Measure Theory and Ergodic Theory (Note)

Based on lectures by Dr Donald Robertson

Notes taken by Yuze Li

September 2024

Contents

1 Background

In this section, we will study the background results and conventions, which mainly relate to sets, functions, and topology.

1.1 Sets and Functions

In this course, we will use the following standard notions from set theory.

Notation 1.1.1. *1.* $\varnothing$ *for the empty set*

2. $\mathbb{N}$ *for the natural numbers set*

3. $\mathbb{Q}$ *for the rational numbers set*

4. $\mathbb{R}$ *for the real numbers set*

5. $\mathbb{C}$ *for the complex number set*

For the power set, we can use the notation $\mathcal{P}(X)$ to represent the power set of a set X, and its definition is as follows.

Definition 1.1.2 (Power Set)**.** *The power set* $\mathcal{P}(X)$ *consists of all the subsets of* X*.*

For instance,

1. $\mathcal{P}(\{1,2\}) = \{\varnothing, \{1\}, \{2\}, \{1,2\}\}$
2. $\mathcal{P}(\varnothing) = \{\varnothing\}$
3. $\mathcal{P}(\{\varnothing\}) = \{\varnothing, \{\varnothing\}\}$

In $\mathbb{R}$, the intervals can be represented as

1. $(a,b) = \{t \in \mathbb{R} : a < t < b\}$
2. $[a,b) = \{t \in \mathbb{R} : a \leq t < b\}$
3. $(a,b] = \{t \in \mathbb{R} : a < t \leq b\}$
4. $[a,b] = \{t \in \mathbb{R} : a \leq t \leq b\}$

for all real numbers a, b. When $a > b$, all the sets are empty. When $a = b$, only $[a,b]$ is not empty.

Also, we have a special notion

$$[0,\infty] = [0,\infty) \cup \{\infty\}.$$

The law of addition, multiplication and order on $[0, \infty]$ follow

1. $t + \infty = \infty = \infty + t$ for all $t > 0$
2. $\infty + \infty = \infty$
3. $t * \infty = \infty$ for all $t > 0$
4. $t < \infty$ for all $t \geq 0$
5. $\infty * 0 = 0$

For function $f : X \to Y$, we can have a new function f^{-1}

$$f^{-1} : \mathcal{P}(Y) \to \mathcal{P}(X),$$

which pull back subsets of Y to get subsets of X. Then we have

1. $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$
2. $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$
3. $f^{-1}(Y \setminus B) = X \setminus f^{-1}(B)$

when $A, B \subset Y$.

1.2 Topology

Definition 1.2.1 (Topology)**.** *Any collection τ of subsets of X satisfying the following properties:*

1. *$\varnothing \subset \tau$*
2. *$X \subset \tau$*
3. *For two sets U and $V \in \tau$, their intersection is also in τ*
4. *For any collection $\alpha \mapsto U_a$ in τ, the set $\bigcup_\alpha U_a$ belongs to τ*

is called a topology on X. (X, τ) formulates a topological space.

Definition 1.2.2 (Open Ball)**.** *Given a metric space $(X, \boldsymbol{d})$, then we have an open ball $\boldsymbol{B}$ centred at x with radius r*

$$\boldsymbol{B}(x, r) = \{y \in X : \boldsymbol{d}(x, y) < r\},$$

which is a collection of metrically open subsets of X.

In the special case $X = \mathbb{R}$ and $\mathbf{d}$ is the Euclidean distance

$$\mathbf{d}(x, y) = |x - y|$$

the open balls are just the open intervals (a, b) with $a < b$. Then we have an explicit description of open subsets of $\mathbb{R}$. Every open set $U \in \mathbb{R}$ is of the form

$$U = \bigcup_{n=1}^{\infty} (a_n, b_n)$$

where the intervals (a_n, b_n) are pairwise disjoint, at most one $a_i = -\infty$ and one $b_i = \infty$.

In the topology, we can reformulate the continuity. Given two topological spaces (X, τ) and (Y, S), a map $f : X \to Y$ is continuous if $f^{-1}(U) \in \tau$ for every $U \in S$. When τ comes from a metric $\mathbf{d}_X$ on X and S comes from a metric $\mathbf{d}_Y$ on Y, the continuity of $f : X \to Y$ can be further rewritten as for every $x \in X$ and every $\epsilon > 0$ there is $\delta > 0$ such that $\mathbf{d}_X(x, y) < \delta$ implies $\mathbf{d}_X(f(x), f(y)) < \epsilon$.

2 Measure Theory

2.1 Length

For an interval (a, b), the length is $b - a$. In this section, we will study how to assign size to more subsets of $\mathbb{R}$. For instance we the Cardinality and length of some sets is listed below.

Set	Cardinality	Length
$\varnothing$	0	0
$\{\sqrt{2}\}$	1	0
$\mathbb{Q}$	$\mathcal{N}_0$	0
$\mathbb{N}$	$\mathcal{N}_0$	0
Middle Thirds Cantor Sets	c	0
$\mathbb{R}$	c	∞

It can be seen from the chat above, the length is a value in $[0, \infty) \cup \{\infty\}$, and we will illustrate how to find the length in this part. Based on the length of interval (a, b), if a set $B \subset \mathbb{R}$ is contained within a union of intervals

$$B \subset (a_1, b_1) \cup (a_2, b_2) \cup \cdots \cup (a_N, b_N)$$

then it is natural that we have

$$Length(B) \leq \sum_{n=1}^{N} (b_n - a_n) \tag{1}$$

the length of B is not larger than the sum of lengths of the intervals. Therefore, the sum

$$\sum_{n=1}^{N} (b_n - a_n)$$

could be considered very close to the length of B. To illustrate the formal definition of length, we can start with an example (Middle Thirds Cantor Sets). The middle thirds cantor sets are

$$\begin{aligned} C_0 &= [0, 1] \\ C_1 &= [0, \frac{1}{3}] \cup [\frac{2}{3}, 1] \\ C_2 &= [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1] \\ &\vdots \end{aligned}$$

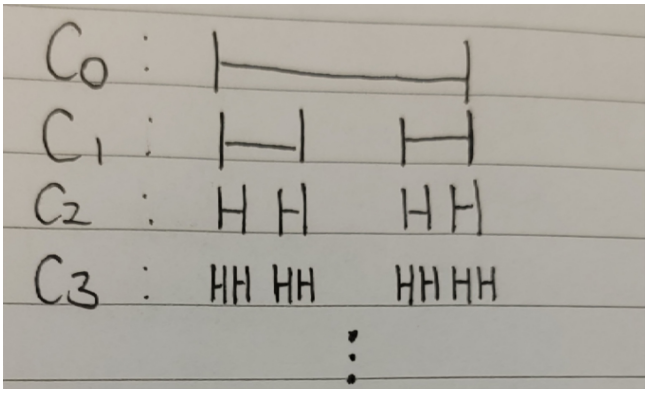

Figure 1: Middle Thirds Counter Sets

and the middle thirds cantor sets can also be drawn as Figure 1. By repeating the procedure, we can have a set

$$\mathcal{C} = \bigcap_{n=1}^{\infty} C_n.$$

Therefore, if $x \in \mathcal{C}$, it belongs to every C_n, and it will never be removed. All end points of each intervals in any C_n will survive at the end, and they belongs to $\mathcal{C}$. In the view of length, for $n \in \mathbb{N}$, C_n has 2^n intervals and each have a length $\frac{1}{3^n}$. Therefore, the length of $C_n = \left(\frac{2}{3}\right)^n$. Since $\mathcal{C} \subset C_n$, the length at most is $\left(\frac{2}{3}\right)^n$. The length of $\mathcal{C}$ should be 0.

Therefore, we can provide a formal definition for length.

Definition 2.1.1 (Length)**.** *Given $E \in \mathbb{R}$, we can define its length $\Lambda(E)$ to be*

$$\inf\{\sum_{n=1}^{\infty}(b_n - a_n) : E \subset \bigcup_{n=1}^{\infty}(a_n, b_n)\},$$

and this takes value in $[0, \infty]$. It is the infimum that picks out the most efficient way of covering E by intervals.

For instance, we can use the Definition 2.1.1 to prove $\Lambda(\sqrt{2}) = 0$.

Proof. The set $\left(\sqrt{2} - \frac{1}{m}, \sqrt{2} + \frac{1}{m}\right)$ contains $\sqrt{2}$. Then we let

$$\begin{aligned}(a_1, b_1) &= \left(\sqrt{2} - \frac{1}{m}, \sqrt{2} + \frac{1}{m}\right)\\ (a_2, b_2) &= (5, 5)\\ (a_3, b_3) &= (5, 5)\\ &\vdots\end{aligned}$$

$\{\sqrt{2}\} \subset \bigcup_{n=1}^{\infty}(a_n, b_n)$, and the length of $\bigcup_{n=1}^{\infty}(a_n, b_n) = \sum_{n=1}^{\infty}(b_n - a_n) = \frac{2}{m}$. Therefore, $\Lambda(\{\sqrt{2}\}) = \inf\{\frac{2}{m} : \{\sqrt{2}\} \subset \bigcup_{n=1}^{\infty}(a_n, b_n)\} = 0$ □

The length (Λ) is a function defined on $\mathcal{P}(\mathbb{R})$, and it has the following properties:

Theorem 2.1.2 (Properties of Length (Λ))**.**

1. $\Lambda(\varnothing) = 0$
2. $\Lambda(A) \leq \Lambda(B)$ *whenever* $A \subset B$
3. $\Lambda(A_1 \cup A_2 \cup \cdots) \leq \Lambda(A_1) + \Lambda(A_2) + \Lambda(A_3) + \cdots$ *for all* A_1, A_2, $A_3, \cdots \subset \mathbb{R}$.
4. $\Lambda(A - t) = \Lambda(A)$ *for all* $A \subset \mathbb{R}$

Based on the first three categories in Theorem 2.1.2, we can summarize the definition of outer measure:

Definition 2.1.3 (Outer Measure)**.** *Given a set* X*, a map* $M : \mathcal{P}(x) \to [0, \infty]$ *is an outer measure if*

1. $M(\varnothing) = 0$
2. $M(A) \leq M(B)$ *whenever* $A \subset B$
3. $M\Big(\bigcup_{n=1}^{\infty} A_n\Big) \leq \sum_{n=1}^{\infty} M(A_n)$ *for every sequence* $n \to A_n$ *in* $\mathcal{P}(X)$

Therefore, we can conclude that length (Λ) is an outer measure.

2.2 Size

In 2.1, we have studied how to assign a length to the subsets of $\mathbb{R}$. In this part, we will assign an outer measure to the set $X = \{0,1\}^{\mathbb{N}}$, which is the set of all functions from $\mathbb{N} \to \{0,1\}$ For example, we have the following sequences

1. $010101010101\cdots$
2. $1111000000000000\cdots$
3. $01110111100000000110001100\cdots$

and these sequences suggest patterns defining sequence in $\{0,1\}^{\mathbb{N}}$. X is important in dynamics and probability, which can be a simple space for coin testing, and we will study how to assign an outer measure on X. Also, we have a notion called Cylinder Set to denote the subsets of $\{0,1\}^{\mathbb{N}}$.

Definition 2.2.1 (Cylinder Set)**.** *Given values $\epsilon_1, \epsilon_2, \cdots, \epsilon_j$ we write*

$$[\epsilon_1, \epsilon_2, \cdots, \epsilon_j] = \{x \in \{0,1\}^{\mathbb{N}} : x(j) = \epsilon_j \text{ for all } 1 \leq j \leq J\}$$

for the cylinder set defined by the initial coordinates $\epsilon_1, \epsilon_2, \cdots, \epsilon_j$.

For instance, we have

1. $[0] = \{x \in X : x(1) = 0\}$ $(J = 1, \epsilon_1 = 0)$
2. $[101] = \{x \in X : x(1) = 1, x(2) = 0, x(3) = 1\}$ $(J = 3, \epsilon_1 = 1, \epsilon_2 = 0, \epsilon_3 = 1)$

The cylinder can be seen as the outcomes of coin testing, and we can use the probability (likelihood) to assign a size. For example $[\epsilon_1, \cdots, \epsilon_J]$ has the likelihood $P^{\sum_{j=1}^{J} \epsilon_j}(1-P)^{J-\sum_{j=1}^{J} \epsilon_j}$. In this case, we can use a new outer measure to denote the size.

