| \; dt \\ &= \int_c^d \|\dot{\alpha}(h(t))\| h'(t) \; dt \\ &= \int_a^b \|\alpha(u)\| \; du = l(\alpha) \end{aligned}$$

2. Unit speed arcs are parametrised by arc length.
   Let α be a unit speed parametrised curve. That is, $\|\dot{\alpha}\| = 1$. Therefore $l(\alpha) = \int_a^b dt = b - a$. Clearly, length of the arc is the length of the parameter interval.

Theorem 18.11.1 (Existence of Global Parametrisation)**.** *Let C be an oriented plane curve. Then there exists a global parametrisation of C if and only if C is connected.*

Proof. **Sufficient Part :** Suppose plane curve C has a global parametrisation, $\alpha : I \to \mathbb{R}^{n+1}$. Let $p, q \in C$, then $p = \alpha(t_1)$ and $q = \alpha(t_2)$ for some $t_1, t_2 \in I$. WLOG $t_1 < t_2$ and clearly, there exists a path from p to q obtained restricting α to $[t_1, t_2]$. Therefore, the plane curve C is path-connected and thus connected.

Necessary Part : Step 1 Construction of α
Let $\mathbf{X}$ be the unit tangent vector field on C (obtained by rotating ∇f by an angle of $-\pi/2$ and orientation $\mathbf{N} = \nabla f / \|\nabla f\|$). Suppose the plane curve $C = f^{-1}(c)$ is connected. Let $p \in C$ and α be the local parametrisation of C at p, which is nothing but the maximal integral curve of $\mathbf{X}$ through p. Let $p_1 \in C$. Parametrisation α is global if $p_1 \in Image\ \alpha$.

Step 2 : Construction of β
The plane curve C is connected. Thus there exists a continuous path β from p to p_1 where $\beta : [a,b] \to C$ with $\beta(a) = p$ and $\beta(b) = p_1$. *We have, $\beta(a) \in Image\ \alpha$. But, $\beta(b) \notin Image\ \alpha$.* Let $t_0 = \sup\{t \in [a,b] : \beta(t) \in Image\ \alpha\}$, the least upper bound of the points which are in $Image\ \alpha$. *That is, if $t > t_0$, then $\beta(t) \notin Image\ \alpha$. And if $t < t_0$, then there exists an $\epsilon > 0$ such that $\beta(t + \epsilon) \in Image\ \alpha$. Otherwise, t_0 is not the supremum.*

Step 3 : Construction of γ
Let γ be the maximal integral curve of $\mathbf{X}$ through $\beta(t_0) = p_0$. Suppose there exists an open rectangle B about $\beta(t_0) = p_0$ such that $C \cap B \subset Image\ \gamma$. Then by the continuity of β, there exists $\delta > 0$ such that $\beta(t) \in Image\ \gamma$ for all $t \in (t_0 - \delta, t_0 + \delta)$. Thus, there exists $\epsilon > 0$ such that $\epsilon < \delta$ and $\beta(t_0 - \epsilon) \in Image\ \gamma$. Since t_0 is the least upper bound, $\beta(t_0 - \epsilon) \in Image\ \alpha$ as well. Thus, α and γ are maximal integral curves through a common point (in the

neighbourhood of p_0). And thus, $Image\ \alpha = Image\ \gamma$. Clearly, $p_1 \in Image\ \alpha$. Therefore, the parametrisation α is global. Therefore it is sufficient to prove that there exists an open box B such that $C \cap B \subset Image\ \gamma$.

Step 4 : Construction of A
Let $u = (\frac{\partial f}{\partial x_2}(p_0), \frac{-\partial f}{\partial x_1}(p_0))$ and $v = (\frac{\partial f}{\partial x_1}(p_0), \frac{\partial f}{\partial x_2}(p_0))$. Clearly, $\mathbf{u} = (p_0, u) \in C_{p_0}$ and $\mathbf{v} = (p_0, v) \perp C_{p_0}$. Consider the open rectangle A given by,

$$A = \{p_0 + ru + sv : |r| < \epsilon_1,\ |s| < \epsilon_2\}$$

where ϵ_1, ϵ_2 are so chosen that A is contained in the domain of f and $\nabla f(q) \cdot (q, v) > 0$ for every $q \in A$. Since $\nabla f(p_0) \cdot \mathbf{v} \neq 0$, by continuity of $\nabla f(q) \cdot (q, v)$ there exists a neighbourhood of p_0 in which this function is positive.

Step 5 : Construction of $\{g_r\}$
Consider the family of functions $\{g_r\}$ defined by $g_r(s) = f(p_0 + ru + sv)$. Since $\nabla f(q) \cdot (q, v) > 0$ for every $q \in A$, for $|r| < \epsilon_1$, $g_r(s)$ is strictly increasing in $(-\epsilon_2, \epsilon_2)$ as $g_r'(s) > 0$. That is, there exists at most one $s \in (-\epsilon_2, \epsilon_2)$ such that $f(p_0 + ru + sv) = c$.

Step 6 : Construction of B
Since the vectors $\mathbf{u}, \mathbf{v}$ spans C_{p_0}, we have $\gamma(t) = p_0 + h_1(t)u + h_2(t)v$ where h_1, h_2 are real-valued functions. Clearly,

$$\begin{aligned}(\gamma(t) - p_0) \cdot \mathbf{u} =& h_1(t)\mathbf{u} \cdot \mathbf{u} + h_2(t)\mathbf{u} \cdot \mathbf{v} \\ =& h_1(t)\|u\|^2 \text{ since } \mathbf{u} \perp \mathbf{v}\end{aligned}$$

$$\begin{aligned}(\gamma(t) - p_0) \cdot \mathbf{v} =& h_1(t)\mathbf{u} \cdot \mathbf{v} + h_2(t)\mathbf{v} \cdot \mathbf{v} \\ =& h_2(t)\|v\|^2\end{aligned}$$

Thus,

$$h_1(t) = \frac{(\gamma(t) - p_0) \cdot \mathbf{u}}{\|u\|^2} \tag{18.25}$$

$$h_2(t) = \frac{(\gamma(t) - p_0) \cdot \mathbf{v}}{\|v\|^2} \tag{18.26}$$

Now, $h_1'(0) = \dot{\gamma}(0) \cdot (p_0, u/\|u\|^2) = \mathbf{X}(p_0) \cdot \mathbf{X}(p_0)/\|\mathbf{X}(p_0)\|^2 = 1$. And $h_1(0) = 0$ and $h_2(0) = 0$. Thus, there exists (t_1, t_2) containing 0 such that $\gamma(t) \in A$ and $h_1'(t) > 0$ for every $t \in (t_1, t_2)$. Set $r_1 = h_1(t_1)$ and $r_2 = h_1(t_2)$. Since h_1 is continuous and strictly increasing, for every $r \in (r_1, r_2)$ there exists $t \in (t_1, t_2)$ such that $r = h_1(t)$ and $s = h_2(t)$.

$$\text{Define } B = \{p_0 + ru + sv : r \in (r_1, r_2), |s| < \epsilon_2\} \tag{18.27}$$

Then $p_0 + ru + su \in B \cap C$ if and only if there exists $t \in (t_1, t_2)$ such that $r = h_1(t)$, and $s = h_2(t)$. That is, $p_0 + ru + sv \in Image\ \gamma$. Therefore, $C \cap B \subset Image\ \gamma$ as required in Step 3. □

Theorem 18.11.2. *Let C be a connected oriented plane curve and let $\beta : I \to C$ be a unit speed global parametrisation of C. Then β is either one to one or periodic. Moreover, β is periodic if and only if C is compact.*

Proof. Suppose $\beta(t_1) = \beta(t_2)$ for some $t_1, t_2 \in I$ $(t_1 \neq t_2)$. Let $\mathbf{X}$ be the unit tangent vector field on C. And let α be the maximal integral curve of $\mathbf{X}$ through $\beta(t_1)$. That is, $\alpha(0) = \beta(t_1) = \beta(t_2)$. Since β is a unit speed global parametrisation of C, β is also a maximal integral curve of $\mathbf{X}$. Thus, β is a reparametrisation of β. Thus, $\beta(t) = \alpha(t - t_1)$ and $\beta(t) = \alpha(t - t_2)$. Let $\tau = t_2 - t_1$. Then $\beta(t) = \alpha(t - t_1) = \alpha(t - t_2) = \alpha(t - t_2 + t_2 - t_1) = \alpha(t + \tau - t_1) = \beta(t + \tau)$. Therefore, β is periodic.

Suppose β is periodic, then its continuous image, $C = \beta[t, t + \tau]$ is compact. Suppose β is not periodic, then β is one to one. Then, C is not compact as β^{-1} doesn't attain its extrema. Thus, it is enough to prove that β^{-1} is continuous.

Given $t_0 \in I$ and $\epsilon > 0$, define $\gamma(t) = \beta(t+t_0)$ such that $|t| < \epsilon$ and $t+t_0 \in I$. Chose an open rectangle B about $p_0 = \beta(t_0)$ such that $C \cap B \subset Image\ \gamma$. Then, $|\beta^{-1}(p) - t_0| = |\gamma^{-1}(p)| < \epsilon$ for every $p \in C \cap B$.

Therefore it is enough to prove that such an open rectangle B exists. The construction of which is given in previous proof. □

Fundamental Domain Let β be a periodic parametrised plane curve with period τ. Then any subset $[t_0, t_0 + \tau]$ of its domain is a fundamental domain of β.

18.11.2 1-Form

Definitions 18.11.1 (1-form)**.** The 1-form is a function $\omega : U \times R^{n+1} \to \mathbb{R}$ where $U \subset \mathbb{R}^{n+1}$ such that for every $p \in U$, $\mathbb{R}^{n+1}_p \subset U \times \mathbb{R}^{n+1}$.

Note : 1-form ω is smooth if ω is a smooth function. A few examples of 1-form are given below: $\omega_{\mathbf{X}}$, df and dx_i.

1-form dual Let $\mathbf{X}$ be a vector field on U. Then 1-form dual of $\mathbf{X}$ is $\omega_{\mathbf{X}}$ given by $\omega_{\mathbf{X}}(p, v) = \mathbf{X}(p) \cdot (p, v)$

differential of f Let $f : U \to \mathbb{R}$ be a smooth function. Then differential of f is df given by $df(\mathbf{v}) = \nabla f(p) \cdot \mathbf{v}$ where $\mathbf{v} = (p, v)$.

Cartesian Coordinate Function Let $U \subset \mathbb{R}^{n+1}$. Then ith cartesian coordinate function x_i is given by $x_i : U \to \mathbb{R},\ x_i(a_1, a_2, \ldots, a_{n+1}) = a_i$.

dx_i The differential of the cartesian coordinate function x_i is given by dx_i where $dx_i : U \times R^{n+1} \to \mathbb{R}$

$$dx_i = \nabla x_i(p) \cdot \mathbf{v} = (p, 0, 0, \ldots, 1, \ldots, 0) \cdot (p, v_1, v_2, \ldots, v_{n+1}) = v_i$$

18.11.3 1-form

Sum of two 1-forms $\omega_1 + \omega_2$

$$\begin{aligned}(\omega_1 + \omega_2)(\mathbf{v}) &= \nabla(\omega_1 + \omega_2) \cdot \mathbf{v} \\ &= \nabla\omega_1 \cdot \mathbf{v} + \nabla\omega_2 \cdot \mathbf{v} \\ &= \omega_1(\mathbf{v}) + \omega_2(\mathbf{v})\end{aligned}$$

Product of function and 1-form $f\omega$

$$\begin{aligned} f\omega(\mathbf{v}) &= \nabla(f\omega)\cdot\mathbf{v} \\ &= f\nabla\omega\cdot\mathbf{v} \\ &= f(p)\omega(\mathbf{v}) \end{aligned}$$

Note : Let $\omega(\mathbf{X}) : U \to \mathbb{R},\ (\omega(\mathbf{X}))(p) = \omega(\mathbf{X}(p))$.
If ω and $\mathbf{X}$ are smooth, then function $\omega(\mathbf{X})$ is also smooth.