Definition 2.2.2 (Size)**.** *Fix $0 < p < 1$. For every $B \subset \{0,1\}^{\mathbb{N}}$ we can define*

$$\Xi_p(B) = \inf\{\sum_{n=1}^{\infty} Prob_p(C_n) : B \subset \bigcup_{n=1}^{\infty} C_n \text{ with each } C_n \text{ a cylinder}\}$$

For each value $0 \leq p \leq 1$ we have a map

$$\Xi_p : \mathcal{P}(\{0,1\}^{\mathbb{N}}) \to [0, \infty]$$

Also Ξ_p has the similar properties with Λ:

Theorem 2.2.3 (Ξ_p Properties)**.** *1.* $\Xi(\varnothing) = 0$

2. $\Xi_p(A) \leq \Xi_p(B)$ *whenever* $A \subset B \subset \{0,1\}^{\mathbb{N}}$

3. $\Xi_p(A_1 \cup A_2 \cup \cdots) \leq \Xi_p(A_1) + \Xi_p(A_2) + \Xi_p(A_3) + \cdots$ *for all* A_1, A_2, $A_3, \cdots \subset \{0,1\}^{\mathbb{N}}$.

In this section, we will study σ-algebra and the measures.

2.3 σ-Algebra

Both Λ and Ξ are badly behaved in one important way because they are not countably additive:

$$A \cap B = \varnothing \Rightarrow \Lambda(A) + \Lambda(B) = \Lambda(A \cup B),$$

The definition of additivity and countable additivity can be formalized as:

Definition 2.3.1. *An outer measure M on X is additive if*

$$M(A_1 \cup \cdots \cup A_r) = M(A_1) + \cdots + M(A_r)$$

when the sets $A_1, \cdots, A_r$ are pairwise disjoint.

M is countably additive if

$$M(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} A_n$$

when the sets $A_1, \cdots, A_r$ are pairwise disjoint.

For the outer measures, additivity and countable additivity are equivalent. However, in the note of week 1, it can be known that Λ is not countably additive. In fact, there is no notion of size on $\mathbb{R}$ with all the following properties:

1. Assigns a size to every subset of $\mathbb{R}$
2. Countable additivity
3. Assigns to each interval its Euclidean length
4. Translation invariance

To develop the theory of length, we will insist on the translation invariance and countable additivity, and abandon the requirement that Λ is defined on all of $\mathcal{P}(\mathbb{R})$. We will identify a rich collection $\mathscr{B}$ of subsets of $\mathbb{R}$ so that

$$\Lambda : \mathscr{B} \to [0, \infty].$$

In this course, we will use the σ-algebra to keep track of the subsets of X whose size we are permitted to calculate.

Definition 2.3.2 (σ-algebra)**.** *Fix a set X, a set $\mathscr{B} \subset \mathcal{P}(X)$ is a σ-algebra if it has all the following properties*

1. *$X \in \mathscr{B}$*
2. *If $B \in \mathscr{B}$ then $X \setminus B \in \mathscr{B}$*
3. *For any sequence $B_1, B_2, B_3, \cdots$ of sets in $\mathscr{B}$ the union*

$$B_1 \cup B_2 \cup B_3 \cup \cdots$$

belongs to $\mathscr{B}$.

The σ-algebra will serve as the domain of measures, and we use a pair $(X, \mathscr{B})$ to denote the measurable space where X is a set and $\mathscr{B}$ is a σ-algebra of subsets of X. Here are some examples of σ-algebra:

1. (The trivial σ-algebra). For any set X the collection

$$\{\varnothing, X\}$$

is a σ-algebra and called the trivial σ-algebra on X.

2. (The full σ-algebra). For any set X the collection

$$\mathcal{P}(x)$$

is a σ-algebra and called the full σ-algebra on X.

3. (The σ-algebra generated by a collection). For any set X and any collection $\mathcal{F} \subset \mathcal{P}(X)$, there is a σ-algebra containing $\mathcal{F}$ that we can believe it is the σ-algebra generated by $\mathcal{F}$:

$$\sigma(\mathcal{F}) = \{A \subset \mathbb{R} : \text{A in every } \sigma\text{-algebra that contains } \mathcal{F}\}$$

It can also be seen as the intersection of all σ-algebra on X contains $\mathcal{F}$, which is cited from [1].

We usually work with $\sigma(\mathcal{F})$, which is also the smallest σ-algebra containing $\mathcal{F}$. If we say a σ-algebra contains $\mathcal{F}$, it is the $\sigma(\mathcal{F})$. For instance if $\mathcal{F} = \{\{1,2\},\{3,4\},\{5,6\}\}$ then $\sigma(\mathcal{F})$ is in the form:

$$\sigma(\mathcal{F}) = \{\varnothing, \{1,2\},\{3,4\},\{5,6\},\{1,2,3,4\},\{1,2,5,6\},\{3,4,5,6\},\{1,2,3,4,5,6\}\}$$

Now we will study the important kind of σ-algebra *Borel* σ-algebra.

Definition 2.3.3 (Borel σ-Algebra)**.** *When X is a topological space we can equip X with a distinguished σ-algebra called Borel σ-algebra. It is denoted $Borel(X)$ and defined as the σ-algebra generated by the collection of all open subsets of X. The subsets of X that belong to $Borel(X)$ are called Borel subsets of X.*

Also according to [1], we can define $\mathcal{O}$ be the class of all open subsets of X, the smallest σ-algebra $\sigma(\mathcal{O})$ is the Borel σ-algebra on X ($Borel(X)$). The elements in $Borel(X)$ can also be called as the Borel subsets of X.

The Borel σ-algebra on $\mathbb{R}$ is the σ-algebra generated by all open intervals $\{(a,b) : a < b\}$ and or by class of all interval $(-\infty, a)$. Given a topological space (x, τ) the Borel σ-algebra X is $G(\tau)$. On $\{0,1\}^{\mathbb{N}}$, the Borel σ-algebra is the σ-algebra generated by cylinders.

The Borel σ-algebra is generated by the open sets, so it contains closed sets, all union of countably many closed sets, all intersections of countably many open sets, and so on. For instance

$$\{\pi\} = \bigcap_{n=1}^{\infty}(\pi - \frac{1}{n}, \pi + \frac{1}{n}) = \Big(\mathbb{R}/(-\infty, \pi)\Big) \cap \Big(\mathbb{R}/(\pi, \infty)\Big)$$

Therefore $\{\pi\}$ is a Borel subset.

Proposition 2.3.4. *Based on the definition of Borel σ-Algebra, we have the σ-algebra on $\mathbb{R}$ generated by the open intervals is the Borel σ-algebra on $\mathbb{R}$.*

Proof. Define ζ as the σ-algebra on $\mathbb{R}$ generated by the open intervals, we have $\zeta \subset Borel(\mathbb{R})$, since every open interval in $\mathbb{R}$ is the subset of $Borel(\mathbb{R})$ and ζ is the smallest σ-algebra generated by every open interval. To prove $Borel(\mathbb{R}) \subset \zeta$, we need to show every *Borel* sets are in ζ. This statement holds naturally because every open set belongs to ζ and the *Borel* set is in the form of the open set (countable union of open sets) in $\mathbb{R}$. □

Then σ-algebras have the following properties:

Theorem 2.3.5. *Given a σ-algebra $\mathscr{B}$, we have*

1. *If $B_1, B_2, \cdots \in \mathscr{B}$ then*

$$\bigcap_{n=1}^{\infty} B_n$$

   belongs to $\mathscr{B}$

2. *If $B_1, \cdots B_n \in \mathscr{B}$ then $B_1 \cup \cdots \cup B_n \in \mathscr{B}$*

3. *If $B_1, \cdots B_n \in \mathscr{B}$ then $B_1 \cap \cdots \cap B_n \in \mathscr{B}$*

4. If $B, C \in \mathscr{B}$ then $B \setminus C \in \mathscr{B}$.

Proof. The first properties can be obtained by

$$\bigcap_{n=1}^{\infty} B_n = X \setminus \left(\bigcup_{n=1}^{\infty} X \setminus B_n \right)$$

According to the definition of σ-algebra, $X \setminus B_n \in \mathscr{B}$, then $\bigcup_{n=1}^{\infty} X \setminus B_n \in \mathscr{B}$, and finally $X \setminus (\bigcup_{n=1}^{\infty} X \setminus B_n) \in \mathscr{B}$. Then, all the other three properties can be verified by the first property and the definition of σ-algebra.

□

2.4 Measures

Our most important example of measurable space are ($\mathbb{R}$ and $Borel(\mathbb{R})$) and ($\{0,1\}^{\mathbb{N}}, Borel(\{0,1\}^{\mathbb{N}})$). For us, the job of measurable space is to act as the thing on which you want to measure. In other words, a measure is a mapping from a σ-algebra to $[0,\infty)$, which assigns length to the σ-algebra. According to [1], it is not possible to define a measure on every subset of $\mathbb{R}$ such as Vitali set in the first week, but we can define the measure on the subsets of a σ-algebra and the set and its corresponding σ-algebra formulates a measurable space.

Definition 2.4.1 (Measure)**.** *Fix a measurable space $(X, \mathscr{B})$, a measure on $(X, \mathscr{B})$ is any map:*

$$\mu : \mathscr{B} \to [0, \infty]$$

with the following properties:

1. $\mu(\varnothing) = 0$
2. $A_1, A_2, A_3, \cdots$ *are pointwise disjoint is in* $\mathscr{B}$

$$\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$$

Note: Λ is not a measure on $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$. We aim to define a measure on $(\mathbb{R}, Borel(\mathbb{R}))$ that assigns intervals their length, which is called the Lebesgue measure and is important for us to study the Lebesgue integration.

Before studying the Lebesgue measure on the real line, we will study some examples of measures:

1. (Counting Measure). Fix a set X. The counting measure on X is defined on $(X, \mathcal{P}(X))$

$$\mu(E) = \sup\{|A| : A \in E \text{ with } 0 \leq |A| < \infty\}$$

   For instance,

$$\mu(\{\sqrt{2}, 4, -\pi\}) = 3$$

   and $\mu(\mathbb{N}) = \infty$.

2. (Point Measure). For any set X and any $a \in X$

$$\delta_a(A) = \begin{cases} 1 \ a \in A \\ 0 \ a \notin A \end{cases}$$

   For instance, if we let $A = \sqrt{2}$, we have $\delta_a(\{\sqrt{2}, 4, -\pi\}) = 1$.

Since we have studied the definition of measures, we will introduce their properties:

Theorem 2.4.2 (Properties of Measures)**.** *Fix a measure space $(X, \mathscr{B}, \mu)$, the measure μ has the following properties*

1. *Monotonicity: If $B, C \in \mathscr{B}$ with $B \subset C$ then*

$$\mu(B) \leq \mu(C)$$

2. *Subadditivity: If $B_1, B_2, \cdots \in \mathscr{B}$ then*

$$\mu\left(\bigcup_{n=1}^{\infty} B_n\right) \leq \sum_{n=1}^{\infty} \mu(B_n)$$

3. *Continuity I: If $B_1 \subset B_2 \subset \cdots$ in $\mathscr{B}$ then*

$$\mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mu(B_n)$$

4. *Continuity II: If $B_1 \supset B_2 \supset \cdots$ in $\mathscr{B}$ and $\mu(B_1) < \infty$ then*

$$\mu\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mu(B_n)$$

Proof. To prove these properties, we need to define a measurable space $(X, \mathscr{B}, \mu)$. Then we will prove these theorems one by one.

1. Given two sets B and $C \in \mathscr{B}$ and $B \subset C$, we can set $D = C \setminus B$, then we have

$$\mu(C) = \mu(B) + \mu(D) \geq \mu(B).$$

2. We can formulate a sequence $C_1 = B_1$ and

$$C_{n+1} = B_{n+1} \setminus (B_1 \cup \cdots B_n)$$

   for all $n \in \mathbb{N}$. Therefore each C_n are pairwise disjoint, and $C_n \subset B_n$, then we have

$$\mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \mu\left(\bigcup_{n=1}^{\infty} C_n\right) = \sum_{n=1}^{\infty} \mu(C_n) \leq \sum_{n=1}^{\infty} \mu(B_n).$$

3. We can define a sequence such that $B_1 = C_1$, and $C_n = B_n \setminus B_{n-1}$ with $n \geq 2$, so we have $B_n = \bigcup_{n=1}^{n} C_n$. As a result, we have:

$$\mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \mu\left(\bigcup_{n=1}^{\infty} C_n\right) = \sum_{n=1}^{\infty} \mu(C_n) = \lim_{N\to\infty} \sum_{n=1}^{N} \mu(C_n) = \lim_{N\to\infty} \mu(B_n)$$

4. Defining $C_n = B_1 \setminus B_n$ for all $n \in \mathbb{N}$, then we can see that C_n is an increasing sequence, such that

$$\mu\left(\bigcup_{n=1}^{\infty} C_n\right) = \lim_{n\to\infty} \mu(C_n),$$

and

$$\mu(C_n) = \mu(B_1) - \mu(B_n).$$

Then we have

$$\mu\left(\bigcup_{n=1}^{\infty} C_n\right) = \mu\left(B_1 \setminus \bigcap_{n=1}^{\infty} B_n\right) = \mu(B_1) - \mu\left(\bigcap_{n=1}^{\infty} B_n\right).$$

Therefore we have

$$\mu\left(\bigcap_{n=1}^{\infty} B_n\right) = \mu(B_1) - \mu\left(\bigcup_{n=1}^{\infty} C_n\right) = \mu(B_1) - \lim_{n\to\infty} \mu\left(C_n\right) = \lim_{n\to\infty} \mu\left(B_n\right)$$

□

2.5 Constructing Measures

It is reasonable to believe that the interval (a, b) has the length $b - a$. In this section, we will construct a measure on $\mathbb{R}$ that satisfies this suggestion. It is known that measures defined on σ-algebra, and σ-algebras contain intervals (a, b) must contained in the $\text{Bor}(\mathbb{R})$, we will construct a measure λ on $\text{Bor}(\mathbb{R})$ such that

$$\lambda((a, b)) = b - a$$

for all $a < b$. In other words, we know Λ is an outer measure on $\mathcal{P}(\mathbb{R})$, our goal is to prove $\lambda = \Lambda \mid \text{Bor}(\mathbb{R})$ to the Borel subsets is a measure.

The central step in constructing s measure μ from an outer measure M on $\mathcal{P}(x)$ is to identify a σ-algebra on which λ defines a measure. In the special case, $X = \mathbb{R}$ and $M = \Lambda$ - a criterion for membership in such a σ-algebra was one of Lebesgue's great insights. If we have such a σ-algebra, it must contain the open intervals, and for any set L in the σ-algebra, we also have $I \cap L$ and $I \setminus L$ in the σ-algebra. Moreover, the measure of I must equal to the sum of measures of $I \cap L$ and $I \setminus L$. Lebesgue defined a set to be measurable if it splits every open interval correctly. In this course, we will study a stronger and more abstract definition of measurable sets, compared with Lebesgue's definition.

Definition 2.5.1 (Carathédory's Theorem)**.** *Fix an outer measure M on a set X. A set $L \subset X$ is Carathédory measurable for M if*

$$M(S) = M(L \cap S) + M(L^c \cap S)$$

for every $S \subset X$. $Car(M)$ is the collection of all subsets of X that are Carathédory measurable for M. The special case is $X = \mathbb{R}$ and $M = \lambda$.

Then we can specify this definition to the Lebesgue measure:

Definition 2.5.2. *A subset L of $\mathbb{R}$ is Lebesgue measurable if*

$$\Lambda(S) = \Lambda(L \cap S) + \Lambda(L^c \cap S) \tag{2}$$

for every $S \subset \mathbb{R}$. $Leb(\mathbb{R})$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}$.

According to the Carathédory's Theorem, we want to prove a set L is Carathédory measurable, we only need to verify:

$$M(S) \geq M(L \cap S) + M(L^c \cap S),$$

because

$$M(S) \leq M(L \cap S) + M(L^c \cap S)$$

holds naturally, due to the properties of outer measure.

With the Carathédory's Theorem, we can define a measure M on the σ-algebra $Car(M)$, which can be formally formulated as the theorem below:

Theorem 2.5.3. *Fix an outer measure M on a set X. The collection $Car(M)$ is a σ-algebra and the restriction of M to $Car(M)$ is a measure.*

Proof. Since M is an outer measure, we can directly have $M(\varnothing) = 0$. Also, we have $\varnothing$ is Carathédory measurable, since

$$M(S) = M(\varnothing \cap S) + M(\varnothing^c \cap S) = M(S)$$

Also, by (2), it can be noticed that if L is Carathédory measurable, then L^c is also Carathédory measurable. If $L, K \subset X$, we will prove $L \cup K$ is Carathédory measurable. It is known that

$$M(S) \leq M((L \cup K) \cap S) + M((L \cup K)^c \cap S).$$

Then

$$\begin{aligned} M(S) &= M(L \cap S) + M(L^c \cap S) \\ &= M(K \cap (L \cap S)) + M(K^c \cap (L \cap S)) + M(K \cap (L^c \cap S)) + M(K^c \cap (L^c \cap S)) \\ &\geq M(S \cap (L \cup K)) + M(S \cap (L \cup K)^c). \end{aligned}$$

If L and K are disjoint, we have

$$M(L \cup K) = M(L \cap (L \cup K)) + M(L^c \cap (L \cup K)) = M(L) + M(K),$$

which means M is additive on $Car(X)$. Further, we set $L_1, L_2, \cdots$ are point-wise disjoint, then we need to prove their union K is also in $Car(M)$ and M is countable additive. For $S \subset X$, we have

$$\begin{aligned} M(S \cap (L_1 \cup \cdots \cup L_N)) &= M(S \cap (L_1 \cup \cdots \cup L_N) \cap L_N) + M(S \cap (L_1 \cup \cdots \cup L_N) \cap L_N^c) \\ &= M(S \cap L_N) + (S \cap (L_1 \cup \cdots \cup L_{N-1})) \end{aligned}$$

so by the idea of induction, we have

$$M(S \cap (L_1 \cup \cdots \cup L_N)) = \sum_{n=1}^{N} M(S \cap L_n).$$

For any $n \in \mathbb{N}$, we have

$$\begin{aligned} M(S) &\geq M(S \cap K^c) + \sum_{n=1}^{N} M(S \cap L_n) \\ &\geq M(S \cap K^c) + M\left(S \cap \bigcup_{n=1}^{\infty} L_n\right) \geq M(S) \end{aligned}$$

Therefore, we have

$$M\left(\bigcup_{n=1}^{\infty} L_n\right) = \sum_{n=1}^{\infty} M(L_n).$$

As a result, the union is Carathédory measurable, and the measure is countable additive. □

Now we know that L is a measure on the σ-algebra $Leb(\mathbb{R})$, in this case, we will prove that the Borel σ-algebra is contained in $Leb(\mathbb{R})$.

Theorem 2.5.4. *The Lebesgue σ-algebra $Leb(\mathbb{R})$ contains the Borel σ-algebra $Bor(\mathbb{R})$.*

Proof. Since $Borel(\mathbb{R})$ is generated by the open interval (a, b) with $a < b$, therefore to prove this theorem, we can show every open interval (a, b) belongs to $Leb(\mathbb{R})$, which means we need to show

$$\Lambda(S) \geq \Lambda(I \cap S) + \Lambda(I^c \cap S) \tag{3}$$

where $I = (a, b)$ and $S \subset \mathbb{R}$. Similar to the section Length 2.1, we can use an countable union of intervals $\bigcup_{n=1}^{\infty} J_n$ to cover S, such that

$$\sum_{n=1}^{\infty} Length(J_n) \leq \Lambda(S) + \epsilon$$

due to the properties of inf. Then to prove the (3), we can rewrite $J_n = (I \cap J_n) \cup (I^c \cap J_n) = ((a, b) \cap J_n) \cup ((-\infty, a] \cap J_n) \cup ([b, \infty) \cap J_n)$ Then the length could be relaxed by

$$Length(J_n) + \frac{2\epsilon}{2^n} \geq Length(I \cap J_n) + Length((-\infty, a + \frac{\epsilon}{2^n}] \cap J_n) + Length([b - \frac{\epsilon}{2^n}, \infty) \cap J_n).$$

Therefore, we have

$$\begin{aligned} \Lambda(S) + 3\epsilon &\geq \sum_{n=1}^{\infty} Length(J_n) + \frac{2\epsilon}{2^n} \\ &\geq \sum_{n=1}^{\infty} Length(I \cap J_n) + \sum_{n=1}^{\infty} Length((-\infty, a + \frac{\epsilon}{2^n}] \cap J_n) + \sum_{n=1}^{\infty} Length([b - \frac{\epsilon}{2^n}, \infty) \cap J_n) \\ &\geq \Lambda(I \cap S) + \Lambda(I^c \cap S) \end{aligned}$$

Since ϵ is arbitrary, therefore we have $\Lambda(S) \geq \Lambda(I \cap S) + \Lambda(I^c \cap S)$. □

2.6 Measurable Function

Since we have studied the structure of σ-algebras, in this section we will study the functions on the sets following this structure. This kind of function has a similar role as the continuous function in topology. They are important in measuring theory since they allow us to perform the integration, and they can be formally defined as

Definition 2.6.1 (Measurable Function)**.** *Given X and Y are two sets, $\mathscr{B}$ is a σ-algebra on the subsets of X and $\mathscr{C}$ is a σ-algebras on the subsets of Y. Then a function $f : X \to Y$ is $(\mathscr{B}, \mathscr{C})$ measurable if $f^{-1}(C) \in \mathscr{B}$ for every $C \in \mathscr{C}$.*

Note: To check a function is measurable, we only need to verify that $f^{-1}(C) \in \mathscr{B}$ for all sets C in a generating set for σ-algebra, which is stated in the theorem:

Theorem 2.6.2. *Let $\mathscr{B}$ and $\mathscr{C}$ be σ-algebras on X and Y respectively. If $\mathscr{C}$ is generated by a collection $\mathcal{F} \subset \mathcal{P}(Y)$, and $f^{-1}(F) \in \mathscr{B}$ for every $F \in \mathcal{F}$ then f is $(\mathscr{B}, \mathscr{C})$ measurable.*

Proof. To prove this, we need to define a collection

$$\mathscr{G} = \{E \subset Y : f^{-1}(E) \in \mathscr{B}\}.$$

$\mathscr{C} = \sigma(\mathcal{F})$, $\mathscr{G}$ contains all sets from Y that their pre-images belongs to $\mathscr{B}$. If we show $\mathscr{G}$ contains $\mathscr{C}$, then we have proved that $f^{-1}(C) \in \mathscr{B}$ for every $C \in \mathscr{C}$. Therefore, we only need to show $\mathscr{G}$ is a σ-algebra, since it contains $\mathscr{F}$, and $\mathscr{C}$ is the smallest σ-algebra generated by $\mathcal{F}$. It could be directly obtained that $Y \in \mathscr{G}$. Then if $E \in \mathscr{G}$, then $Y \setminus E$ also belongs to $\mathscr{G}$, since $f^{-1}(Y \setminus E) = f^{-1}(Y) \setminus f^{-1}(E) = X \setminus f^{-1}(E) \in \mathscr{B}$ due to $\mathscr{B}$ is a σ-algebra. If $E_1, E_2, \cdots \in$, then their countable union also belongs to , because

$$f^{-1}\left(\bigcup_{n=1}^{\infty} E_n\right) = \bigcup_{n=1}^{\infty} f^{-1}(E_n) \in \mathscr{B}$$

Therefore $\mathscr{G}$ is a σ-algebra. □

To study the integration, we are interested in the real-valued functions, which is from the given set X to the real line $\mathbb{R}$. Since the real line has a unique σ-algebra $Borel(\mathbb{R})$, given a measurable space $(X, \mathscr{B})$, we would like to determine whether a function $f : X \to \mathbb{R}$ is $(\mathscr{B}, Bor(\mathbb{R})$ measurable.

Proposition 2.6.3. *A function is said to be $(\mathscr{B}, Bor(\mathbb{R}))$ measurable, if $f^{-1}(U) \in \mathscr{B}$ for every open set $U \subset \mathbb{R}$, given a measurable space $(X, \mathscr{B})$ and a function $f : X \to \mathbb{R}$.*

Proof. Since $Borel(\mathbb{R})$ is generated by all open sets in $\mathbb{R}$, this proposition holds due to the previous theorem. □

Therefore, we can provide more explicit criteria for $(\mathscr{B}, Bor(\mathbb{R}))$ mensurability.

Theorem 2.6.4. *For a measurable space $(X, \mathscr{B})$, a function $f : X \to \mathbb{R}$ is $(\mathscr{B}, Bor(\mathbb{R}))$ measurable when any of the followings is true:*

1. *$f^{-1}((a, b)) \in \mathscr{B}$ for all $a < b$;*
2. *$f^{-1}((a, \infty)) \in \mathscr{B}$ for all $a < \infty$.*

We will omit this proof and provide some examples of how to apply this theorem.