Proposition 18.11.1 (1-form representation)**.** *For every* 1*-form* ω *on* U*, there exists unique functions* $f_i : U \to \mathbb{R}$ *such that*

$$\omega = \sum_{i=1}^{n+1} f_i dx_i \tag{18.28}$$

And ω *is smooth if and only if each* f_i *is smooth.*

Proof. Let $\mathbf{X}_j$ be a smooth vector field on U where $X_j(p) = (p, 0, 0, \ldots, 1, \ldots, 0)$. Then, $dx_i(X_j) = \delta_{ij}$. Suppose $\omega = \sum f_i dx_i$. Then

$$\omega(\mathbf{X}_j) = \sum_{i=1}^{n+1}(f_i dx_i)\mathbf{X}_j = \sum_{i=1}^{n+1} f_i(p) dx_i(\mathbf{X}_j) = \sum_{i=1}^{n+1} f_i(p)\delta_{ij} = f_j(p)$$

Thus f_j is unique for each j if they exist. And for each 1-form ω there exist functions f_i such that $\omega = \sum f_i dx_i$ exists where $f_i = \omega(X_i)$. And thus, ω is smooth if and only if each f_i is smooth. □

Corollary 18.11.2.1. *Let* $f : U \to \mathbb{R}$*. Then*

$$df = \sum_{i=1}^{n+1} \frac{\partial f}{\partial x_i} dx_i \tag{18.29}$$

Proof.

$$df(\mathbf{X}_j) = \nabla f \cdot \mathbf{X}_j = \sum_{i=1}^{n+1} \frac{\partial f}{\partial x_i}\delta_{ij} = \frac{\partial f}{\partial x_j} = f_j \implies df = \sum_{i=1}^{n+1} \frac{\partial f}{\partial x_i} dx_i$$

□

18.11.4 Line Integral

Definitions 18.11.2 (Line Integral)**.** Let ω be a 1-form and $\alpha : [a, b] \to U$ be a parametrised curve on U. Then the line integral of ω over α is given by,

$$\int_\alpha \omega = \int_a^b \omega(\dot{\alpha}(t)) \tag{18.30}$$

Note : The line integral of 1-form ω over a parametrised curve is invariant of reparametrisation.

Let $\beta : [c,d] \to U$ be reparametirsation of $\alpha : [a,b] \to U$ where $\beta(t) = \alpha(h(t))$ and function $h : [c,d] \to [a,b]$. Then ,

$$\begin{aligned}\int_\beta \omega &= \int_c^d \omega(\dot{\beta}(t)) \\ &= \int_c^d \omega(\dot{\alpha}(h(t))h'(t)) \\ &= \int_c^d \omega(\dot{\alpha}(h(t)))h'(t) \\ &= \int_a^b \omega(\dot{\alpha}(u)) = \int_\alpha \omega\end{aligned}$$

Theorem 18.11.3. *Let η be a 1-form on $\mathbb{R}^2 - \{0\}$ defined by*

$$\eta = \frac{-x_2}{x_1^2 + x_2^2} dx_1 + \frac{x_1}{x_1^2 + x_2^2} dx_2$$

Then for any closed, piecewise smooth, parametrised smooth curve $\alpha : [a,b] \to \mathbb{R} - \{0\}$, the line integral of η over α

$$\int_\alpha \eta = 2\pi k$$

Proof. Consider $\varphi : [a,b] \to \mathbb{R}$ defined by $\varphi(t) = \varphi(a) + \int_{\alpha_t} \eta$ where α_t is the restriction of α to $[a,t]$ and $\varphi(a)$ is so chosen that $\alpha(a)/\|\alpha(a)\| = (\cos\varphi(a), \sin\varphi(a))$. We claim that

$$\frac{\alpha(t)}{\|\alpha(t)\|} = (\cos\varphi(t), \sin\varphi(t)) \,,\ \forall t \in [a,b] \tag{18.31}$$

Let $t_0 = \sup\{t \in [a,b] : \alpha(t)/\|\alpha(t)\| = (\cos\varphi(t), \sin\varphi(t))\}$. By continuity the claim should be true for t_0 as well. Therefore it is enough to prove that $t_0 = b$.

Define $v = -\alpha(t_0)/\|\alpha(t_0)\|$ and $V = \mathbb{R}^2 - \{rv : r \geq 0\}$. And $\theta_V : V \to \mathbb{R}$ is defined in such a way that $v = (\cos\theta_v, \sin\theta_v)$ and $\theta_v \in [0, 2\pi)$. In other words, θ_V a real-valued function which give the angle of the ray from origin through the points.

Then $(\cos\theta_V(\alpha(t_0)), \sin\theta_V(\alpha(t_0))) = \alpha(t_0)/\|\alpha(t_0)\| = (\cos\varphi(t_0), \sin\varphi(t_0))$. Clearly, $\varphi(t_0) - \theta_V(\alpha(t_0)) = 2\pi m$ for some integer m.

Consider a δ neighbourhood of t_0 such that $\alpha(t) \in V$ for every point t in it. Then, α is smooth in that neighbourhood.

$$\frac{d}{dt}\left(\varphi(t) - \theta_V(\alpha(t))\right) = \eta(\dot{\alpha}(t)) = d\theta_V(\dot{\alpha}(t)) = 0$$

Thus, $\varphi(t) - \theta_V(\alpha(t)) = 2\pi m$ for every $t \in (t_0 - \delta, t_0 + \delta)$. Therefore, the claim is true. Suppose $t_0 < b$, then the claim is true for $t_0 + \epsilon \in (t_0 - \delta, t_0 + \delta)$ which is a contradiction to the choice of t_0 as least upper bound of all such values.

Thus, $(\cos\varphi(a), \sin\varphi(a)) = (\cos\varphi(b), \sin\varphi(b))$. Therefore, $\varphi(b)-\varphi(a) = 2\pi k$ for some integer k. And $\int_\alpha \eta = \int_a^b \eta(\dot\alpha(t))dt = \varphi(b) - \varphi(a) = 2\pi k$.

□

Winding Number The winding number of α is $k(\alpha) = \frac{1}{2\pi}\int_\alpha \eta$

Note : Winding number is the number of times α winds around origin.

18.12 Curvature of Surfaces

Normal Curvature The number $\kappa(\mathbf{v}) = L_p(\mathbf{v}) \cdot \mathbf{v}$ is the normal curvature of the n-surface S at point p in the direction $\mathbf{v}$.

If $\kappa(\mathbf{v}) > 0$, then S bends toward orientation at p. And if $\kappa(\mathbf{v}) < 0$, then S bends away from orientation at p.

We have, $\kappa : S_p \to S$. When $n = 1$, then $\kappa(\mathbf{v}) = \kappa(p)$. Clearly S_p is compact and κ attains its extrema, which are the eigen values of the Weingarten Map.

Normal Section The normal section of an n-surface S with orientation N is $\mathcal{N}(\mathbf{v}) \subset \mathbf{R}^{n+1}$ given by

$$\mathcal{N}(\mathbf{v}) = \{q \in \mathbf{R}^{n+1} : q = p + xv + yN(p),\ (x, y) \in \mathbb{R}^2\} \tag{18.32}$$

You take an apple and place a knife at any point on that apple in any direction tangent to its surface and cut it. The apple cuts into two parts. The 2-dimensional space of that cut, is a **Nomal Section** of that apple. For any $n \geq 2$, the normal section of an n-surface is always 2 dimensional.

$S \cap \mathcal{N}(\mathbf{v})$ is the intersection of both the surface and its normal section.

In above example, $S \cap \mathcal{N}(\mathbf{v})$ is plane curve traced by the skin around cross-section of that apple.

Theorem 18.12.1 (Components of $S \cap \mathcal{N}(\mathbf{v})$)**.** *Let S be an oriented n-surface and $\mathbf{v}$ is a tangent direction at p, $\mathbf{v} \in S_p$. Then there exists an open V containing p such that $S \cap \mathcal{N}(\mathbf{v}) \cap V$ is a plane curve. Moreover, the curvature at p of this curve is the normal curvature $\kappa(\mathbf{v})$ of S at p in the direction $\mathbf{v}$.*

Proof. Let $f : U \to \mathbb{R}$ such that $S = f^{-1}(c)$ and $\nabla f(q) \neq \mathbf{0}$, $\forall q \in S$. Let $\nabla f(q) = \left(q, \widetilde{\nabla f}(q)\right)$. Let $p \in S$ and $\mathbf{v} \in S_p$. Let $i : \mathbb{R}^2 \to \mathbb{R}^n$ defined by $i(x, y) = p + xv + yN(p)$. Clearly, the image of the function i is the normal section, $\mathcal{N}(\mathbf{v}) = i(\mathbb{R}^2)$.

Let V be the space spanned by v and $N(p)$. That is,

$$V = \left\{q \in U : \widetilde{\nabla f}(q) \cdot v \neq 0 \text{ OR } \widetilde{\nabla f}(q) \cdot N(p) \neq 0\right\} \tag{18.33}$$

Then $\nabla(f \circ i)(x, y) = \left(x, y, \widetilde{\nabla f}(i(x, y)) \cdot v, \widetilde{\nabla f}(i(x, y)) \cdot N(p)\right) \neq \mathbf{0}$ for every $(x, y) \in i^{-1}(V)$. Therefore, $C = i^{-1}(S \cap \mathcal{N}(\mathbf{v})) = (f \circ i)^{-1}(c) \ \cap \ i^{-1}(V)$ is a plane curve. In other words, $S \cap \mathcal{N}(\mathbf{v}) \cap V$ is a plane curve. □

Lemma 18.12.2 (Extrema of Curvature are Eigenvalue of Weingarten Map)**.** *Let V be a finite dimensional vector space with dot product and let $L : V \to V$ be a self-adjoint linear transformation on V. Let $S = \{v \in V : v \cdot v = 1\}$ and define $f : S \to \mathbb{R}$ by $f(v) = L(v) \cdot v$. Suppose f is stationary at $v_0 \in S$. Then $L(v_0) = f(v_0)v_0$.*

Proof. Since f is stationary at v_0, $f \circ \alpha)'(0) = 0$ for every parametrised curves in S with $\alpha(0) = v_0$. Let $v \perp v_0$. Then $\alpha(t) = (\cos t)v_0 + (\sin t)v$. Then,

$$\begin{aligned}
0 =&(f \circ \alpha)'(0)\\
=&\frac{d}{dt}\left(L(\alpha(t)) \cdot \alpha(t)\right)(0)\\
=&\frac{d}{dt}\left(\cos^2 t\ L(v_0) \cdot v_0 + 2\sin t \cos t\ L(v_0) \cdot v + \sin^2 t\ L(v) \cdot v\right)\\
=&\left(-\sin 2t\ L(v_0) \cdot v_0 + 2\cos 2t\ L(v_0) \cdot v + \sin 2t\ L(v) \cdot v\right)_{|t=0}\\
=&2L(v_0) \cdot v
\end{aligned}$$

Clearly, $L(v_0) \perp v$ for every $v \in v_0^{\perp}$. In other words, $L(v_0) \perp v_0^{\perp}$. Thus, v_0 can span the space $L(v_0)$. That is, $L(v_0) = \lambda v_0$ for some $\lambda \in \mathbb{R}$. Thus v_0 is an eigen vector of $L(v_0)$. And the corresponding eigen vector λ can be obtained, $\lambda = \lambda v_0 \cdot v_0 = L(v_0) \cdot v_0 = f(v_0)$. $\square$

Theorem 18.12.3. *Let V be a finite dimensional vector space with dot product and let $L : V \to V$ be a self-adjoint linear transformation on V. Then there exists an orthonormal basis for V consisting of eigen vectors of L.*

Proof. The proof is by mathematical induction. For $n = 1$, the theorem is true by lemma.Suppose that the theorem is true for $n = k$. It is enough to prove that theorem is then true for $n = k + 1$.