1. $f(x) = \mathbf{1}_{\mathbb{R}\setminus\mathbb{Q}}(x)$ is $(Bor(\mathbb{R}), Bor(\mathbb{R}))$ measurable. It can be easily verified. When $a < 0$, $f^{-1}((a, \infty)) = \mathbb{R}$. When $0 \leq a < 1$, $f^{-1}((a, \infty)) = \mathbb{R} \setminus \mathbb{Q}$. When $a \geq 1$, $f^{-1}((a, \infty)) = \varnothing$ Both of them belong to $Bor(\mathbb{R})$, so f is $(Bor(\mathbb{R}), Bor(\mathbb{R}))$ measurable.
2. $f(x) = \mathbf{1}_{(0,1)}(x)\frac{1}{x}$ is $(Bor(\mathbb{R}), Bor(\mathbb{R}))$ measurable. When $a < 0$, $f^{-1}((a, \infty)) = \mathbb{R}$. When $a > 0$, $f^{-1}((a, \infty)) = (0, min(\frac{1}{a}, 1))$. Both of them belong to $Bor(\mathbb{R})$.

We have studied the basic definition of measurable function and the methods to check the mensurability of a function. We will study that the sum of measurable functions are measurable.

Theorem 2.6.5. *Given a measurable space $(X, \mathscr{B})$, for all the $(\mathscr{B}, Bor(\mathbb{R}))$ measurable functions f, g from $X \to \mathbb{R}$, the sum $f + g$ is also $(\mathscr{B}, Bor(\mathbb{R}))$ measurable.*

Proof. To prove this theorem, we will utilize the theorem 2.6.4 to show

$$(f + g)^{-1}((a, \infty)) = \{x \in X : f(x) + g(x) > a\}$$

belongs to $\mathscr{B}$. There exists a $q \in \mathbb{Q}$ such that

$$f(x) > q > a - g(x).$$

Therefore,

$$\{x \in X : f(x) + g(x) > a\} = \bigcup_{q\in\mathbb{Q}} f^{-1}((q, \infty)) \cap g^{-1}((a - q, \infty))$$

It is a countable union of intersections, therefore it belongs to $\mathscr{B}$. □

From the worksheet of week 4, we have many a group of statements about the measurable function.

Theorem 2.6.6 (Non-decreasing & Measurable)**.** *If $f : [a, b] \to \mathbb{R}$ is non-decreasing, then f is $(Bor([a, b]), Bor(\mathbb{R}))$ measurable.*

Theorem 2.6.7 (Continuous & Measurable)**.** *Every continuous function $f : \mathbb{R} \to \mathbb{R}$ is $(Bor(\mathbb{R}), Bor(\mathbb{R}))$ measurable.*

Theorem 2.6.8 (Pointwise & Measurable)**.** *Let $f_1, f_2, \cdots$ be a sequence of $(Bor(\mathbb{R}), Bor(\mathbb{R}))$ measurable functions. Suppose $f_n \to g$ pointwise, then g is $(Bor(\mathbb{R}), Bor(\mathbb{R}))$ measurable.*

We have studied the concepts of measurable functions, in next section, we will study the Lebesgue integration.

2.7 Integration

Before formally studying the integration, we will briefly study the concept of indicator functions and simple functions.

The simple functions have already been used in the previous section, in this section we will provide its formal definition and some lemma since the simple functions are formulated by the indicator functions.

Definition 2.7.1 (Indicator Function)**.** *Fix a set X and $A \subset X$. The indicator function of A is the function*

$$\mathbf{1}_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

from $X \to \mathbb{R}$.

Lemma 2.7.2. *Fix a measurable function $(X, \mathscr{B})$, for every $B \in \mathscr{B}$, the function $\mathbf{1}_B : X \to \mathbb{R}$ is $(\mathscr{B}, \mathcal{P}(\mathbb{R}))$ measurable.*

Proof. Since the pre-image of this functions are $\{\varnothing, B, X \setminus B, X\}$, since $B \in \mathscr{B}$, all the elements belongs to $\mathscr{B}$. Therefore, the indicator function is is $(\mathscr{B}, \mathcal{P}(\mathbb{R}))$ measurable. □

Based on the indicator functions, we can formulate a standard definition of a simple function.

Definition 2.7.3 (Simple Functions)**.** *Fix a measurable space $(X, \mathscr{B})$. A function $f : X \to \mathbb{R}$ is simple if*

$$f = a_1 \mathbf{1}_{B(1)} + \cdots + a_k \mathbf{1}_{B(k)}$$

where $B(1), \cdots, B(k)$ in $\mathscr{B}$ and $a_1, \cdots, a_k$ are real numbers.

For the simple function, we have a lemma:

Lemma 2.7.4. *Fix a measurable function $(X, \mathscr{B})$, a function $f : X \to \mathbb{R}$ is simple iff its range is finite and it is $(\mathscr{B}, \mathcal{P}(\mathbb{R}))$ measurable.*

The first step for us is to integrate the non-negative simple function.

Definition 2.7.5. *Fix a measure space $(X, \mathscr{B}, \mu)$. We can define an integral of a simple, measurable function $f : X \to [0, \infty)$ wrt μ to be*

$$\int f d\mu = \sum_{t \in \mathbb{R}} t \times \mu(f^{-1}(\{t\})) = \sum_{i=1}^{n} a_i \mu(f^{-1}(\{a_i\}))$$

where $\{a_1, \cdots, a_n\}$ is the range of f.

Theorem 2.7.6. *Let $(X, \mathscr{B}, \mu)$ be a measure space. Let $f, g : X \to [0, \infty)$ be simple and measurable. Then*

$$\int (f+g)d\mu = \int f d\mu + \int g d\mu$$
$$\alpha \int f d\mu = \int \alpha f d\mu$$

for all $\alpha \geq 0$.

Proof. For the first property, we can define:

$$f = \sum_{i=1}^{m} a_i \mathbf{1}_{A(i)}$$
$$g = \sum_{j=1}^{r} b_j \mathbf{1}_{B(j)}$$

where both of $A(i)$ and $B(i)$ are pairwise disjoint sequences and they cen further write as:

$$A(i) = \bigcup_{j=1}^{r} A(i) \cap B(j)$$
$$B(j) = \bigcup_{i=1}^{m} A(i) \cap B(j)$$

Then we can rewrite the above sequence as:

$$f = \sum_{i=1}^{m} \sum_{j=1}^{r} a_i \mathbf{1}_{A(i)\cap B(j)}$$
$$g = \sum_{j=1}^{r} \sum_{i=1}^{m} b_j \mathbf{1}_{A(i)\cap B(j)}$$

Therefore, $(f+g)(x) = \sum_{i=1}^{m} \sum_{j=1}^{r} (a_i + b_j) \mathbf{1}_{A(i)\cap B(j)}$. The second property holds naturally. □

We know how to integrate simple functions, and in the following part, we will study integrating the non-negative functions.

Definition 2.7.7 (Integration of Non-negative Functions)**.** *Given a measurable space $(X, \mathscr{B}, \mu)$ and a measurable function $f : X \to [0, \infty]$, the integral of f with respect to μ can be defined as:*

$$\int f d_\mu = \sup \left\{ \int \phi d\mu : \ \phi \text{ a simple function } 0 \leq \phi \leq f \right\} \tag{4}$$

Note: This means that we can use an integration of simple functions to approximate the integration of non-negative functions.

Proposition 2.7.8. *Fix $f, g : X \to [0, \infty]$ measurable,*

1. *If $f \leq g$, then $\int f d\mu \leq \int g d\mu$*
2. *If $\alpha \geq 0$ then $\alpha \int f d\mu = \int \alpha f d\mu$*

To show the integral is linear, we need to propose the Monotone Convergence Theorem.

Theorem 2.7.9 (Monotone Convergence Theorem)**.** *Given a measurable space $(X, \mathscr{B}, \mu)$ and any non-decreasing sequence $f_n : X \to [0, \infty]$ of a measurable functions, the equality*

$$\lim_{n\to\infty} \int f_n d\mu = \int \lim_{n\to\infty} f_n d\mu$$

holds.

Proof. We can set f_n is a function has the following properties:

1. f_n is non-decreasing;
2. $\lim_{n\to\infty} f_n = g$.

Then we must have

$$\begin{aligned} \int f_n d\mu &\leq \int g d\mu \text{ (for any } n) \\ \lim_{n\to\infty} \int f_n d\mu &\leq \int g d\mu \\ \lim_{n\to\infty} \int f_n d\mu &\leq \int \lim_{n\to\infty} f_n d\mu \end{aligned}$$

To prove the reverse inequality, we need to define a simple function ϕ satisfying $0 \leq \phi \leq g$, and a set such that:

$$E(n) = \{x \in X : f_n(x) \geq \alpha\phi(x)\}$$

where $\alpha \in (0, 1)$, and $\phi \leq f$. Since f_n is an increasing sequence, $E(n)$ is also increasing as well. Therefore, we have

$$\int \mathbf{1}_{E(n)} \phi d\mu = \sum_{i=1}^{r} a_r \mu(\phi^{-1}(a_r) \cap E(n)),$$

where $\{a_1, a_2, \cdots, a_n\}$ are the range of ϕ. As $n \to \infty$, $\mu(\phi^{-1}(a_r) \cap E(n)) \to \mu(\phi^{-1}(a_r))$. Therefore, we can have

$$\lim_{n\to\infty} \int f_n d\mu \geq \lim_{n\to\infty} \int \mathbf{1}_{E(n)} f_n d\mu \geq \lim_{n\to\infty} \int \mathbf{1}_{E(n)} \alpha\phi(x) d\mu = \alpha \int \phi d\mu$$

Since ϕ's supreme is g, therefore $\alpha \in (0, 1)$, we have:

$$\lim_{n\to\infty} \int f_n d\mu \geq \int \lim_{n\to\infty} f_n d\mu.$$

□

With the help of the monotone convergence theorem 2.7.9, we have

Theorem 2.7.10. *When $f, g : X \to [0, \infty]$ are measurable, the equality*

$$\int (f + g) d\mu = \int f d\mu + \int g d\mu$$

holds.

Proof. Using two non-decreasing sequences to approximate f and g when $n \to \infty$. Then using the linearity of simple function integration, we can work out this theorem. □

2.8 Lebesgue Space

In this section, we will study the Lebesgue space, defined by certain integrals' regularity. It will help us to understand the integration of real-valued functions.

Definition 2.8.1 (Measurable Function Integration)**.** *Given a measurable space $(X, \mathscr{B}, \mu)$. The integral of a measurable function $f : X \to \mathbb{R}$ is*

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu \tag{5}$$

when at least one of f^+ or f^- does not have an integral wrt μ equals ∞.

f^+ and f^- in the integration (5) can be written as

$$f^+(x) = \begin{cases} f(x) & f(x) \geq 0 \\ 0 & f(x) < 0 \end{cases}$$

and

$$f^-(x) = \begin{cases} 0 & f(x) \geq 0 \\ -f(x) & f(x) < 0 \end{cases}$$

Note: The integral of $|f|$ is always defined:

$$\int |f| d\mu = \int f^+ d\mu + \int f^- d\mu.$$

For the Lebesgue integral, we have a very useful theorem to exchange the order of integration:

Theorem 2.8.2 (Dominated Convergence Theorem)**.** *Given a measurable space $(X, \mathscr{B}, \mu)$ and a sequence $f_1, f_2, \cdots$ of measurable functions from $X \to \mathbb{R}$ that converges pointwise. If there is a function $b : X \to [0, \infty]$ with*

1. *$\int b d\mu < \infty$*
2. *$|f_n| \leq b$ for all $n \in \mathbb{N}$*

then

$$\lim_{n\to\infty} \int f_n d\mu = \int \lim_{n\to\infty} f_n d\mu$$

holds.

In the area of metric space, we have a norm

$$\|v\| = \sqrt{V_1^2 + V_2^2 + \cdots + V_d^2}$$

on $\mathbb{R}_d$, and $\|v\|$ refers to the Euclidean length of a vector. If we set μ refers to the counting measure on $\{1, \cdots, d\}$, then we have

$$\|v\| = \left(\int |v|^2 d\mu \right)^{\frac{1}{2}}.$$

Also, this formula can be extended to a more general form:

$$\|v\| = \left(\int |v|^p d\mu \right)^{\frac{1}{p}}, \tag{6}$$

where $p \geq 1$. Therefore, we can define the norm as (6) for a function $f : X \to \mathbb{R}$, and on a measurable space $(X, \mathscr{B}, \mu)$:

$$\|f\|_p = \left(\int |f|^p d\mu \right)^{\frac{1}{p}},$$

but it only works when f is integrable. Therefore, we need to define a space such that we can define a norm on the space of all functions from $X \to \mathbb{R}$ that are measurable:

$$\mathscr{L}^p(X, \mathscr{B}, \mu) = \{f : X \to \mathbb{R} \text{ measurable} : \int |f|^p d\mu < \infty\}$$

which is the $\mathscr{L}^p$. Based on the knowledge of $\mathscr{L}^p$, we can have the following inequalities.