By lemma, there exists a unit vector $v_1 \in V$ which is an eigen vector of L. This is done by selecting v_1 such that $L(v_1) \cdot v_1 \geq L(v) \cdot v$ for every unit vector $v \in V$. Define $W = v_1^{\perp}$. Then $L(w) \cdot v_1 = w \cdot L(v_1) = \lambda w \cdot v_1 = 0$ for every $w \in W$. *(Remember :* L self-adjoint, $L(w) \cdot v = w \cdot L(v)$)

Thus the restriction of L, $L_{|W} : W \to W$ is surjective. And clearly self-adjoint. The dimension of W is k, thus by induction hypothesis there exists an orthonormal basis $\{v_2,\ v_3,\ \ldots,\ v_{k+1}\}$ for W with eigen vectors of $L_{|W}$. Clearly, these are eigen vectors of L and $\{v_1,\ v_2,\ \ldots, v_{k+1}\}$ is an orthonormal basis of V. $\square$

Principal Curvature Directions are the vectors in the orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ for S_p obtained as eigen vectors of the Weingarten Map. (Hint : $\mathbb{R}^n$ is a vector space with dot product and L_p is self-adjoint, linear transformation)

Principal Curvature are the eigen values of the Weingarten Map. That is, $k_i(p)$ is the eigen value of $L_p : S_p \to S_p$ such that $L_p(\mathbf{w}) \cdot \mathbf{v}_i = \kappa_i(p)\mathbf{w} \cdot \mathbf{v}_i$.

Theorem 18.12.4. *Let S be an oriented n-surface in $\mathbb{R}^{n+1}$. Let $pinS$ and $\{\kappa_1(p), \kappa_2(p), \ldots, \kappa_n(p)\}$ be the principal curvatures of S at p with corresponding orthogonal principal curvature directions $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$. Then the normal* *curvature $\kappa(\mathbf{v})$ in the direction $\mathbf{v} \in S_p$ is given by*

$$\kappa(v) = \sum_{i=1}^{n} \kappa_i(p)(\mathbf{v} \cdot \mathbf{v}_i)^2 = \sum_{i=1}^{n} \kappa_i(p) \cos^2 \theta_i \tag{18.34}$$

where θ_i is the angle between $\mathbf{v}$ and $\mathbf{v}_i$.

Proof. We have $v \in V$ and $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ is an orthonormal basis for V. Then,

$$\mathbf{v} = \sum_{i=1}^{n} (\mathbf{v} \cdot \mathbf{v}_i)\mathbf{v}_i = \sum_{i=1}^{n} \cos\theta_i \mathbf{v}_i$$

We have,

$$\begin{aligned}
\kappa(\mathbf{v}) =& L_p(\mathbf{v}) \cdot v \\
=& \sum_{i=1}^{n} \cos\theta_i L_p(\mathbf{v}_i) \cdot \mathbf{v} \\
=& \sum_{i=1}^{n} \kappa_i(\mathbf{v}_i)\mathbf{v}_i \cdot \mathbf{v} \\
=& \sum_{i=1}^{n} \kappa_i \cos^2 \theta_i
\end{aligned}$$

□

Direction Cosines Let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be an orthonomal basis for S_p. Then θ_i are the angles $\mathbf{v}$ makes with each vector in that basis. And direction cosines are the components of $\mathbf{v}$ in those directions. That is, $\cos\theta_i = \mathbf{v} \cdot \mathbf{v}_i$.

Quadratic Form Let V be a finite dimensional vector space with dot product and $L : V \to V$ be a self-adjoint, linear transformation on V. Then quadratic form associated with L is a $\mathcal{L} : V \to \mathbb{R}$ defined by $\mathcal{L}(v) = L(v) \cdot v$.

First Fundamental Form, $\mathscr{I}_p$ of surface S is the quadratic form associated with the identity transformation on S_p.

$$\mathscr{I}_p(v) = id(v) \cdot v = v \cdot v = \|v\|^2$$

Second Fundamental Form $\mathscr{S}_p$ of surface S is the quadratic form associated with the Weingarten Map.

$$\mathscr{S}_p(v) = L_p(v) \cdot v$$

A quadratic forms is

positive definite if $\mathscr{L}(v) > 0, \ \forall v \neq 0$.

negative definite if $\mathscr{L}(v) < 0, \ \forall v \neq 0$.

definite if either positive definite or negative definite.

indefinite if not definite.

positive semi-definite if $\mathscr{L}(v) \geq 0,\ \forall v \neq 0$.

negative semi-definite if $\mathscr{L}(v) \leq 0,\ \forall v \neq 0$.

semi-definite if either positive semi-definite or negative semi-definite.

For example, first fundamental form of any surface is postiive definite.

Theorem 18.12.5. *On each compact, oriented n-surface S in $\mathbb{R}^{n+1}$ there exists a point p such that the second fundamental form at p is definite.*

Proof. Let S^n be an n-sphere containing S and S^n touches S at p. Then $\mathscr{S}_p$ is definite.

Define $g : \mathbb{R}^n \to \mathbb{R}$ by $g(x_1, x_2, \ldots, x_n) = x_1^2 + x_2^2 + \cdots + x_n^2$. Since S is compact, g attains maximum at $p \in S$. By Lagrange multiplier theorem, $\nabla g(p) = \lambda \nabla f(p) = \mu N(p)$ where $\lambda, \mu \in \mathbb{R}$ and $\mu = \pm\lambda \|\nabla f(p)\|$. WLOG Suppose that $\mu < 0$. Then $\mu = -|\mu| = -\|\mu N(p)\| = -\|\nabla g(p)\| = -2\|p\|$. Thus, $N(p) = \frac{1}{\mu}\nabla g(p) = \frac{-1}{\|p\|}(p, p)$.

Let $\mathbf{v} \in S_p$ such that $\|\mathbf{v}\| = 1$. Let $\alpha : I \to S$ defined by $\dot{\alpha}(t_0) = \mathbf{v}$. Then $g \circ \alpha(t_0) \geq g \circ \alpha(t),\ \forall t \in I$.

$$
\begin{aligned}
0 \geq & \frac{d^2}{dt^2}(g \circ \alpha)(t_0) \\
= & \frac{d}{dt}(\nabla g(\alpha(t)) \cdot \dot{\alpha}(t))(t_0) \\
= & \frac{d}{dt}\left(2\alpha(t) \cdot \frac{d\alpha}{dt}(t)\right)(t_0) \\
= & 2\dot{\alpha}(t_0) \cdot \dot{\alpha}(t_0) + 2\alpha(t_0) \cdot \ddot{\alpha}(t_0) \\
= & 2\left(\|\dot{\alpha}(t_0)\|^2 + (\alpha(t_0), \alpha(t_0)) \cdot \ddot{\alpha}(t_0)\right) \\
= & 2\left(\|\mathbf{v}\|^2 + (p, p) \cdot \ddot{\alpha}(t_0)\right) \\
= & 2\left(1 - \|p\|\mathbf{N}(p) \cdot \ddot{\alpha}(t_0)\right) \\
= & 2\left(1 - \|p\|\kappa(\mathbf{v})\right)
\end{aligned}
$$

Thus, $0 \geq 2 - 2\|p\|\kappa(\mathbf{v}) \implies 2\|p\|\kappa(\mathbf{v}) \geq 2 \implies \kappa(\mathbf{v}) \geq \frac{1}{\|p\|},\ \forall \mathbf{v} \in S_p$. If S is oriented so that $\mu > 0$, then $\kappa(\mathbf{v}) \leq -\frac{1}{\|p\|}$. □

Gauss Kronecker Curvature, $K(p)$ of surface S in $\mathbb{R}^{n+1}$ at a point p is $K(p) = \det L_p$. It is equal to the product of the principal curvatures at p.

$$K(p) = k_1(p)k_2(p)\cdots k_n(p) \tag{18.35}$$

Mean Curvature $H(p)$ of of surface S in $\mathbb{R}^{n+1}$ at a point p is the average value of principal curvatures at p.

$$H(p) = \frac{1}{n}(\kappa_1(p) + \kappa_2(p) + \cdots + \kappa_n(p)) \tag{18.36}$$

Theorem 18.12.6. *Let S be an n-surface in $\mathbb{R}^{n+1}$. Let $p \in S$. Let $\mathbf{Z}$ be any non-zero normal vector field on S such that $\mathbf{N} = \mathbf{Z}/\|\mathbf{Z}\|$ and let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be any basis for S_p. Then*

$$K(p) = \frac{(-1)^n \det \begin{pmatrix} \nabla_{v_1}\mathbf{Z} \\ \nabla_{v_2}\mathbf{Z} \\ \vdots \\ \nabla_{v_n}\mathbf{Z} \\ \mathbf{Z}(p) \end{pmatrix}}{\|\mathbf{Z}(p)\|^n \det \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \\ \mathbf{Z}(p) \end{pmatrix}} \tag{18.37}$$

Proof. $\mathbf{Z} = \|\mathbf{Z}\|\mathbf{N}$.

$$\begin{aligned}
\det \begin{pmatrix} \nabla_{v_1}\mathbf{Z} \\ \nabla_{v_2}\mathbf{Z} \\ \vdots \\ \nabla_{v_n}\mathbf{Z} \\ \mathbf{Z}(p) \end{pmatrix} &= \det \begin{pmatrix} (\nabla_{v_1}\|\mathbf{Z}\|)\,\mathbf{N}(p) + \|\mathbf{Z}(p)\|\nabla_{v_1}\mathbf{N} \\ (\nabla_{v_2}\|\mathbf{Z}\|)\,\mathbf{N}(p) + \|\mathbf{Z}(p)\|\nabla_{v_2}\mathbf{N} \\ \vdots \\ (\nabla_{v_n}\|\mathbf{Z}\|)\,\mathbf{N}(p) + \|\mathbf{Z}(p)\|\nabla_{v_n}\mathbf{N} \\ \|\mathbf{Z}(p)\|\mathbf{N}(p) \end{pmatrix} \\
&= \|\mathbf{Z}(p)\|^n \det \begin{pmatrix} \nabla_{v_1}\mathbf{N} \\ \nabla_{v_2}\mathbf{N} \\ \vdots \\ \nabla_{v_n}\mathbf{N} \\ \|\mathbf{Z}(p)\|\mathbf{N}(p) \end{pmatrix} \\
&= (-1)^n \|\mathbf{Z}(p)\|^n \det \begin{pmatrix} L_p(\mathbf{v}_1) \\ L_p(\mathbf{v}_2) \\ \vdots \\ L_p(\mathbf{v}_n) \\ \mathbf{Z}(p) \end{pmatrix} \\
&= (-1)^n \|\mathbf{Z}(p)\|^n \det \begin{pmatrix} & \cdots & & 0 \\ & A^t & & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \\ \mathbf{Z}(p) \end{pmatrix} \\
&= (-1)^n \|\mathbf{Z}(p)\|^n \det A \det \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \\ \mathbf{Z}(p) \end{pmatrix}
\end{aligned}$$

$$=(-1)^n \|\mathbf{Z}(p)\|^n K(p) \det \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \\ \mathbf{Z}(p) \end{pmatrix}$$

where A is the matrix for L_p with respect to the orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ for S_p and A^t is its transpose. $\square$

local property is a property of S which is valid in the neighbourhood of a particular point.

global property is a property of S which is valid everywhere on S.

local theorem is a theorem on local behaviour of S.

global theorem is a theorem on global behaviour of S.