Theorem 2.8.3 (Young's Inequality)**.** *we have*

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q} \tag{7}$$

when $x, y \geq 0$ *and* $p, q > 1$ *with* $\frac{1}{p} + \frac{1}{q} = 1$.

Theorem 2.8.4 (Hölder's Inequality)**.** *If* f *belongs to* $\mathscr{L}^p(X, \mathscr{B}, \mu)$ *and* g *belongs to* $\mathscr{L}^q(X, \mathscr{B}, \mu)$ *then*

$$\int |fg| d\mu \leq \left(\int |f|^p d\mu \right)^{1/p} \left(\int |g|^q d\mu \right)^{1/q}$$

where $p, q > 1$ *with* $\frac{1}{p} + \frac{1}{q} = 1$.

Theorem 2.8.5 (Minkowski's Inequality)**.** *For every* $p \geq 1$ *we have*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p \tag{8}$$

when $f, g \in \mathscr{L}^p(X, \mathscr{B}, \mu)$.

However $\mathscr{L}^p(X, \mathscr{B}, \mu)$ is not a norm, since it will contain non-empty sets of measure zero. For instance, when $f = \mathbf{1}_{(\{\pi\})}$, $\int |f| d\lambda = 0$ but $f \neq 0$. In order to generate a normed vector space, we need to define a set

$$\mathscr{Z}(X, \mathscr{B}, \mu) = \{f \in \mathscr{L}^p(X, \mathscr{B}, \mu) : \|f\|_p = 0\}.$$

Finally, we can work out a precise definition for Lebesgue space:

Definition 2.8.6 (Lebesgue Space)**.** *The Lebesgue Space (*$\mathbf{L}^p$*) can be defined as*

$$\mathbf{L}^p = \mathscr{L}^p(X, \mathscr{B}, \mu) \setminus \mathscr{Z}(X, \mathscr{B}, \mu)$$

Theorem 2.8.7. *For every measure space* $(X, \mathscr{B}, \mu)$ *and every* $p \geq 1$ *the Lebesgue space* $\mathbf{L}^p(X, \mathscr{B}, \mu)$*is complete as a normed vector space.*

2.9 Representing Measures

In this section, we will study the linear functional, which assigns numbers to continuous functions and act as measures in disguise.

In this section, we will start working on the function of taking complex values:

$$\int f d\mu = \int Re(f) d\mu + i \int Im(f) d\mu.$$

$f : X \to \mathbb{C}$ is measurable iff $Re(f)$ and $Im(f)$ are measurable. The Borel σ-algebra on $\mathbb{C}$ can be written as $Bor(\mathbb{C})$. For detail,

$$\{\{z \in \mathbb{C} : |z - a| < r\} : a \in \mathbb{C}, r > 0\}$$

Similar to the previous section, continuous functions from X to $\mathbb{C}$ are always $(Bor(X), Bor(\mathbb{C}))$ measurable.

Given a measure μ on $(X, Bor(X))$, every continuous function from X to $\mathbb{C}$ is bounded, if $\mu(X) < \infty$ then every continuous function belongs to $\mathscr{L}^1(X, \mu; \mathbb{C})$. Therefore, it could think μ as a functional from continuous functions to $\mathbb{C}$. In this section, we will study representing measures by functional. The space $C(X)$ of continuous functions from X to $\mathbb{C}$ is a normed vector space if we take:

$$\|f\|_u = \sup\{|f(x)| : x \in X\}.$$

For instance, we have $f(t) = \cos t + i \sin t$, which belongs to $C([0, 1])$, because $C(X)$ refers to the set of all continuous functions from X to $\mathbb{C}$, and $f(t)$ has the continuous real and imaginary part. $C(X)$ is also a vector space over $\mathbb{C}$, since:

1. $(\alpha f)(x) = \alpha f(x)$;
2. $(f + g)(x) = f(x) + g(x)$;
3. There is also a norm on $C(X)$, which is in the form of $\|f\|_u$.

In this section, we will mainly focus on the theorem:

Theorem 2.9.1. *$(C(X), \|.\|_u)$ is complete and measures can be thought as functional on $C(X)$.*

Here we have a definition for functional on $C(X)$:

Definition 2.9.2. *A functional on $C(X)$ is any bounded and linear map from $C(X)$ to $\mathbb{C}$. A map $\psi : C(X) \to \mathbb{C}$ is bounded if there is $R \geq 0$ such that*

$$|\psi(f)| \leq R\|f\|_u$$

and is linear if

1. $\psi(f+g) = \psi(f) + \psi(g)$
2. $\psi(\alpha f) = \alpha\psi(f)$

for any $f, g \in C(X)$ *and all* $\alpha \in \mathbb{C}$.

For example: Given $X = [0,1]$, we can define $\psi(f) = f(0.3)$. We have $\psi(f+g) = (f+g)(0.3) = f(0.3) + g(0.3) = \psi(f) + \psi(g)$, $\psi(\alpha f) = (\alpha f)(0.3) = \alpha f(0.3) = \alpha\psi(f)$, and finally $|\psi(f)| = |f(0.3)| \leq 1\|f\|_u$.

ψ comes from a measure. Let μ be a point mass at 0.3:

$$\mu(E) = \begin{cases} 1 & 0.3 \in E \\ 0 & \text{Otherwise} \end{cases}$$

Therefore, we have $\int f d\mu = f(0.3) = \psi(f)$. In fact, every Borel measure on X that is finite i.e. $\mu(X) < \infty$, then it determines a functional. Define

$$\psi_\mu(f) = \int f d\mu,$$

which is linear, because integration is linear and bounded

$$|\psi_\mu(f)| = |\int f d\mu| \leq \int |f| d\mu \leq \int \|f\|_\mu d\mu = \mu(X)\|f\|_\mu$$

In fact, all functionals come from measures, at least all non-negative ones. Formally,

Definition 2.9.3. *A functional* $\psi : C(x) \to \mathbb{C}$ *is non-negative if* $\forall f \geq 0$, $\psi(f) \geq 0$.

Theorem 2.9.4 (Markov)**.** *Fix* X *a compact metric space. For every non-negative functional on* $C(x)$*, there is a Borel measure* μ *on* X*, such that*

$$\int f d\mu = \psi(f)$$

Example: The Riemann integral on $[0,1]$ assigns values to all f in $C([0,1])$. It defines a linear functional on $C([0,1])$. By the representation theorem, there is a Borel measure μ on $[0,1]$ with

$$\int_0^1 f(t)dt = \int f d\mu$$

for all $f \in C([0,1])$.

3 Ergodic Theory

From this section, we will study the ergodic theory, which concerns about the limiting of the behavior over time. We will start with studying the uniform distributions.

3.1 Uniform Distribution

We will study the uniform distribution by observing the statistical properties of irrational rotation dynamics:

$$T(x) = x + \alpha \bmod 1 \tag{9}$$

where α is an irrational number, $x \in [0,1)$, and $T \in [0,1)$. We can have T^n as a result of composing T a total of n times, such that

$$\begin{aligned} T^2(x) &= T(T(x)) = x + \alpha \bmod 1 + \alpha \bmod 1 = x + 2\alpha \bmod 1 \\ T^3(x) &= T(T(T(x))) = x + 2\alpha \bmod 1 + \alpha \bmod 1 = x + 3\alpha \bmod 1 \\ &\vdots \end{aligned}$$

We can call $\{T^n : n \in \mathbb{N}\}$ the orbit of $x \in [0,1)$. We can also find its statistical properties: given $E \in [0,1)$, then

$$\frac{|\{1 \le n \le N : T^n(x) \in E\}|}{N},$$

which refers the frequency with which the orbit visit E as $N \to \infty$, which is the uniform distribution and will be proved in the following paragraph. Also, we have $|\{1 \le n \le N : T^n(x) \in E\}| = \sum_{n=1}^{N} \mathbf{1}_E(T^n(x))$.

Theorem 3.1.1. *Fix $\alpha \in \mathbb{R}$ irrational and $f : [0,1) \to \mathbb{C}$ continuous, then*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(T^n(x)) = \int f d\lambda = \int_0^1 f(t)dt$$

for $\forall x \in [0,1)$.

Proof. 1. There are special functions for which the theorem is easy to prove. For each $k \in \mathbb{Z}$ define

$$X_k(x) = e^{2\pi i k x} = \cos(2\pi k x) + i \sin(2\pi k x).$$

Fix $k \in \mathbb{Z}$ and $x \in [0,1)$,

$$\begin{aligned}\sum_{n=1}^{N} X_k(T^n(x)) &= \sum_{n=1}^{N} X_k(x+na) \\ &= \sum_{n=1}^{N} X_k(x)X_k(a)^n \\ &= X_k(x)\sum_{n=1}^{N} X_k(a)^n \\ &= X_k(x)\frac{(1-X_k(\alpha))^{n+1}}{1-X_k(\alpha)}\end{aligned}$$

In this transformation, we require $X_k(\alpha) \neq 1$. Then we have:

$$\frac{1}{N}\sum_{n=1}^{N} X_k(T^n(x)) = \frac{1}{N}X_k(x)\frac{(1-X_k(\alpha))^{n+1}}{1-X_k(\alpha)}$$

$|X_k(\alpha)| = 1 \Rightarrow |1 - X_k(\alpha)^{n+1}| \leq 2$. Therefore,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} X_k(x) = 0 = \int_0^1 X_k(t)dt$$

When $k = 0$, $X_0(x) = 1$ for any $x \in \{0,1\}$:

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} X_0(x) = 1 = \int_0^1 X_0(t)dt.$$

Theorem 3.1.2. *For every $f \in ([0,1])$ and every $\epsilon > 0$, there is $a_{p(1)}, \cdots, a_{p(r)} \in \mathbb{C}$ with*

$$\| a_{p(1)}X_{q(1)} + \cdots + a_{p(r)}X_{q(r)} - f \|_u \leq \epsilon$$

$$|\int_0^1 f(t)dt - \int_0^1 a_{p(1)}X_{q(1)} + \cdots + a_{p(r)}X_{q(r)}dt| < \epsilon$$

$$|\sum_{n=1}^{N} f(T^n(x)) - \sum_{n=1}^{N} a_{p(1)}X_{q(1)} + \cdots + a_{p(r)}X_{q(r)}| \leq N\epsilon \tag{10}$$

By using these two theorems, we have shown the orbit follows uniform distribution. □

Theorem 3.1.3 (Uniform Distribution)**.** *Fix α irrational, for all $0 \leq a < b \leq 1$ we have*

$$\lim_{N\to\infty}\frac{|\{0 \leq n \leq N : T^n(x) \in [a,b)\}|}{N} = b - a$$

3.2 The Law of Large Numbers

In this section, we will study the law of large numbers which is an important result in probability theory.

Consider the outer measure Ξ_p on $\{0,1\}^{\mathbb{N}}$, then we can have a probability assigning the size:

$$p^{|1\leq i\leq r:\epsilon(i)=1|}(1-p)^{|1\leq i\leq r:\epsilon(i)=0|},$$

to the cylinder sets $[\epsilon(1)\cdots\epsilon(r)]$. By the Carathédory's theorem, we can define a measure ξ_p be the Borel measure on X from Ξ_p, such that

$$\xi_p[\epsilon(1)\cdots\epsilon(r)]p^{|1\leq i\leq r:\epsilon(i)=1|}(1-p)^{|1\leq i\leq r:\epsilon(i)=0|}$$

In probability, we would like to understand the empirical average as the frequency:

$$\frac{|\{1\leq n\leq N: x(N)=1\}|}{N},$$

which is the frequency of heads of N coin tosses. As $N\to\infty$, we would like to claim we have the expected frequency, which is the 'expected' frequency and this idea could be formalised as the strong law of large numbers:

Theorem 3.2.1 (Strong Law of Large Numbers)**.** *For all $0\leq p\leq 1$. then*

$$\xi_p(\{x\in\{0,1\}^{\mathbb{N}}:\lim_{N\to\infty}\frac{|\{1\leq n\leq N: x(n)=1\}|}{N}=p\})=1 \tag{11}$$

To prove theorem 3.2.1, we will also need to study the following statements:

Lemma 3.2.2. *(Borel-Cantelli Lemma) Fix a measure space $(X,\mathscr{B},\mu)$ with $\mu(X)<\infty$. Given $A_1,A_2,A_3,\cdots\in\mathscr{B}$ if*

$$\sum_{n=1}^{\infty}\mu(A_n)<\infty$$

then

$$\mu\left(\bigcap_{m\in\mathbb{N}}\bigcup_{N>m}A_n\right)=0$$

$\left(\bigcap_{m\in\mathbb{N}}\bigcup_{N>m}A_n\right)$ *often refers to*

$$\{x\in X: x\in\inf A_n\}$$

Lemma 3.2.3 (Markov Inequality)**.** *Fix a measurable space $(X,\mathscr{B},\mu)$ with $\mu(x)<\infty$. For every measurable $f:X\to[0,\infty)$, we have:*

$$\mu(\{x\in X: f(x)\geq S\})\leq\frac{1}{S}\int f d\mu$$

Here is a sketch of how to prove the strong law of large numbers 3.2.1:

Proof. 1. Define a new variable $Y_n = x(n) - p$, then we can prove:

$$\lim_{N\to\infty} \frac{Y_1(x) + \cdots + Y_N(x)}{N} = 0 \tag{12}$$

which is equivalent to showing the equation (11) holds.

2. Using the setting of Y_n, we can define a set:

$$B = \left\{ x \in \{0,1\}^{\mathbb{N}} : \lim_{N\to\infty} \frac{Y_1(x) + \cdots + Y_N(x)}{N} \neq 0 \right\}$$

Proving $\xi_p(B) = 0$ is equivalent to showing equation (12). The set B can also be written as

$$B = \bigcup_{r\in\mathbb{N}} \bigcap_{M\in\mathbb{N}} \bigcup_{N\geq M} \{x \in \{0,1\}^N : |\frac{Y_1(x) + \cdots + Y_N(x)}{N}| \geq \frac{1}{r}\}$$

Since $\bigcap$ means exist, therefore it is sufficient to prove

$$C = \bigcap_{M\in\mathbb{N}} \bigcup_{N\geq M} \{x \in \{0,1\}^N : |\frac{Y_1(x) + \cdots + Y_N(x)}{N}| \geq \frac{1}{r}\}$$

has a zero measure.

3. To prove C (13) has a zero measure, we can use the Borel-Cantelli lemma 3.2.2 and in this case

$$A_n = \{x \in \{0,1\}^N : |\frac{Y_1(x) + \cdots + Y_N(x)}{N}| \geq \frac{1}{r}\}.$$

which requires us to prove $\xi_p(A_n)$ is summable.

4. To prove $\xi_p(A_n)$, we can use the Markov inequality (3.2.3) to show it is bounded.

□

3.3 Measure-Preserving Transformation

Dynamics is the study of iterations. Given a map $T : X \to X$, we want to understand for points x in X in the sequence obtained by repeated application of T:

$$x, T(x), T(T(x)), T(T(T(x))), \cdots \tag{13}$$

by the dynamics. This sequence is also known as the orbit of x. For simplicity, we will use $T^n(x)$ to denote the point obtained by n application of T.

Ergodic theory is to study the probabilistic and statistical properties of the dynamical systems. We have studied the first aspects the empirical averages (frequency):

$$\frac{|\{1 \leq n \leq N : T^n(x) \in E\}|}{N},$$

which the orbit belongs to a given set $E \in X$. In ergodic theory, we start with the analysis of such frequencies in situations where the map T was related to the problems in statistical physics. It has been studied in the previous two sections. In this section, we will study the second aspects invariant measures.

In ergodic theory we have a tuple $(X, \mathscr{B}, \mu, T)$, in which $(X, \mathscr{B}, \mu)$ is a measure space, μ is the probability measure. The invariant measures mean that $T : X \to X$ is measurable and measure preserving. Before formally stating what is measure preserving, we can begin with two examples:

1. We have the following settings: $X = [0, 1)$, $\mathscr{B} = Borel(x)$, $\mu = \lambda$, and $T(x) = x + \alpha \mod 1$. Fix $0 \leq a < b \leq 1$:

   $$T^{-1}([a, b]) = \{x \in [0, 1) : T(x) \in [a, b)\} = [a - \alpha, b - \alpha]$$

   or unions of two half-open intervals, which has been illustrated by the graph 2.

2. The settings are $X = \{0, 1\}^{\mathbb{N}}$, $\mathscr{B} = Borel(x)$, $\mu = \xi_p$, and$(T(x))(n) = x(n + 1)$. Fix a cylinder $[\epsilon(1), \cdots, \epsilon(r)]$, then we need to find the $T^{-1}([\epsilon(1), \cdots, \epsilon(r)]) = \{x \in \{0, 1\}^{\mathbb{N}} : T(x) \in [\epsilon(1), \cdots, \epsilon(r)]\}$:

   $$(T(x))(1) = \epsilon(1), (T(x))(2) = \epsilon(2), \cdots, (T(x))(r) = \epsilon(r)$$

   which equivalent to

   $$x(2) = \epsilon(1), x(3) = \epsilon(2), \cdots, x(r + 1) = \epsilon(r).$$

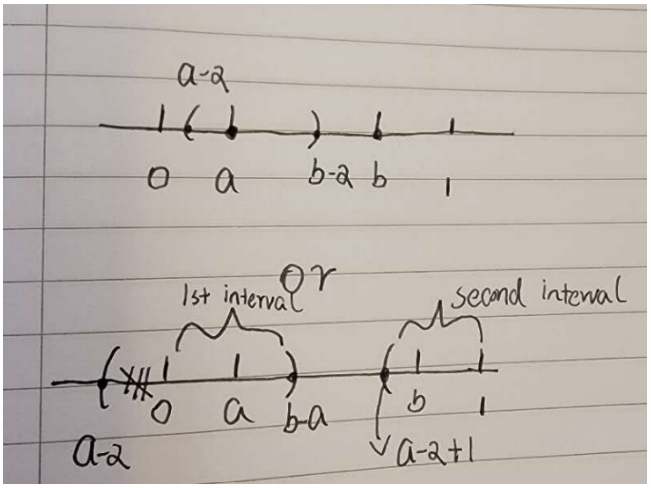

Figure 2: This graph illustrates the $T^{-1}([a,b])$, which shows that it is a closed interval $[a-\alpha, b-\alpha]$ or union of two half-open intervals $(0, b-\alpha] \cup [a-\alpha+1, 1)$. Both of them have same length with $[a,b]$.

Therefore, we have

$$T^{-1}([\epsilon(1), \cdots, \epsilon(r)]) = [0, \epsilon(1), \cdots, \epsilon(r)] \cup [1, \epsilon(1), \cdots, \epsilon(r)].$$

Based on the above examples, it could be found that given a map $T : X \to X$ is $(\mathscr{B}, \mathscr{B})$ measurable, and the measure μ or ξ_p is T-invariant, since the T does not change the sizes of subsets of X in the sense that

$$\mu(T^{-1}(B)) = \mu(B)$$
$$\xi_p(T^{-1}(B)) = \xi_p(B)$$

for all $B \in \mathscr{B}$. In the probabilistic view, we could understand that an event is no more or less likely than it was yesterday. We can further generalize this idea into theories:

Definition 3.3.1 (Measure-preserving)**.** *Fix a probability space $(X, \mathscr{B}, \mu)$, a measurable map $T : X \to X$ is measure preserving if*

$$\mu(T^{-1}(B)) = \mu(B)$$

for all $B \in \mathscr{B}$.

and

Definition 3.3.2 (Invariant Measure)**.** *Fix a measure space $(X, \mathscr{B})$ and a measurable map $T : X \to X$. A measure μ on $(X, \mathscr{B})$ is invariant for T if*

$$\mu(T^{-1}(B)) = \mu(B)$$

for all $B \in \mathscr{B}$.

We can set an example to illustrate the theories above. For instance, we have $X = [0, 1)$, $B = [0, log(2))$, then we have a measure μ such that $\mu(B)$ is the likelihood for a random point in X less than $log(2)$. Given a map T, we have a set:

$$T^{-1}(B) = \{x \in X : T(x) \in B\}$$

which contains points $x \in X$ and will be in B for a moment later. If the map T is measure-preserving, the future $T(x)$ has same possibility with x for less than $log(2)$.

Before going further, we need to learn the tool named $\pi - \lambda$ theorem, which can help us to check whether a given measurable map is measure preserving or not. The $\pi - \lambda$ theorem tells us that two measures agree on special collections of sets in $\mathscr{B}$ known as π systems.

Definition 3.3.3 (π System)**.** *Fix a set* X*, a collection* $D \subset \mathcal{P}(X)$ *is a* π *system if* $\mathcal{D}$ *is non-empty and whenever* A, B *belong to* $\mathcal{D}$ *one has* $A \cap B$ *in* $\mathcal{D}$ *as well.*

In this course, we have two particular examples:

1. (The π system of intervals). Take $X = [0, 1)$, the collection:

$$\{[a, b) : 0 \leq a < b \leq 1\} \cup \{\varnothing\}$$

   is a π system on X.

2. (The π system of cylinders). Take $X = \{0, 1\}^{\mathbb{N}}$, the collection:

$$\{C \subset X : C \text{ is a cylinder}\} \cup \{\varnothing\}$$

   is a π system on X.

These two examples will be important in this course. Then we can use the $\pi - \lambda$ system to verify the measurable map is measure-preserving.

Theorem 3.3.4 (The $\pi - \lambda$ Theorem)**.** *Fix a measurable space* $(X, \mathscr{B})$*,* μ *and* ν *are two measures on this space. If* D *is a* π *system with*

1. $\sigma(D) \supset \mathscr{B}$
2. $D \in \mathcal{D} \Rightarrow \mu(D) = \nu(D)$

then $\mu = \nu$

Corollary 3.3.5. *Fix a measure space $(X, \mathscr{B}, \mu)$ and a π system $\mathcal{D}$ with $\sigma(\mathcal{D}) \supset \mathscr{B}$. A measurable map $T : X \to X$ is measure-preserving if*

$$\mu(D) = \mu(T^{-1}(D))$$

for all $D \in \mathcal{D}$.

Proof. We can define $\nu(B) = \mu(T^{-1}(B))$ and ν is a measure on $(X, \mathscr{B})$. By the $\pi - \lambda$ theorem, we have

$$\mu(T^{-1}(B)) = \nu(B) = \mu(B).$$

□

Going back to the examples at the beginning of this section, it could be observed that $\mu([a, b]) = \mu(T^{-1}([a, b]))$ and $T^{-1}([\epsilon(1), \cdots, \epsilon(r)]) = [0, \epsilon(1), \cdots, \epsilon(r)] \cup [1, \epsilon(1), \cdots, \epsilon(r)]$ has the same measure with $[\epsilon(1), \cdots, \epsilon(r)]$.

We have studied the first result about the measure-preserving dynamical systems. According to the Poincaré recurrence, the events $B \in \mathscr{B}$ will happen infinitely often as one continues to iterate the dynamics.

Theorem 3.3.6 (Poincaré Recurrence)**.** *We have a system $(X, \mathscr{B}, \mu, T)$, then $\mu(B) > 0$ for every $B \in \mathscr{B}$, and there is $n \in \mathbb{N}$ such that $\mu(B \cap (T^n)^{-1}(B)) > 0$.*

Proof. This theorem can be proved by the contradiction. Suppose for any $n \in \mathbb{N}$, we have

$$\mu(B \cap (T^n)^{-1}(B)) = 0.$$

Due to T is measure preserving, therefore we have

$$\mu\big((T^k)^{-1}(B) \cap (T^{k+n})^{-1}(B)\big) = 0.$$

for all $n \in \mathbb{N}$, which means B and each $(T^n)^{-1}(B)$ is pointwise disjoint. Then we have

$$1 = \mu(X) \geq \mu(B \cup T^{-1}(B) \cup \cdots \cup (T^N)^{-1}(B)) = N\mu(B)$$

However if $N > \mu(B)^{-1}$, we have a contradiction. □

3.4 Ergodicity

We have studied the idea pf measurable and measure-preserving maps $T : X \to X$ where $(X, \mathscr{B}, \mu)$ is a fixed probability space. Also, the Poincaré recurrence theorem is a powerful tool, which tells us that when $\mu(B) > 0$ one has

$$\mu(B \cap (T^n)^{-1}(B)) > 0$$

for some $n \in \mathbb{N}$. In this section, we will study the ergodicity. We will start with the invariant sets which help us to justify the ergodicity.