Theorem 18.12.7 (Global Characterisation of Surfaces with Definite Second Fundamental Form)**.** *Let S be a compact, connected, oriented n-surface in $\mathbb{R}^{n+1}$. Then $\forall p \in S,\ K(p) \neq 0 \iff \forall p \in S,\ \mathscr{S}_p$ is defінte.*

Proof. If $\mathscr{S}_p$ is definite then normal curvature $\kappa(\mathbf{v}) = \mathscr{S}_p(\mathbf{v})$ is nonzero in every direction $\mathbf{v} \in S_p$. Thus, all principal curvatures are nonzero and their product Gauss Kronecker Curvatures $K(p)$ is also nonzero.

Let $\kappa_1 \leq \kappa_2 \leq \ldots \kappa_n$. Let $\mathscr{S}_{p_0}$ be definite. Since every compact, connected, oriented n-surface S has a point p_0 at which its second fundamental form $\mathscr{S}_{p_0}$ is definite. Suppose $\mathscr{S}_{p_0}$ is positive definite. Then minimal principal curvature κ_1 is positive at p_0. Clearly, κ_i is nowhere zero and continuous. Thus, every principal curvature κ_i are positive everywhere on S. Therefore, their product $K(p)$ is positive everywhere on S. That is, $\mathscr{S}_p$ is postiive definite.

Suppose $\mathscr{S}_{p_0}$ is negative definite. Then the maximal principal curvature $\kappa_n(p)$ is negative at p_0. And κ_n is nonzero everywhere and continuous. Thus, κ_n is negative definite. Thus every principal curvature is negative definite. Thus Gauss Kronecker Curvature $K(p)$ is either everywhere negative or postive depending on the parity on n. Therefore, $K(p)$ is definite. $\square$

18.14 Parameterized Surfaces

tangent bundle is the set $T(U) = U \times \mathbb{R}^{n+1}$ where U is an open subset of $\mathbb{R}^{n+1}$ such that $\forall p \in U,\ \mathbb{R}^{n+1}_p \subset U \times \mathbb{R}^{n+1}$.

Definitions 18.14.1 (Differential of parametrisations)**.** Let $\varphi : U \to \mathbb{R}^{n+1}$ be parametrisation of a surface. Then its differential $d\varphi : T(U) \to T(\mathbb{R}^{n+1})$ is given by

$$d\varphi(\mathbf{v}) = (\varphi\dot{\circ}\alpha)(t_0) \tag{18.38}$$

Note : $d\varphi$ is independent of the choice of α.

Definitions 18.14.2 (Differential of surface parametrisation)**.** Let S be an n-surface in $\mathbb{R}^{n+1}$. Then $d\varphi_p$ is the restriction of $d\varphi$ to S_p.

There are two different meanings for $d\varphi$ depending on its domain.

1. $d\varphi : I \to \mathbb{R}$
2. $d\varphi : I \times \mathbb{R} \to \mathbb{R}^2$

$$d\varphi(t, u) = (\varphi(t), d\varphi(t, u)) \tag{18.39}$$

Author could have used dark font for this second version, $\boldsymbol{d\varphi}$. But he chose not to.

18.14.1 Parametrized Surfaces

parametrised n-surface is a smooth map $\varphi : U \to \mathbb{R}^n$ where U is a connected and open subset of $\mathbb{R}^n$.

regular surface A parametrised n-surface is regular if $d\varphi$ is non-singular.

parametrised 2-sphere $\varphi : U \to \mathbb{R}^3$ defined by $\varphi(\theta, \phi) = (r\cos\theta\sin\phi, r\sin\theta\sin\phi, r\cos\phi)$

parametrised n-plane Let $L : \mathbb{R}^n \to \mathbb{R}^{n+k}$ be non-singular linear map. Then $\varphi : U \to \mathbb{R}^{n+k}$ defined by $\varphi(p) = L(p) + w$ is a parametrised n-plane through w in $\mathbb{R}^{n+k}$.

cylinder over n-surface φ Let $\varphi : U \to \mathbb{R}^{n+k}$ be a parametrised n-surface in $\mathbb{R}^{n+k}$ where U is an open subset of $\mathbb{R}^n$. The cylinder over φ is an $(n+1)$-surface. $\tilde{\varphi} : U \times \mathbb{R} \to \mathbb{R}^{n+k+1}$ defined by

$$\tilde{\varphi}(u_1, u_2, \ldots, u_{n+1}) = (\varphi(u_1, u_2, \ldots, u_n), u_{n+1}) \tag{18.40}$$

Surface of revolution Let $\alpha : I \to \mathbb{R}^2$ be a parametrised curve in $\mathbb{R}^2$ whose image lies above x_1 axis. The surface obtained by revolving $\alpha(t) = (x_1(t), x_2(t))$ around x_1 axis is an 2-surface in $\mathbb{R}^3$.

$$\varphi : I \times \mathbb{R} \to \mathbb{R}^3 \text{ defined by } \varphi(t, \theta) = (x_1(t), x_2(t)\cos\theta, x_2(t)\sin\theta) \tag{18.41}$$

Torus in $\mathbb{R}^3$ Consider the circle $\alpha(\phi) = (a + b\cos\phi, b\sin\phi)$ with center at $(a, 0)$ with radius b $(b > a)$. Then the surface of revolution of α about x_2 axis,

$$\varphi(\phi, \theta) = (a + b\cos\phi\cos\theta, a + b\cos\phi\sin\theta, b\sin\phi) \tag{18.42}$$

Torus in $\mathbb{R}^4$ is a parametrised 2-surface $\varphi : \mathbb{R}^2 \to \mathbb{R}^4$ defined by

$$\varphi(\theta, \phi) = \{(\cos\theta, \sin\theta, \cos\phi, \sin\phi) : \theta, \phi \in \mathbb{R}\} \tag{18.43}$$

Möbius Band is a parametrised 2-surface $\varphi : I \times \mathbb{R} \to \mathbb{R}^3$ defined by

$$\varphi(t, \theta) = \left(\left(1 + t\cos\frac{\theta}{2}\right)\cos\theta, \left(1 + t\cos\frac{\theta}{2}\right)\sin\theta, t\sin\frac{\theta}{2}\right) \tag{18.44}$$

18.14.2 Vector Field along φ

Definitions 18.14.3. Let $\varphi : U \to \mathbb{R}^{n+k}$ be a smooth function. A vector field along φ is a map $\mathbf{X}$ which assigns a vector $\mathbf{X}(p) \in \mathbb{R}^{n+k}_{\varphi(p)}$ to each point $p \in U$.

Definitions 18.14.4. A vector field along φ is a tangent vector field along φ if there exists a vector field $\mathbf{Y}$ on U such that $\mathbf{X}(p) = d\varphi_p(\mathbf{Y}(p))$.

Definitions 18.14.5. Let $\varphi : U \to \mathbb{R}^{n+k}$ be a smooth map where U is an open subset of $\mathbb{R}^n$. The coordinate vector fields along φ are the tangent fields $\mathbf{E}_i$ defined by

$$\mathbf{E}_i(p) = d\varphi_p(p, 0, 0, \ldots, 1, \ldots, 0) \tag{18.45}$$

Theorem 18.14.1 (Basis for Tangent Space Image $d\varphi_p$)**.** *Let $\varphi : U \to \mathbb{R}^{n+k}$ be smooth. Then the set of all coordinate vector fields $\{\mathbf{E}_i(p) : i = 1, 2, \ldots, n\}$ form a basis for the tangent space image $d\varphi_p$.*

Definitions 18.14.6. Let $\varphi : U \to \mathbb{R}^{n+k}$ be smooth. And $\mathbf{X}$ be a smooth vector field along φ. Let $p \in U$ and $\mathbf{v} \in \mathbf{R}^n_p$. The derivative $\nabla_{\mathbf{v}}\mathbf{X} \in \mathbb{R}^{n+k}_{\varphi(p)}$ defined by

$$\nabla_{\mathbf{v}}\mathbf{X} = \left(\varphi(p), \left(\frac{d}{dt} X \circ \alpha\right)(t_0)\right) \tag{18.46}$$

Definitions 18.14.7 (orientation vector field)**.** Let $\varphi : U \to \mathbb{R}^{n+1}$ be smooth parametrised n-surface in $\mathbb{R}^{n+1}$. Let $p \in U$. An orientation vector field along φ is the unique vector field $\mathbf{N}$ along φ such that $\mathbf{N}(p)$ are unit vectors and $\mathbf{N}(p) \perp d\varphi_p(\mathbf{v})$. And $\mathbf{N}$ is consistently oriented if the following determinant is positive.

$$det \begin{pmatrix} \mathbf{E}_1(p) \\ \mathbf{E}_2(p) \\ \vdots \\ \mathbf{E}_n(p) \\ \mathbf{N}(p) \end{pmatrix} > 0 \tag{18.47}$$

18.14.3 Weingarten Map

Definitions 18.14.8 (Weingarten map)**.** The Weingarten map is the linear function $L_p : \text{Image } d\varphi_p \to \text{Image } d\varphi_p$ defined by $L_p(d\varphi(\mathbf{v})) = -\nabla_{\mathbf{v}}\mathbf{N}$.

L_p is self-adjoint $L_p(\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot L_p(\mathbf{w})$

principal curvatures are the eigen values of L_p

principal curvature directions are the unit eigen vectors of L_p

Gauss-Kronecker Curvature is the determinant.

Mean Curvature is $1/n$ times its trace.

Subject 19

ME800402 Algorithmic Graph Theory

19.1 Networks

19.1.1 An Introduction to Networks

Definitions 19.1.1. A **network** N is a digraph D with two special vertices source s and sink t together with a capacity function $c : E(D) \to \mathbb{Z}$ such that for every arc $a = (u, v)$ of the digraph, $c(u, v)$ is non-negative.

Remark. Mathematical Modeling using Network,

1. There is no restriction on indegree/outdegree of source/sink vertices of the digraph D of a network N.

2. Applications of Network : Transportation problem.

$c(u, v)$ is the capacity of the arc (u, v) of D

$N^+(x) = \{y \in V(D) : (x, y) \in E(D)\}$ is the out-neighbourhood of x.

$N^-(x) = \{y \in V(D) : (y, x) \in E(D)\}$ is the in-neighbourhood of x.