After studying the Poincaré recurrence theorem, we are curious about that given two disjoint sets $A, B \in \mathscr{B}$ with $\mu(A) > 0$ and $\mu(B) > 0$ do we always have:

$$\mu(A \cap (T^n)^{-1}B) > 0$$

for some $n \in \mathbb{N}$?

If we assume this statement is not correct, it means that:

$$A \cap \Big(\bigcup_{n\in\mathbb{N}} (T^n)^{-1}B \Big)$$

has a zero measure. We set

$$C = \bigcup_{n\in\mathbb{N}} (T^n)^{-1}B$$

that satisfies $T^{-1}(C) \subset C$. Since $\mu(T^{-1}(C)) = \mu(C)$, we have

$$Y = \bigcap_{j\in\mathbb{N}} (T^j)^{-1}C = \bigcap_{j\in\mathbb{N}} \bigcup_{n>j} (T^n)^{-1}(B)$$

has a positive measure equal to $\mu(C)$. Moreover, if $x \in Y$ then $T(x) \in Y$ as well. We therefore have within X a set Y of positive measure but not full measure that the dynamics never escape. Therefore, we claim this dynamic is in isolation on y.

If the statement is correct, it means that all points in X will go almost everywhere that they will visit every sets with a positive measure. This means that X cannot be broken up into pieces that are themselves measure-preserving systems. In this case, we can say that this T is an ergodic, and we can provide a formalized definition:

Definition 3.4.1 (Ergodicity)**.** *A measure-preserving transformation T on a probability space $(X, \mathscr{B}, \mu)$ is ergodic if*

$$\mu(B) > 0 \Rightarrow \mu\left(\bigcup_{n\in\mathbb{N}} (T^n)^{-1}B \right) = 1$$

for every $B \in \mathscr{B}$.

Based on this definition, we have a theorem to show a a measure-preserving transformation T is an ergodic.

Theorem 3.4.2. *A measure-preserving transformation T on a probability space $(X, \mathscr{B}, \mu)$ is ergodic iff*

$$B = T^{-1}(B) \Rightarrow \mu(B) \in \{0, 1\}$$

for every $B \in \mathscr{B}$.

Note: In this theorem, we can conclude that being an ergodic means that we cannot decompose X into a group of T-invariant subsets with non-trivial measures. If we can find an $B \in \mathscr{B}$ such that $0 < \mu(B) < 1$, T on $(X, \mathscr{B}, \mu)$ is not ergodic.

Since we have studied the criteria of how to justify whether a mapping (transformation) is ergodic or not, in the rest paragraphs of this section, we will justify the irrational rotations and full shifts are ergodic.

Fix α irrational, take $X = [0, 1)$, we have an irrational rotation:

$$T(x) = x + \alpha \mod 1$$

for all $x \in [0, 1)$. Then we have a statement:

Theorem 3.4.3. *The system $(X, \mathscr{B}, \lambda, T)$ is ergodic, where λ is a Lebesgue measure to [0,1).*

We will show it as an example of how to apply the theorem to justify a mapping is ergodic.

Proof. To prove this, we need to check

$$\lambda(I \cap B) = \lambda(B)\lambda(I)$$

for any $I = [a, b)$.

$$\begin{aligned}
\lambda(I \cap B) &= \int \mathbf{1}_I \mathbf{1}_B d\lambda \\
&= \frac{1}{N} \sum_{n=1}^{N} \int \mathbf{1}_I \circ T \cdot \mathbf{1}_B d\lambda \\
&= \lim_{N \to \infty} \int \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}_I(T^n x) \cdot \mathbf{1}_B(x) d\lambda(x) \\
&= \lambda(I)\lambda(B)
\end{aligned}$$

by the DCT and the uniform distribution theorem. Now we can define a new measure ν:

$$\nu(E) = \frac{\lambda(B \cap E)}{\lambda(B)}$$

for all $E \in \mathscr{B}$. Then we have:

$$\nu(I) = \lambda(I)$$

$I = [a, b)$. Then the $\pi - \lambda$ system states that $\nu = \lambda$. Therefore,

$$\nu(B) = \frac{\lambda(B \cap B)}{\lambda(B)} = 1.$$

Therefore T is an ergodic. □

Take $X = \{0, 1\}^{\mathbb{N}}$ and Let $T : X \to X$ be shift map. Put

$$\mu = \frac{1}{2}\delta_a + \frac{1}{2}\delta_b$$

where

$$a = 000000000000\cdots$$

$$b = 111111111111\cdots.$$

are the constants sequences in X. With respect to μ, the map T is not ergodic, since the sets $\{a\}$ and $\{b\}$ are T invariant ($T(a) = a$ and $T(b) = b$) but both have the measure $\frac{1}{2}$.

Next, we can look at the measures μ_p for $0 \leq p \leq 1$ fixed. Let $B \in \mathscr{B}$ be $T-$invariant set. We have

$$\begin{aligned}
\mu(B \cap [1]) &= \lim_{N\to\infty} \sum_{n=1}^{N} \mu(B \cap (T^n)^{-1}[1]) \\
&= \lim_{N\to\infty} \int \mathbf{1}_B \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}_{(T^n)^{-1}[1]} d\mu \\
&= \int \mathbf{1}_B \frac{1}{N} \lim_{N\to\infty} \sum_{n=1}^{N} \mathbf{1}_{(T^n)^{-1}[1]} d\mu \text{ (DCT)} \\
&= p \int \mathbf{1}_B d\mu = \mu([1])\mu(B) \text{ (Strong Law of Large Number)}
\end{aligned}$$

In fact, by a strengthening of the strong law of large numbers, we have

$$\mu(B \cap C) = \mu(B)\mu(C),$$

for every cylinder set C. We can define a new measure $v(C) = \frac{\mu(B\cap C)}{\mu(C)}$. By the $\Pi - \lambda$ theorem, we have $V(B) = 1 = \mu(B)$. Therefore, shift map is ergodic.

3.5 The Pointwise Ergodic Theorem

The ergodicity means that before the point returns to its origin under the mapping, it will 'run' through the whole space. Given $(X, \mathscr{B}, \mu, T)$ to be ergodic, we have

$$\mu(\{x \in X : T^n(x) \in B \text{ infinitely often}\}) = 1 \tag{14}$$

when $B \in \mathscr{B}$ satisfies $B > 0$. Although the equation (14) means that almost every point $x \in X$ will visit B infinitely often. However, we can find a 'correct' frequency with this behaviour, which is also the content of the pointwise ergodic theorem.

Theorem 3.5.1 (Pointwise Ergodic Theorem)**.** *Let $(X, \mathscr{B}, \mu)$ be a probability space and T is a measure-preserving map $T : X \to X$ that is ergodic. For every measurable and integrable function $f : X \to \mathbb{R}$ there is a set $\Omega \in \mathscr{B}$ with $\mu(\Omega) = 1$ and*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n(x)) = \int f d\mu.$$

for all $x \in \Omega$.

Taking $f = \mathbf{1}_E$ says that

$$\lim_{N\to\infty} \frac{|\{1 \le n \le N : T^n(x) \in E\}|}{N} = \mu(E).$$

for μ almost-every $x \in X$.

Based on the pointwise ergodic theorem, we can summarize the strong law of large numbers:

Theorem 3.5.2 (Strong Law of Large Numbers)**.** *Given $X = \{0,1\}^{\mathbb{N}}$, $B = Borel(X)$, $\mu = \xi_p$ $(0 \le p \le 1)$, and $(T(x))(n) = x(n+1)$, then we have*

$$\xi_p\left(\left\{X = \{0,1\}^{\mathbb{N}} : \frac{X(1) + \cdots + X(N)}{N} \to p\right\}\right) = 1.$$

This is a consequence of the pointwise ergodic theorem since we can set f.

Also, we can rewrite the uniform distribution:

Theorem 3.5.3 (Uniform Distribution)**.** *Given $X = [0,1)$, $\mathscr{B} = Borel([0,1))$, μ is the Lebesgue measure o $x + \alpha \mod 1$, then the uniform distribution says that for any $x \in [0,1)$, we have*

$$\lim_{N\to\infty} \frac{|\{1 \le n \le N : T^n(x) \in [a,b)\}|}{N} = b - a$$

for any $[a,b) \subset [0,1)$.

Note: This is statement does not follow the pointwise ergodic theorem. A bit more special, if we check that T is ergodic, we can apply the pointwise ergodic theorem, but different conclusion.

3.6 Normal Number

In this section, we will apply the pointwise ergodic theorem to deduce the result of Borel that says almost every number is normal in base 10. In this note, we will do this by studing the map

$$T(x) = 10x \mod 1$$

on $[0, 1)$ equipped with Lebesgue measure λ on the Borel subsets of $[0, 1)$.

Theorem 3.6.1. *The map $T(x) = 10x \mod 1$ is ergodic wrt Lebesgue measure.*

Proof. Fix an interval $[a, b)$ with $0 \leq a < b \leq 1$, combining it with an empty set, we can get a $\pi-$system. Also,

$$T^{-1}([a, b)) = \bigcup_{i=0}^{9} [\frac{a+i}{10}, \frac{b+i}{10})$$

which is disjoint union of intervals whose length equals to $b - a$. Then, we can claim λ is an invariant measure for T.

To utilize the $\pi - \lambda$ theorem showing T is ergodic, we can define a new measure v by

$$v(E) = \frac{\lambda(B \cap E)}{\lambda(B)}, \tag{15}$$

on the Borel subset of $[0, 1)$. We will check that $v(I) = \lambda(I)$ for every interval of the form:

$$I = [\frac{a}{10^r}, \frac{a+1}{10^r})$$

with $r \in \mathbb{N}$ and $0 \leq a < 10^r$. This interval I is called decimal interval and can generate the Borel σ-algebra. Then $v(B) = \lambda(B) = 1$. To show (15), we need to prove

$$\lambda(B \cap E) = \lambda(B)\lambda(E).$$

For the interval I, we have

$$\lambda(B \cap I) = \lambda(B \cap (T^n)^{-1} I)$$

since λ is T invariant and $T^{-1}(B) = B$. Next, fixing $\epsilon > 0$ we choose decimal intervals J_1, J_2, $\cdots$ with

$$B \subset J_1 \cup J_2 \cup \cdots$$

and

$$\lambda(B) + \epsilon \geq \lambda(J_1) + \lambda(J_2) + ... \geq \lambda(J_1 \cup J_2 \cup \cdots) \geq \lambda(B).$$

Using the definition of Lebesgue measure, we have

$$\left|\lambda(B\cap I)-\lambda\Big(\bigcup_{t=1}^{\infty}J_t\cap(T^n)^{-1}I\Big)\right|<\epsilon$$

for all $n\in\mathbb{N}$. One calculates for every $K\in\mathbb{N}$ that

$$\lim_{n\to\infty}\lambda\Big(\bigcup_{t=1}^{K}J_t\cap(T^n)^{-1}I\Big)=\lambda\Big(\bigcup_{t=1}^{K}J_t\Big)\lambda(I)$$

and thus we have

$$\left|\lambda(B\cap I)-\sum_{t=1}^{\infty}\lambda(J_t)\lambda(B)\right|<\epsilon.$$

Since ϵ is arbitrary, we can prove $\lambda(B\cap I)=\lambda(B)\lambda(I)$. □

Since T is ergodic, then we can apply the pointwise ergodic theorem. Letting $f=\mathbf{1}_{[0,0.1]}$, we have

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}\mathbf{1}_{[0,0.1]}(T^n(x))=\int\mathbf{1}_{[0,0.1]}d\lambda=\frac{1}{10}.$$

This means that almost every point $x\in[0,1)$ has the property that the sequence

$$n\to T^n(x)$$

spends about 10% of its time in the interval $[0,0.1]$. Formally, the pointwise ergodic theorem tells us a set $\Omega_0\subset[0,1)$ with $\lambda(\Omega_0)=1$ such that every $x\in\Omega_0$ has the desired property.

We can reinterpret this by expressing

$$x=\sum_{j=1}^{\infty}\frac{d_j(x)}{10^j}\tag{16}$$

where each $d_j(x)$ is natural numbers between zero and nine i.e., by writing the decimal expansion of x. In those terms we have $x\in[0,0.1)$ iff $d_1(x)=0$. Moreover

$$\begin{aligned}T(x)&=10\Big(\frac{d_1(x)}{10}+\frac{d_2(x)}{10^2}+\frac{d_3(x)}{10^3}+\cdots\Big)\mod 1\\&=d_1(x)+\frac{d_2(x)}{10}+\frac{d_3(x)}{10^2}+\cdots\\&=\frac{d_2(x)}{10}+\frac{d_3(x)}{10^2}+\cdots\end{aligned}$$

and $T(x)$ belongs to $[0,0.1]$ iff the second digit in the decimal expansion of x is a zero. By repeating applying T as above, we have

$$\sum_{n=0}^{N-1}\mathbf{1}_{[0,0.1]}(T^n(x))$$

counts how many of the first N digits in the decimal expansion of x are equal to 0.