Definitions 19.1.2. A **flow f in a network N** is function $f : E(D) \to \mathbb{Z}$ such that 1. each edge satisfies capacity constraint and 2. each vertex except source and sink satisfies conservation equation.

capacity constraint

$$0 \leq f(a) \leq c(a) \text{ for every arc } a \in V(D) \tag{19.1}$$

conservation equation

$$\sum_{y \in N^+(x)} f(x, y) = \sum_{y \in N^-(x)} f(y, x), \ \forall \text{vertex } x \in V(D) - \{s, t\} \tag{19.2}$$

net flow out of x

$$\sum_{y \in N^+(x)} f(x, y) - \sum_{y \in N^-(x)} f(y, x)$$

net flow into x

$$\sum_{y\in N^-(x)} f(y,x) - \sum_{y\in N^+(x)} f(x,y)$$

Definitions 19.1.3. The **flow f in a network N** is the net flow out of source s.

Remark. 1. net flow out of/into $x \in V(D) - \{s,t\}$ is zero.

2. Without loss of generality[1], underlying digraph is always assymetric.

(X,Y) $= \{(x,y) \in E(D) : x \in X,\ y \in Y\}$.
Let X, Y be non-empty subsets of $V(D)$ such that X, Y are disjoint. Then (X,Y) is the set of all arcs from X to Y.

flow from X **to** Y is the sum of flow on each arc in (X,Y)

$$f(X,Y) = \sum_{(x,y)\in(X,Y)} f(x,y) \tag{19.3}$$

capacity of the partition (X,Y) is the total capacity of arcs in (X,Y)

$$c(X,Y) = \sum_{(x,y)\in(X,Y)} c(x,y) \tag{19.4}$$

cut Let $P \subset V(D)$ such that $s \in P$ and $t \notin P$ and $\bar{P} = V(D) - P$, then $(P,\bar{P})$ is a cut.

flow from P **to** $\bar{P}$ is the sum of flow on each arc in $(P,\bar{P})$.

$$f(P,\bar{P}) = \sum_{(x,y)\in(P,\bar{P})} f(x,y) \tag{19.5}$$

flow from $\bar{P}$ **to** P is the sum of flow on each arc in $(\bar{P},P)$

$$f(\bar{P},P) = \sum_{(x,y)\in(\bar{P},P)} f(x,y) \tag{19.6}$$

capacity of the cut $(P,\bar{P})$ is the total capacity of the arcs in $(P,\bar{P})$

$$c(P,\bar{P}) = \sum_{(x,y)\in(P,\bar{P})} c(x,y) \tag{19.7}$$

Theorem 19.1.1. *For any cut $(P,\bar{P})$, the flow in N is $f(N) = f(P,\bar{P}) - f(\bar{P},P)$.*

Synopsis. The net flow out of source s is the flow $f(N)$ in the network N. Let $(P,\bar{P})$ be a cut of N, then $s \in P$ and $t \notin P$. Suppose $P = \{s\}$, then the theorem is true. Suppose P is not singleton, then for each vertex $x \in P,\ x \neq s$, the net flow out of x is zero by flow conservation equation. And flow between vertices in P cancels out each other. Thus adding net flow out of each vertex in P, will be same as the net flow out of source which is the flow in the network, $f(N)$.

[1] If underlying digraph of a network is symmetric, then by replacing an arc (u,v) with a new vertex w and two arcs $(u,w),(w,v)$ gives an assymetric digraph.[Gray Chartrand,]pp.131

Proof.

$$\text{Flow, } f = \sum_{y \in N^+(s)} f(s,y) - \sum_{y \in N^-(s)} f(y,s) \tag{19.8}$$

By conservation equation, we have $\forall x \in P,\ x \neq s$,

$$\sum_{y \in N^+(x)} f(x,y) - \sum_{y \in N^-(x)} f(y,x) = 0 \tag{19.9}$$

By above equations,

$$\begin{aligned} \text{Flow, } f &= \sum_{x \in P} \sum_{y \in N^+(x)} f(x,y) - \sum_{x \in P} \sum_{y \in N^-(x)} f(y,x) \\ &= \sum_{(x,y) \in (P,\bar{P})} f(x,y) - \sum_{(y,x) \in (\bar{P},P)} f(y,x) \end{aligned} \tag{19.10}$$

□

Corollary 19.1.1.1. *Flow cannot exceed the capacity of any cut $(P, \bar{P})$. Further, $f(N) \leq \min c(P, \bar{P})$.*

Synopsis. Let $(P, \bar{P})$ be a cut in network N, then by theorem the flow $f(N) =$ flow from P to $\bar{P}$ - flow from $\bar{P}$ to P. Since the flow from $\bar{P}$ to P is non-negative, $f(N) \leq$ flow from P to $\bar{P}$. Clearly, $f(x,y) \leq c(x,y)$ by the capacity constaint. Thus $f(N) \leq f(P, \bar{P}) \leq c(P, \bar{P}) \leq \min c(P, \bar{P})$.

Proof.

$$\begin{aligned} f(N) &= \sum_{(x,y) \in (P,\bar{P})} f(x,y) - \sum_{(y,x) \in (\bar{P},P)} f(y,x) \\ &\leq \sum_{(x,y) \in (P,\bar{P})} f(x,y) = f(P, \bar{P}) \\ &\leq \sum_{(x,y) \in (P,\bar{P})} c(x,y) = c(P, \bar{P}), \quad \because \forall x, y \in V(D),\ f(x,y) \leq c(x,y) \\ &\leq \min c(P, \bar{P}) \end{aligned}$$

□

Corollary 19.1.1.2. *In a network N flow is the net flow into the sink of N.*

Synopsis. Let $\bar{P} = \{t\}$, then by theorem $f(N)$ is the net flow into the sink.

Proof. Suppose $P = V(D) - \{t\}$. Then by theorem, we have

$$\begin{aligned} f(N) &= \sum_{(x,y) \in (P,\bar{P})} f(x,y) - \sum_{(y,x) \in (\bar{P},P)} f(y,x) \\ &= \sum_{x \in N^-(t)} f(x,t) - \sum_{x \in N^+(t)} f(t,x) \end{aligned}$$

□

Remark. Exercise 5.1

4. Let N be a network with underlying digraph D which has a vertex $v \in V(D) - \{s, t\}$ with zero indegree. Clearly the flow into v is zero. Thus flow out of v is also zero by flow conservation equation. Let N' be the network obtained from N by deleting the vertex v. Then $f(N) = f(N')$.

19.1.2 The Max-Flow Min-Cut Theorem

maximum flow A flow f in network N is maximum flow in N, if $f(N) \geq f'(N)$ for each flow f' in N.

minimum cut A cut $(P, \bar{P})$ in network N is minimum cut of N, if $c(P, \bar{P}) \leq c(X, \bar{X})$ for each cut $(X, \bar{X})$ in N.

f-unsaturated Let f be a flow in network N with underlying digraph D, and $Q = u_0, a_1, u_1, a_2, \cdots, u_{n-1}, a_n, u_n$ be a semipath in D such that every forward arc $a_i = (u_{i-1}, u_i)$ has flow not upto its capacity, $f(a_i) < c(a_i)$ and every reverse arc $a_i = (u_i, u_{i-1})$ has some positive flow in it, $f(a_i) > 0$

f-augmenting semipath Let f be a flow in a network N with underlying digraph D. Suppose semipath $Q = s, a_1, u_1, a_2, \cdots, u_{n-1}, a_n, t$ (from source to sink) is f-unsaturated, then Q is an f-augmenting semipath.

Theorem 19.1.2. *Let f be a flow in a network N with underlying digraph D. The flow f is maximum in N iff there is no f-augmenting semipath in D.*

Synopsis. Suppose Q is an f-augmenting semipath in D, then there exists a flow f^* in N such that $f(N) + \Delta = f^*(N)$. Therefore, f is not a maximum flow in N. Suppose there is no f-augmenting semipath in D, then there exists a cut $(P, \bar{P})$ such that $f(a) = c(a)\ \forall a \in (P, \bar{P})$ and $f(a) = 0\ \forall a \in (\bar{P}, P)$. Suppose f^* in a maximum flow in N, then $f(N) \leq f^*(N) \leq c(P, \bar{P}) = f(N)$.

Proof. Let f be a flow in a network N with underlying digraph D and $Q = s, a_1, u_1, a_2, u_2, \cdots, u_{n-1}, a_n, t$ be an f-augmenting semipath in D.

$$\text{define } \Delta_i = \begin{cases} c(a_i) - f(a_i) \text{ for every forward arc } a_i \in Q, \\ f(a_i) \text{ for every reverse arc } a_i \in Q, \end{cases}$$

Define $\Delta = \min\{\Delta_i\}$. Also define $f^* : E(D) \to \mathbb{Z}$ such that

$$f^*(a_i) = \begin{cases} f(a) + \Delta, \text{ for every forward arc } a_i \in Q, \\ f(a) - \Delta, \text{ for every reverse arc } a_i \in Q, \\ f(a_i), \text{ for every arc of } D \text{ which are not in } Q. \end{cases}$$

Since Q is an f-augmenting semipath in D, $\Delta > 0$ and $f(N) + \Delta = f^*(N)$.

Clearly $f(N) < f^*(N)$, and it is enough to show that f^* is a flow in N. f^* is a flow if it satisfies 1. capacity constraint and 2. conservation equation. For any arc $a_i \notin Q$, $f^*(a_i) = f(a_i) \leq c(a_i)$. Suppose $a_i \in Q$. If $a_i = (u_{i-1}, u_i)$, a_i is a forward arc and we have $f^*(a_i) = f(a_i) + \Delta \leq f(a_i) + \Delta_i = f(a_i) + c(a_i) - f(a_i) = c(a_i)$. If $a_i = (u_i, u_{i-1})$, then a_i is a reverse arc and we have $f^*(a_i) = \Delta \leq \min\{\Delta_i\} = \Delta_i = c(a_i)$. Thus f^* satisfies capactity constraint on every arc of D.

Let $x \in V(D) - \{s, t\}$. Suppose $x \notin Q$,

$$\begin{aligned}\text{Net flow out of } x &= \sum_{y \in N^+(x)} f^*(x,y) - \sum_{y \in N^-(x)} f^*(y,x) \\ &= \sum_{y \in N^+(x)} f(x,y) - \sum_{y \in N^-(x)} f(y,x) \\ &= 0\end{aligned}$$

Suppose $x = u_i \in Q$, then Q has two arc having vertex x say, a_{i-1}, and a_i. There are four possibilities for these two arcs,

1. Both a_{i-1}, a_i are forward arcs.
2. Arc a_{i-1} is forward, but arc a_i is reverse.
3. Arc a_{i-1} is reverse, but arc a_i is forward.
4. Both a_{i-1}, a_i are reverse arcs.

Case 1 $a_{i-1} = (u_{i-1}, u_i)$ and $a_i = (u_i, u_{i+1})$.

$$\begin{aligned}\text{Net flow out of } x &= \sum_{y \in N^+(x)} f^*(x,y) - \sum_{y \in N^-(x)} f^*(y,x) \\ &= \sum_{\substack{y \in N^+(x) \\ y \neq u_{i+1}}} f^*(x,y) + f^*(u_i, u_{i+1}) - \left(\sum_{\substack{y \in N^-(x) \\ y \neq u_{i-1}}} f^*(y,x) + f^*(u_{i-1}, u_i) \right) \\ &= \sum_{\substack{y \in N^+(x) \\ y \neq u_{i+1}}} f(x,y) + f(u_i, u_{i+1}) + \Delta - \left(\sum_{\substack{y \in N^-(x) \\ y \neq u_{i-1}}} f(y,x) + f(u_{i-1}, u_i) \right) - \Delta \\ &= \sum_{y \in N^+(x)} f(x,y) - \sum_{y \in N^-(x)} f(y,x) \\ &= 0\end{aligned}$$

Case 2 $a_{i-1} = (u_{i-1}, u_i)$ and $a_i = (u_{i+1}, u_i)$.

$$\begin{aligned}\text{Net flow out of } x &= \sum_{y \in N^+(x)} f^*(x,y) - \sum_{y \in N^-(x)} f^*(y,x) \\ &= \sum_{\substack{y \in N^+(x) \\ y \neq u_{i+1}, u_{i-1}}} f^*(x,y) + f^*(u_i, u_{i+1}) + f^*(u_i, u_{i-1}) - \sum_{y \in N^-(x)} f^*(y,x) \\ &= \sum_{\substack{y \in N^+(x) \\ y \neq u_{i+1}, u_{i-1}}} f(x,y) + f(u_i, u_{i+1}) + \Delta + f(u_i, u_{i-1}) - \Delta - \sum_{y \in N^-(x)} f(y,x) \\ &= \sum_{y \in N^+(x)} f(x,y) - \sum_{y \in N^-(x)} f(y,x) \\ &= 0\end{aligned}$$

Case 3 $a_{i-1} = (u_i, u_{i-1})$ and $a_i = (u_i, u_{i+1})$.

$$\begin{aligned}
\text{Net flow out of } x &= \sum_{y \in N^+(x)} f^*(x,y) - \sum_{y \in N^-(x)} f^*(y,x) \\
&= \sum_{y \in N^+(x)} f^*(x,y) - \left(\sum_{\substack{y \in N^-(x) \\ y \neq u_{i-1}, u_{i+1}}} f^*(y,x) + f^*(u_{i-1}, u_i) + f^*(u_{i+1}, u_i) \right) \\
&= \sum_{y \in N^+(x)} f^*(x,y) - \left(\sum_{\substack{y \in N^-(x) \\ y \neq u_{i-1}, u_{i+1}}} f(y,x) + f(u_{i-1}, u_i) + \Delta + f(u_{i+1}, u_i) - \Delta \right) \\
&= \sum_{y \in N^+(x)} f(x,y) - \sum_{y \in N^-(x)} f(y,x) \\
&= 0
\end{aligned}$$

Case 4 $a_{i-1} = (u_i, u_{i-1})$ and $a_i = (u_{i+1}, u_i)$.

$$\begin{aligned}
\text{Net flow out of } x &= \sum_{y \in N^+(x)} f^*(x,y) - \sum_{y \in N^-(x)} f^*(y,x) \\
&= \sum_{\substack{y \in N^+(x) \\ y \neq u_{i-1}}} f^*(x,y) + f^*(u_i, u_{i-1}) - \left(\sum_{\substack{y \in N^-(x) \\ y \neq u_{i+1}}} f^*(y,x) + f^*(u_{i+1}, u_i) \right) \\
&= \sum_{\substack{y \in N^+(x) \\ y \neq u_{i-1}}} f(x,y) + f(u_i, u_{i-1}) - \Delta - \left(\sum_{\substack{y \in N^-(x) \\ y \neq u_{i+1}}} f(y,x) + f(u_{i+1}, u_i) \right) + \Delta \\
&= \sum_{y \in N^+(x)} f(x,y) - \sum_{y \in N^-(x)} f(y,x) \\
&= 0
\end{aligned}$$

Therefore, f^* is a flow on N. We have $f(N) < f^*(N)$. Thus f is not maximum flow in N due to the existence of an f-augmenting semipath in D.

Conversely, assume that there is no f-augmenting semipath in D. Now, we construct a cut $(P, \bar{P})$ of N. Let P be the set of all vertices $x \in V(D)$ such that there is an f-unsaturated $s - x$ semipath in D. Trivially, $s \in P$. And $t \notin P$ since there are no f-augmenting semipath in D. [2] Clearly, $(P, \bar{P})$ is a cut of the network N.

We claim that $c(P, \bar{P}) = f(N)$. Suppose there is a forward arc $(x, y) \in (P, \bar{P})$, then flow in it is saturated. If $f(x, y) < c(x, y)$, then there is an f-unsaturated $s - y$ semipath in D. ie, $s - x$ semipath + arc (x, y). Thus every forward arc $(x, y) \in (P, \bar{P})$ is saturated. Suppose there is a reverse arc $(y, x) \in$

[2] An f-augmenting semipath is an f-unsaturated $s - t$ semipath in D.

$(\bar{P}, P)$, then there is no flow in it(saturated reversed arc). If $f(y, x) > 0$, then there is an f-unsaturated $s - y$ semipath in D. ie, $s - x$ semipath + arc (y, x). Thus every reverse arc $(y, x) \in (\bar{P}, P)$ is saturated. And we have,

$$\sum_{(x,y)\in(P,\bar{P})} f(x, y) = \sum_{(x,y)\in(P,\bar{P})} c(x, y)$$

$$\sum_{(y,x)\in(\bar{P},P)} f(y, x) = 0$$

$$\begin{aligned} f(N) &= \sum_{(x,y)\in(P,\bar{P})} f(x, y) - \sum_{(y,x)\in(\bar{P},P)} f(y, x) \\ &= \sum_{(x,y)\in(P,\bar{P})} c(x, y) \\ &= c(P, \bar{P}) \end{aligned}$$

Suppose f^* is maximum flow in network N and $(X, \bar{X})$ is minimum cut of N. Then $f(N) \le f^*(N)$. Thus we have, $f(N) \le f^*(N) \le c(X, \bar{X}) \le c(P, \bar{P}) = f(N)$. Therefore, $f(N) = f^*(N)$. ie, the flow f is maximum in network N if there are no f-augmenting semipaths in D. □

Theorem 19.1.3 (maximum-flow, min-cut)**.** *In every network, the value of maximum flow equals capacity of minimum cut.*

Proof. Suppose flow f in network N in maximum, then by previous theorem there is no f-augmenting semipath in D. And $f(N) \le c(X, \bar{X})$ for any cut $(X, \bar{X})$ in N. We can construct a cut $(P, \bar{P})$ in N such that $f(N) = c(P, \bar{P})$. Let P be the set of all vertices x in D such that there is an f-unsaturated $s - x$ semipath in D. Clearly $s \in P$ and $t \notin P$. Also $f(P, \bar{P}) = c(P, \bar{P})$ and $f(\bar{P}, P) = 0$. Then the cut $(P, \bar{P})$ is minimum cut of N. Suppose there is a cut $(X, \bar{X})$ such that $c(X, \bar{X}) < c(P, \bar{P})$. Then $f(N) = f(P, \bar{P}) - f(\bar{P}, P) = c(P, \bar{P}) < c(X, \bar{X})$ which is a contradiction. Therefore, the value of maximum flow equals capacity of minimum cut. □

Remark. Exercise 5.2

1. Suppose $(X, \bar{X})$ is a cut of N such that $f(a) = c(a)$, $\forall a \in (X, \bar{X})$ and $f(a) = 0$, $\forall a \in (\bar{X}, X)$. By the definition of cut, $s \in X$ and $t \in \bar{X}$. Thus there is no f-augmenting semipath in D. Suppose there is an f-augmenting semipath Q in D, then there is either (a) a forward arc $(x, y) \in (X, \bar{X})$ such that $f(x, y) < c(x, y)$ or (b) a reverse arc $(y, x) \in (\bar{X}, X)$ such that $f(y, x) > 0$ which is a contradition. Therefore, the flow $f(N)$ is maximum and the given cut $(X, \bar{X})$ is minimum as shown in the proof of the maximum-flow min-cut theorem.

3. The algorithm suggested in the hint of this exercise won't work if two subnetworks have a common arc such that the direction of flow in which is not consistent. Suppose, the generalized network is not supposed to have any common arcs. Then construct subnetworks for each pair (s, t) with all those arcs which are on some $s - t$ semipath. Define subnetwork capacity function $c'(a) = c(a)$ for every arc in N'.

Let N be a generalized network with set of sources S and set of sinks T. A flow in N is maximum if there is not f-augmenting $s-t$ semipath for each pair $(s,t) \in S \times T$.

19.1.3 A max-flow min-cut algorithm

Theorem 19.1.4. *Let N be a network with underlying digraph D, source s, sink t, capacity function c and flow f. Let D' be the digraph with same vertex set as D and arc set defined by $E(D') = \{(x,y) : (x,y) \in E(D),\ c(x,y) > f(x,y)$ or $(y,x) \in E(D),\ f(y,x) > 0\}$. ie, D' has only the unsaturated arcs of D. Then D' has an $s-t$ directed path iff D has an f-augmenting semipath. Moreover, shortest $s-t$ path in D' has the same length as shortest f-augmenting semipath in D.*

Synopsis. Each directed $s-t$ path in D' has respective f-augmenting semipath in D and vice versa. Clearly, they have the same length.

Proof. Let N be a network with underlying digraph D, capacity c and flow f. Let D' be the digraph with vertex set $V(D') = V(D)$ and arc set $E(D') = \{(x,y) : \text{either } (x,y) \text{ or } (y,x) \text{ is unsaturated in } N\}$.

Suppose D' has a directed $s-t$ path $Q' : s, u_1, u_2, \cdots, u_{n-1}, t$. Then by the construction of D', for each $u_i \in Q$, there exists an f unsaturated arc a_i in D. ie, either forward arc $a_i = (u_{k-1}, u_k)$ such that $f(u_{k-1}, u_k) < c(u_{k-1}, u_k)$ or reverse arc $a_i = (u_k, u_{k-1})$ such that $f(u_k, u_{k-1}) > 0$. Therefore, we have an $s-t$ semipath $Q : s, a_1, u_1, a_2, \cdots, u_{n-1}, a_n, t$ in D such that Q is an f-augmenting semipath since every arc in Q is f-unsaturated. Clearly, Q, Q' are of the same length.

Conversely, suppose that the digraph D has an f-augmenting semipath $Q : s, a_1, u_1, a_2, \cdots, u_{n-1}, a_n, t$. Then each arc $a_i \in Q$ are f-unsaturated and by the construction of D', there exists a directed $s-t$ path $Q' = s, u_1, u_2, \cdots, u_{n-1}, t$ in D'. And Q, Q' are of the same length.

There is a one-one correspondence betweeen the directed $s-t$ paths in D' and f-augmenting semipaths in D. Clearly, they have the same length. Thus shortest directed $s-t$ path in D and shortest f-augmenting semipath in D' are of the same length. □

saturation arc of N with respect to the flow f is an arc a_j in an f-augmenting semipath Q with $\Delta_j = \Delta$.

augmentation path is an f-augmenting semipath Q in D.

Algorithm 19.1.1 (max-flow min-cut)**.** *An algorithm to find maximum flow and minimum cut of a network N with underlying digraph D, source s, sink t, capacity function c and initial flow f.*

1. *Construct digraph D' with vertex set $V(D') = V(D)$ and arc set $E(D') = \{(x,y) : (x,y) \in E(D) \& f(x,y) < c(x,y)$ or $(y,x) \in E(D) \& f(y,x) > 0\}$*

2. *Find (shortest) s—t directed path in D' using Moore's breadth first search(BFS) algorithm. If D' doesn't have an $s-t$ path, then proceed to step 5. Otherwise, let $Q' : s, u_1, u_2, \cdots, u_{n-1}, t$ be a (shortest) $s-t$ path in D'.*

3. *Let $Q : s, a_1, u_1, a_2, \cdots, u_{n-1}, a_n, t$ be the respective semipath in D such that $f(a_j) < c(a_j)$ for forward arcs and $f(a_i) > 0$ for reverse arcs. Let $\Delta_j = c(a_j) - f(a_j)$ for forward arcs and $\Delta_j = f(a_j)$ for reverse arcs. And let $\Delta = \min\{\Delta_j\}$. And augment flow f by Δ ie, $f(a_j) \leftarrow f(a_j) + \Delta$ for forward arcs and $f(a_j) \leftarrow f(a_j) - \Delta$ for reverse arcs.*

4. *Goto step 1 (Proceed with new flow f and find whether there are any directed $s-t$ paths in D'. If any, augment the flow along the new augmentation path Q by saturating the flow along the saturation arc.)*

5. *There is no $s-t$ directed path in D'. Thus there is no f-augmenting semipath in D. Therefore the flow f in N is maximum. Let P be the set of all vertices in D' with non-zero breadth first index(bfi) from Moore's BFS algorithm applied in step 2. $(P, \bar{P})$ is minimum cut of N.*

Remark. Validity of the algorithm is proved in the previous theorem.

19.1.5 Connectivity and Edge-Connectivity

edge cutset is the set U subset of $E(G)$ such that $G - U$ is disconnected.

vertex cutset is the set S subset of $V(G)$ such that $G - S$ is disconnected.

edge connectivity $\lambda(G)$ is the minimum cardinality of all edge cutsets of G.

connectivity $\kappa(G)$ is the minimum cardinality of all vertex cutsets of G.

Theorem 19.1.5. *For every graph G, $\kappa(G) \leq \lambda(G) \leq \delta(G)$*

Proof. Suppose graph G is disconnected then $\kappa(G) = \lambda(G) = \delta(G) = 0$. Let G be a connected graph. Then G has at least one vertex v with degree $\delta(G)$. Therefore $\lambda(G) \leq \delta(G)$ since edges incident with v form an edge cutset of G and $\lambda(G)$ is the cardinality of all edge cutsets.

Let G be a graph with edge connectivity $\lambda(G) = c$. Let U be a edge cutset with cardinality c and let edge $uv \in U$. Construct a set of vertices $S \subset V(G)$ such that (S is of minimal cardinality and) for each edge in U other uv, S has a vertex incident with it. Cardinality of S is atmost $c - 1$, since we can select one vertex each for each edge in U other than u, v. If $G - S$ is a disconnected graph, then $\kappa(G) < \lambda(G)$. Suppose $G - S$ is a connected graph, then delete a non-pendent vertex u or v from $G - S$, say v. Since $G - S$ is a connected graph with a singleton edge cutset, $\{uv\}$. We have a vertex cutset $S \cup \{v\}$ of G. Therefore, $\kappa(G) \leq c = \lambda(G)$. ☐

Theorem 19.1.6. *If G is a graph of diameter 2, then $\lambda(G) = \delta(G)$*

n-edge connected G is n-edge connected if $\lambda(G) \geq n$.

n **connected** G is n-connected if $\kappa(G) \geq n$.

Theorem 19.1.7. *Let G be a graph of order p and n be an integer such that $1 \leq n \leq p-1$. If $\delta(G) \geq \frac{p+n-2}{2}$, then G is n-connected.*

connection number $c(G)$ is the smallest integer such that $2 \leq c(G) \leq p$ and every subgraph of order n in G is connected.

l-connectivity $\kappa_l(G)$ is minimum number of vertices whose removal will produce a disconnected graph with at least l components or a graph with fewer than l vertices.

(n, l)-connected A graph G is (n, l)-connected if $\kappa_l(G) \geq n$.

Remark. Exercises 5.5

1. $\lambda(K_{m,n}) = \kappa(K_{m,n}) = m$

8. $c(K_p) = 2$, $c(K_{m,n}) = n+1$, $c(C_p) = p-1$
   Every two vertices of complete graph of order p are adjacent. For complete bi-partitie graph $K_{m,n}$ such that $1 \leq m \leq n$, there exists a totally disconnected subgraph of order n. Therefore $c(K_{m,n}) \geq n+1$. And with $n+1$ vertices, both partitions have at least two vertices each and therefore the graph is connected and $c(K_{m,n}) \leq n+1$. For cycle C_p, any subgraph is disconnected if two non-adjacent vertices are deleted. Therefore $c(C_p) \geq p-1$. And C_p remains connected even after deletion of any vertex, therefore $c(C_p) \leq p-1$.

9. $$\delta(G) \geq \frac{p + (l-1)(n-2)}{l} \implies \kappa_l(G) \geq n$$

19.1.6 Menger's Theorem

Theorem 19.1.8. *For a non-trivial graph G, $\lambda(u, v) = M'(u, v)$ for every pair (u, v) of vertices of G.*

Corollary 19.1.8.1. *Graph G is n-edge connected iff every two vertices of G are connected by at least n edge disjoint paths.*

Theorem 19.1.9. *For every pair of non-adjacent vertices u, v in graph G, $\kappa(u, v) = M(u, v)$.*

Corollary 19.1.9.1. *Graph G is n-connected iff every pair of vertices of G are connected by at least n internally disjoint paths.*

Algorithm 19.1.2 (connectivity $\kappa(G)$)**.** .

1. *If degree of every vertex is $p-1$, then output $\kappa = p-1$ and stop. Otherwise, continue.*

2. *If G is disconnected, output $\kappa = 0$ and stop. Otherwise, continue.*

3. $\kappa \leftarrow p$

4. $i \leftarrow 0$

5. *If $i \leq \kappa$, then $i \leftarrow i+1$ and continue. Otherwise, output κ and stop.*

6. $j \leftarrow i + 1$

7. (1) *If $j = p + 1$, then return to step 5. Otherwise continue.*

   (2) *If $v_i v_j \notin E(G)$, construct network N with digraph D as follows : for each vertex $v \in V(G)$, there are two vertices $v', v'' \in V(D)$ and an arc $(v', v'') \in E(D)$. And for each edge $uv \in E(G)$, there are two arcs $(u'', v), (v'', u) \in E(D)$. The capacity function is given by, $c(v', v'') = 1$ for every $v \in V(G)$ and $c(a) = \infty$ for every other arc in D. Set source $s = v_i''$ and sink $t = v_j'$ and find maximum flow in N using max-flow min-cut algorithm. Otherwise proceed to step 7d*

   (3) *If $f(N) < \kappa$, then $\kappa \leftarrow f(N)$. Otherwise, continue.*

   (4) *$j \leftarrow j + 1$ and return to step 7a*

19.2 Matchings and Factorizations

19.2.1 An Introduction to Matching

Marriage Problem Given a collection of men and women, where each womon knows some of the men. Can every women marry a man she knows ?

Assignment Problem Given several job openings and applicants for one or more of these positions. Find an assignment so that maximum positions are filled ?

Optimal Assignment Problem Given several job openings and applicants for one or more of these positions. The benefits of employing these applicants on those positions are also given. Find an assignment of maximum benefit to the company ?

matching in G is a 1-regular [3] subgraph of G.

maximum matching in G is a matching of G with maximum cardinality.

perfect matching in G is a matching of cardinality $p/2$. ie, $p/2$ edges.

maximum weight matching in a weighted graph G is a matching with maximum weight.

Definitions 19.2.1. Let M be a matching in a graph G,

matched edge is an edge in subgraph M of G.

unmatched edge is an edge of G that doesn't belong to M.

matched vertex with respect to M is a vertex incident with an edge of M.

single vertex is a vertex that is not incident with any edge of M.

alternating path in G is a path with edges alternately matched and unmatched.

[3] A graph G is k-regular, if every vertex of G has degree k.

augmenting path in G is a non-trivial alternating path with single vertices as end vertices.

Theorem 19.2.1. *Let M_1, M_2 be two matchings in G such that there is a spanning subgraph H of G with edges that are either in M_1 or M_2, but not both. Then the components of H are either 1. isolated vertex 2. even cycle with edge alternately from M_1 and M_2 3. a non-trivial path with edges alternately from M_1 and M_2 such that each end vertex is single with respect to either M_1 or M_2, but not both.*

Synopsis. $\Delta(H) \leq 2$ by Pigeonhole principle. Any component of H is either a path or a cycle. A cycle with edge alternately from M_1 and M_2 is even. If an end vertex of a non-trivial path is matched with respect to M_1(WLOG), then it is there in $M_1 - M_2$ ie, it is not there in M_2. If there is another edge in M_2 incident with it, then it has to be in H and it will cease to be an end vertex of path component. Therefore, it is unmatched with respect to M_2.

Theorem 19.2.2. *A matching M in a graph G is maximum iff there is no augmenting path with respect to M in G.*

Synopsis. If M is maximum matching and P an M-augmenting path. Since both end-vertices are single, length of P is odd. Let M', M'' be edges of P which are in M and not in M respectively. Then $M - M' + M''$ is a matching of cardinality one greater than that of M which is a contradiction since M is maximum.

Conversely, suppose M be a matching such that there no M-augmenting paths in G. Let M' be a maximum matching in G. Then a nontrivial path component of the graph induced by $M \Delta M'$ is of even length otherwise both end-vertices are matched with respect to one of the matching M or M' which is a contradiction. Again every cycle components are even. Therefore $|M| = |M'|$, since $M \Delta M'$ doesn't have a nontrivial component of another kind.

Definitions 19.2.2. Let U_1, U_2 be two nonempty, disjoint, subsets of the vertex set of a graph G. Then U_1 **is matched to** U_2 if there exists a matching M in G such that every edge in M incident with a vertex in U_1 and a vertex in U_2. And every vertex of U_1 (or U_2) is incident with some edge in M. Suppose M^* be a matching such that $M \subset M^*$, then U_1 **is matched under** M^* **to** U_2.

Definitions 19.2.3. Let U be a nonempty set of vertices of a graph G. U is **nondeficient**,[4] if $|N(S)| \geq |S|$ for every nonempty subset S of U.

Theorem 19.2.3. *Let G be a bipartite graph with partite sets V_1, V_2. The set V_1 can be matched to a subset of V_2 iff V_1 is nondeficient.*

Corollary 19.2.3.1. *Every r-regular bipartite multigraph has a perfect matching.*

Theorem 19.2.4. *A collection $S_1, S_2, \cdots, S_n$ of finite non-empty sets has a system of distinct representatives iff for each k, $0 \leq k \leq n$, the union of any k of these sets contains at least k elements.*

[4] N(S) is the neighbourhood set of all vertices adjacent to some vertex in S

Remark (Hall's Marriage Theorem). Suppose there are n women. Then every women can marry a man she knows iff each subset of k women ($1 \leq k \leq n$) colletively knows atleast k men.

Remark. Let W be a set of n women. Then there are $2^n - 1$ nonempty subset for W. Thus, Hall's Marriage Theorem suggests that we ensure $|N(S)| \geq |S|$ for every nonempty subset S of W. This method has complexity $O(2^n)$.

19.2.2 Maximum Matching in Bipartite Graphs

Definitions 19.2.4. Let M be a matching in a graph G and P is an augmenting path with respect to M. Let M' be set of edge in P and M. And M'' be the set of edges in P and not in M. Then $M_1 = (M - M') \cup M''$ is the **matching obtained by augmenting M along path P**.

Remark. $|M_1| = |M| + 1$

Theorem 19.2.5. *Let M be a a matching of a graph G that is not maximum, and let v be a single vertex with respect to M. Let M_1 denote the mathing obtained by augmenting M along some augmenting path. If G contains an augmenting path with respect to M_1 that has v and an end-vertex, then G contains an augmenting path with respect to M that has v as an end-vertex*

Corollary 19.2.5.1. *Let M be a matching of a graph G. Suppose that $M = M_1, M_2, \cdots, M_k$ is a finite sequence of matchings of G such that M_i ($2 \leq i \leq k$) is obtained by augmenting M_{i-1} along some augmenting path. Suppose v is a single vertex with respect to M for which there exists no augmenting path starting at v. Then G does not contain an augmenting path with respect to M_i ($2 \leq i \leq k$) that has v as an end-vertex.*

Definitions 19.2.5. An **alternating tree** with respect to a matching M is a tree such that every path from it's root are alternating path with respect to M.

Algorithm 19.2.1 (Maximum Matching Algorithm for Bipartite Graphs)**.** .

1. *$i \leftarrow 1$ and $M \leftarrow M_1$*
2. *If $i < p$, then continue; otherwise stop.*
3. *If v_i is matched, then $i \leftarrow i + 1$ and return to Step 2;*
   otherwise, $v \leftarrow v_i$ and Q is initialized to contain v only.
4. (1) *For $j = 1, 2, \cdots, p$ and $j \neq i$, let $TREE(v_j) \leftarrow F$.*
   Also, $TREE(v_j) \leftarrow T$.
   (2) *If $Q = \phi$, then $i \leftarrow i + 1$ and return to Step 2;*
   otherwise, delete a vertex x from Q and continue.
   (3) (1) *Suppose that $N(x) = \{y_1, y_2, \cdots, y_k\}$. Let $j \leftarrow 1$.*
   (2) *If $j \leq k$, then $y \leftarrow y_j$; otherwise return to Step 4.2*
   (3) *If $TREE(y) = T$, then $j \leftarrow j + 1$ and return to Step 4.3.2. Otherwise, continue.*
   (4) *If y is incident with a matching edge yz, then $TREE(y) \leftarrow T$, $TREE(z) \leftarrow T$, $PARENT(y) \leftarrow x$, $PARENT(z) \leftarrow y$ and add z to Q, $j \leftarrow j + 1$ and return to Step 4.3.2. Otherwise, y is a single vertex and continue.*

(5) *Use $PARENT$ to determine the alternating $v - x$ path P' in the alternating tree. Let P be the augmenting path obtained from P' by adding the path x, y. Proceed to Step 5*

5. *Augment M along P to obtain a new matching M'. Let $M \leftarrow M'$, $i \leftarrow i + 1$, and return to Step 2.*

Definitions 19.2.6. Let G be a weighted complete bipartite graph with partite sets V_1 and V_2. A **feasible vertex labeling** is a real function $l : V(G) \to \mathbb{R}$ on vertex set of G such that $l(v) + l(u) \geq w(vu)$ where $v \in V_1$ and $u \in V_2$.

Definitions 19.2.7. Consider the function $l : V(G) \to \mathbb{R}$ such that $\forall v \in V_1,\ l(v) = \max\{w(vu) : u \in V_2\}$ and $\forall u \in V_2,\ l(u) = 0$. Then l is a feasible vertex labeling on $V(G)$. And,

E_l is the set of all edge of the weighted complete bipartite graph G such that $l(v) + l(u) = w(vu)$.

H_l is the spanning subgraph of G induced by the edge set E_l.

Theorem 19.2.6. *Let l be a feasible vertex labeling of a weighted complete bipartite graph G. If H_l contains a perfect matching M', then M' is a maximum weight matching of G.*

Algorithm 19.2.2 (Kuhn-Munkres)**.** .

1. (1) *For each $v \in V_1$, let $l(v) \leftarrow \max\{w(uv) : v \in V_2\}$.*
   (2) *For each $u \in V_2$, let $l(u) \leftarrow 0$.*
   (3) *Let H_l be the spanning subgraph of G with edge set E_l.*
   (4) *Let G_l be the underlying graph of H_l.*

2. *Apply Algorithm 19.2.1 to determine a maximum matching M in G_l.*

3. (1) *If every vertex v of V_1 is matching with respect to M, output M and stop. Otherwise, continue.*
   (2) *Let x be the first single vertex of V_1.*
   (3) *Construct an alternating tree with respect to M that is rooted at x. If an augmenting path P is discovered, then augmenting M along P and return to Step 3.1. Otherwise, let T be the alternating tree with respect to M and rooted at x that cannot be expanded further in G_l.*

4. *Compute $m_l \leftarrow \min\{l(v) + l(u) - w(vu) : v \in V_1 \cap V(T), u \in V_2 - V(T)\}$. Let*
$$l'(v) = \begin{cases} l(v) - m_l \text{ for } v \in V_1 \cap V(T) \\ l(v) + m_l \text{ for } v \in V_2 \cap V(T) \\ l(v) \text{ otherwise} \end{cases}$$

5. *Let $l \leftarrow l'$, construct G_l and return to Step 3.3.*

19.2.4 Factorizations

Definitions 19.2.8. A **factor** of a graph G is a spanning[5] subgraph of G.

Definitions 19.2.9. Let $G_1, G_2, \cdots, G_n$ be edge-disjoint factors of G such that $E(G) = \cup_{i=1}^{n} E(G_i)$. Then G is **factorable** and $G = G_1 \oplus G_2 \oplus \cdots \oplus G_n$.

Definitions 19.2.10. An r-regular factor of G is an **r-factor** of G.

Definitions 19.2.11. If G has a factorisation to r-factors, then G is **r-factorable**.

Remark. $K_{3,3}$ is 1-factorable. K_5 is 2-factorable.

Definitions 19.2.12. An **odd component of G** is a component of G with odd number of vertices. And an **even component of G** is a component of G of with even number of vertices.

Theorem 19.2.7 (Tutte)**.** *A nontrivial graph G has a 1-factor iff for every proper subset S of $V(G)$, the number of odd components of $G - S$ does not exceed $|S|$.*

Remark. There exist cubic graphs that doesn't have a 1-factor.

Theorem 19.2.8 (Petersen)**.** *Every bridgeless cubic graph contains a 1-factor.*

Remark. Every brideless cubic graphs has a 1-factor. Let G be a bridgeless cubic graph. Consider every pair of factors G_1, G_2 such that $G = G_1 \oplus G_2$ where G_1 is a 1-factor and G_2 is a 2-factor. G is not 1-factorable only if every such G_2 doesn't have a 1-factor.

Theorem 19.2.9. *Petersen graph is not 1-factorable.*

Theorem 19.2.10. *Every r-regular bipartite multigraph $(r \geq 1)$ is 1-factorable.*

Remark (Application of 1-factorisation)*.* For even number p, a 1-factorisation of K_p corresponds to the schedule of a round of the round robin tournament among p teams. If p is odd, consider K_{p+1} where v_{p+1} is an imaginary team called bye team. A game with bye team is a bye.

Definitions 19.2.13. A **hamiltonian cycle** is a spanning cycle. And, **Hamiltonian graph** is a graph containing a hamiltonian cycle.

Theorem 19.2.11. *Complete graph K_{2n+1} can be factored into n hamiltonian cycles.*

Remark. For $n = 3$, K_7 can be factored into three hamiltonian cycles.

Theorem 19.2.12. *Let $0 \leq r < p$. Then there exists an r-regular graph of order p iff pr is even.*

Definitions 19.2.14. Let $\{E_1, E_2, \cdots, E_n\}$ be partition of $E(G)$. And let H_i be subgraph of G induced by the edge set E_i. A **decomposition** of a graph G is a colletion of these subgraphs $H_1, H_2, \cdots, H_n$. And $G = H_1 \oplus H_2 \oplus \cdots \oplus H_n$.

Definitions 19.2.15. Let $G = H_1 \oplus H_2 \oplus \cdots \oplus H_n$ be a decomposition of G such that $H \cong H_i$. Then G is H-decomposible.

Remark. $K_{3,3}$ is $3K_2$-decomposible. K_5 is C_5-decomposible. K_{2n} is nK_2-decomposible. K_{2n+1} is C_{2n+1}-decomposible.
Every graph is K_2-decomposible. Every complete bipartite graph $K_{m,n}$ is $K_{1,m}$-decomposible and $K_{1,n}$-decomposible.

[5] Spanning subgraph of a graph G has every vertex of G

19.2.5 Block Designs

Definitions 19.2.16. A block design on a set V is a collection of k-element subsets of V such that each element of V appears exactly in r subsets.

variety The elements of V are called varieties.

block k-element subsets of V are called blocks.

balanced design If each variety appears in exactly r blocks and each pair of varieties appears in exactly λ blocks.

incomplete design If blocks are proper subsets of V. ie, $k < v$.

Definitions 19.2.17. A balanced incomplete block design of v varities in b blocks of cardinality k such that each variety appears in exactly r blocks and each pair of varities appears in exactly λ blocks is a (b, v, r, k, λ)-design.

Theorem 19.2.13. $bk = vr$

Theorem 19.2.14. $\lambda(v - 1) = r(k - 1)$

Corollary 19.2.14.1. $\lambda < r$

Theorem 19.2.15 (Fisher's Inequality)**.** $b \geq v$

symmetric design If $b = v$

Theorem 19.2.16. *In a symmetric (b, v, r, k, λ)-design with even v, $r - \lambda$ is a perfect square.*

steiner triple system (b, v, r, k, λ)-design with $k = 3$, $\lambda = 1$.

Remark. (b, v, r, k, λ)-designs are incomplete. However, a complete block design (ie, $v = 3$) is also included as a Steiner triple system.

Theorem 19.2.17. *Steiner triple system with v varieties exists iff $v = 6n + 1$ or $v = 6n + 3$ or $v = 3$.*

Definitions 19.2.18 (Kirkman's Schoolgirls Problem)**.** A class of 15 girls. Parade 15 girls in five rows (3 girls in a row). Is it possible to plan 7 days parade so that two girls are together in a row exactly once ?

kirkman triple system Steiner triple system with $v = 6n + 3$.

Remark. It is proved that kirkman triple system exists with $v = 6n+3$ for every $n \geq 0$.

Theorem 19.2.18. *The code consisting of the rows of the incidence matrix of a (b, v, r, k, λ)-design ($b = v$, r=k) is t-error correcting, where $t = k - \lambda - 1$.*

Subject 20

ME800403 Combinatorics

Subject 21

Probability Theory

Subject 22

Operational Research

Subject 23

Coding Theory

Subject 24

Commutative Algebra

Subject 25

Ordinary Differential Equations

Subject 26

Classical Mechanics

Bibliography

[Apostol, 1973] Apostol, T. (1973). *Mathematical Analysis, 2nd edition.* Narosa Publishing House.

[Fraleigh, 2013] Fraleigh, J. B. (2013). *A First Course in Abstract Algebra, 7th edition.* Pearson Education.

[Gray Chartrand,] Gray Chartrand, O. O. (?). *Applied and Algorithmic Graph Theory.* Tata McGraw Hill Company.

[Joshi, 1983] Joshi, K. D. (1983). *Introduction to General Topology.* Wiley Eastern Ltd.

[Kiusalaas, 2013] Kiusalaas, J. (2013). *Numerical Methods in Engineering with Python3.* Cambridge University Press.

[Kreyszig, 2014] Kreyszig, E. (2014). *Introductory Functional Analysis with applications.* John Wiley and Sons.

[Munkres, 2003] Munkres, J. R. (2003). *Topology, 2nd edition.* Pearson.

[Rudin, 1976] Rudin, W. (1976). *Principles of Mathematical Analysis, 3rd edition.* McGraw Hill Book Company, International Editions.