For example, we can set $f = \mathbf{1}_{[0.47,0.48]}$, the sum

$$\sum_{n=0}^{N-1} \mathbf{1}_{[0.47,0.48]}(T^n(x)),$$

which counts the occurrences of 47 amongst the decimal digits of x, and it can be computed by pointwise ergodic theorem letting Ω_{47} has the full measure.

In fact, for every finite string σ of digits between zero and nine we get a set Ω_σ with $\mu(\Omega_\sigma) = 1$ and the property that every $x \in \Omega_\sigma$ has the expected frequency $1/10^l$ where l is the number of digits in σ- of appearances of σ in its decimal expansion. A number with this property is sometimes called normal in base 10.

3.7 Koopman Operator

Let T be a measure preserving transformation on a probability space $(X, \mathscr{B}, \mu)$, and T is also a map from X to itself. Therefore, we can use a composition to induce a map on spaces of functions on X. For instance, given $f : X \to \mathbb{C}$ we can define a new function Tf by

$$(Tf)(x) = f(T(x))$$

for all $x \in X$. Thus Tf can be written as $f \circ T$. In the previous section, we have studied T by its direct effect on points in X. In this section, we will study T by its indirect effect on functions. At the beginning of this section, we can provide some basic facts. Given any two functions f, g and any $\alpha \in \mathbb{C}$ we have

$$T(f + g) = Tf + Tg$$
$$T(\alpha f) = \alpha(Tf)$$

where T acts linearly on functions. In this section, we will explore how T affects functions in the Lebesgue space $\mathbf{L}^2(X, \mathscr{B}, \mu)$.

Lemma 3.7.1. *Fix a measurable space $(X, \mathscr{B})$ and a measurable map $T : X \to X$. If $f : X \to \mathbb{R}$ is $(\mathscr{B}, Bor(\mathbb{C}))$ measurable then $f \circ T$ is also $(\mathscr{B}, Bor(\mathbb{C}))$ measurable.*

Proof. Given $B \in Bor(\mathbb{C})$, we have

$$(f \circ T)^{-1}(B) = T^{-1}(f^{-1}(B)).$$

$f^{-1}(B)$ is always in $\mathscr{B}$, and then we can always have $(f \circ T)^{-1}(B) \in \mathscr{B}$. □

Lemma 3.7.2. *Fix a probability space $(X, \mathscr{B}, \mu)$ and a measure-preserving map $T : X \to X$. If f belongs to $\mathscr{L}^2(X, \mathscr{B}, \mu)$ then Tf does as well.*

Based on these two lemmas, we have

Theorem 3.7.3. *Fix a measure-preserving transformation T of a probability space $(X, \mathscr{B}, \mu)$. For every $f \in \mathbf{L}^2(X, \mathscr{B}, \mu)$ the composition Tf also belongs to $\mathbf{L}^2(X, \mathscr{B}, \mu)$.*

The mapping $\mathbf{L}^2(X, \mathscr{B}, \mu) \to \mathbf{L}^2(X, \mathscr{B}, \mu)$ defined by $f \to Tf$ is the Koopman operator.

Since we have defined the Koopman operator, we will start studying the dynamical properties of T through the Koopman operator. In the previous sections, we define T to be ergodic iff all invariant sets have measure 0 or measure 1. Since

$$T\mathbf{1}_E = \mathbf{1}_{T^{-1}E} = 1_E,$$

when E is invariant. Therefore, it seems reasonable to look at those functions in $\mathbf{L}^2(X, \mathscr{B}, \mu)$ that fixed by the Koopman operator.

Theorem 3.7.4. *A measure-preserving transformation T on $(X, \mathscr{B}, \mu)$ is ergodic iff all T invariant elements of $\mathbf{L}^2(X, \mathscr{B}, \mu)$ are represented by constant functions.*

Proof. Suppose every invariant function f in $\mathbf{L}^2(X, \mathscr{B}, \mu)$ is equal to almost-everywhere to a constant function. This means the following: if f is invariant then there is $c \in \mathbb{C}$ such that

$$\mu(\{x \in X : f(x) = c\}) = 1$$

holds. Fix $E \subset X$ with $T^{-1}E = E$. Then $T_{1_E} = 1_E$, and 1_E is a member of $\mathbf{L}^2(X, \mathscr{B}, \mu)$ and fixed by the Koopman operator. According to the hypothesis, we have $1_E = c$ almost everywhere and c is a constant from $\mathbb{C}$. As 1_E only takes value in $\{0, 1\}$, and it must the case $c \in 0, 1$ and therefore $\mu(E) \in \{0, 1\}$.

For the reverse statement, we are given T is ergodic and that f from $\mathbf{L}^2(X, \mathscr{B}, \mu)$ is fixed by the Koopman operator. Suppose f is real-valued function, fix $k \in \mathbb{N}$, we have

$$E(k, N) = \{x \in X : \frac{N}{2^k} \leq f(x) \leq \frac{N+1}{2^k}\}$$

for each $N \in \mathbb{Z}$. We have

$$\mu(E(k, N) \Delta T^{-1} E(k, N)) = 0$$

and therefore $\mu(E(k, N)) \in \{0, 1\}$. Exactly one of these sets can have full measure. As we increase k we converge to a specific value $c \in \mathbb{R}$ with

$$\mu(\{x \in X : f(x) = c\} = 1$$

and therefore f is a constant everywhere. □

In analogy with linear algebra, we will look for eigenfunction of T.

Definition 3.7.5. *By an eigenfunction of a measure-preserving transformation T we mean a member f of $\mathbf{L}^2(X, \mathscr{B}, \mu)$ such that $Tf = \eta f$ for some $\eta \in C$.*

We can provide an example for this definition. Given an irrational rotation $T(x)$ there are lots of eigenfunction. Indeed

$$\psi_k(x + \alpha) = e^{2\pi i k(x+\alpha)} = e^{2\pi i k\alpha} \psi_k(x).$$

Therefore

$$T\psi_k(x) = e^{2\pi ik\alpha}\psi_k(x)$$

and ψ_k is an eigenfunction of T with eigenvalue $e^{2\pi ik\alpha}$.

In contrast with the existence of eigenfunctions, we have the property of mixing.

Definition 3.7.6. *A measure-preserving transformation T on a probability space $(X, \mathscr{B}, \mu)$ is mixing if*

$$\lim_{n\to\infty} \langle f, T^n g\rangle = \langle f, 1\rangle\langle 1, g\rangle$$

for all $f, g \in \mathbf{L}^2(X, \mathscr{B}, \mu)$.

3.8 Fourier Series

We have seen that a measure-preserving transformation T on a probability space $(X, \mathscr{B}, \mu)$ induces an isometry of space $\mathbf{L}^2(X, \mathscr{B}, \mu)$ by composition. That is, if f belongs to $\mathbf{L}^2(X, \mathscr{B}, \mu)$ then so too does

$$Tf = f \circ T$$

and moreover $\|Tf\|_2 = \|f\|_2$.

In this section, we will focus on the special case where our probability space is $[0, 1)$ equipped with Lebesgue measure λ. We will produce an orthonormal basis of $\mathbf{L}^2(X, \mathscr{B}, \mu)$ that will allow us more easily verify dynamical properties of measure-preserving transformations on the unit interval.

We can define a function

$$\chi_k(x) = e^{2\pi ikx} = \cos(2\pi kx) + i\sin(2\pi kx)$$

for each $x \in [0, 1)$. Since $|\chi_k| = 1$ for each $K \in \mathbb{Z}$ and $\lambda([0, 1)) = 1$ each of the functions χ_k belongs to $\mathbf{L}^2([0, 1), \mathscr{B}, \lambda)$. Recall the inner product, we have the following theorem

Theorem 3.8.1. *Fix $f, g \in \mathbf{L}^2$, we have*

$$\langle f, g\rangle = \int f\overline{g}d\mu$$

and

$$\langle Tf, Tg\rangle = \langle f, g\rangle$$

Therefore, we have an inner product

$$\langle \chi_k, \chi_j\rangle = \int \chi_k\overline{\chi_j}d\lambda = \int_0^1 e^{2\pi i(k-j)x}dx = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}$$

which means the collection $\{\chi_k : k \in \mathbb{Z}\}$ of functions is an orthonormal system. We have two conclusions:

$$f = \sum_{n\in\mathbb{Z}} \langle f, \chi_k\rangle\chi_k$$

$$\lim_{N\to\infty} \left\| f - \sum_{n=-N}^{N} \langle f, \chi_k\rangle\chi_k \right\|_2 = 0$$

for every $f \in \mathbf{L}^2([0, 1), \mathscr{B}, \lambda)$.

We can use this Fourier series to prove the ergodicity.

Proposition 3.8.2. *Every irrational rotation is ergodic.*

Proof. Given $T(x) = x + \alpha \mod 1$ for some irrational α. Fix f in $\mathbf{L}^2([0,1), \mathscr{B}, \lambda)$ that is T invariant. We can write

$$f = \sum_{n \in \mathbb{Z}} c(n)\chi_n,$$

where $c_n = \langle f, \chi_n \rangle$ and calculate that

$$Tf = \sum_{n \in \mathbb{Z}} c(n) T\chi_n = \sum_{n \in \mathbb{Z}} c(n) e^{2\pi i n a} \chi_n.$$

We have $c(n) = c(n)e^{2\pi i n a}$ for all $n \in \mathbb{Z}$. As α is irrational we always $c_n = 0$ for all non-zero n. Therefore $f = c(0)$ is a constant and by Theorem 3.7.4, we can say T is ergodic. □

Similarly, we have

Proposition 3.8.3. *The map $T(x) = 2x \mod 1$ on $[0,1)$ is ergodic.*

For the irrational rotation T each of the functions χ_n is an eigenfunction. We define the class of dynamical systems with this property.

Definition 3.8.4. *A measure-preserving transformation T on a probability space $(X, \mathscr{B}, \mu)$ has discrete spectrum when $\mathbf{L}^2(X, \mathscr{B}, \lambda)$ has an orthonormal Hilbert basis consisting of eigenfunction.*

Note: Ergodic transformations with discrete spectrum can be modelled by systems that look like irrational rotations.

Theorem 3.8.5. *If a measure-preserving transformation T is ergodic and has discrete spectrum then it is isomorphic to a measure-preserving transformation S on a compact, Abelian group defined by $S(g) = g + a$ for some a in G with the property that $\{na : n \in \mathbb{Z}\}$ is dense.*

Could the doubling map $T(x) = 2x \mod 1$ be modelled by an irrational rotation? If it can, we need to define what it means for measure-preserving transformation to be the same.

Definition 3.8.6 (Isomorphism)**.** *Let $(X, \mathscr{B}, \mu, T)$ and $(Y, \mathscr{C}, v, S)$ be two measure-preserving transformations. We say that they are isomorphic if one can find $W \subset X$ and $Z \subset Y$ and a bijection $\psi : W \to Z$ such that*

- *$\mu(W) = 1$ and $v(Z) = 1$*
- *W is T invariant*

- Z *is* S *invariant*
- ψ *and* ψ^{-1} *are measurable*
- $S \circ \psi = \psi \circ T$
- $\mu(\psi^{-1}(E)) = v(E)$ *for all* $e \in \mathscr{C}$

all hold.

However, $T(x) = x + \alpha \mod 1$ and $S(x) = 2x \mod 1$ are not isomorphic. We can explain this by the Koopman operator. We know T has lots of eigenfunction. If S has very few eigenfunction, then T and S can not be isomorphic. This is because, were they isomorphic by ψ then $\psi^{-1} \circ \chi_k \circ \psi$ would be an eigenfunction of S with eigenvalue $e^{2\pi ik\alpha}$.

$$\begin{aligned} S(\psi^{-1} \circ \chi_k \circ \psi) &= \psi^{-1} \circ \chi_k \circ \psi \circ S \\ &= \psi^{-1} \circ (\chi_k \circ T) \circ \psi \\ &= \psi^{-1} \circ (e^{2\pi ik\alpha}\chi_k) \circ \psi \\ &= e^{2\pi ik\alpha}(\psi^{-1} \circ \chi_k \circ \psi) \end{aligned}$$

Lemma 3.8.7. *If* $(X, \mathscr{B}, \mu, T)$ *is mixing then there are no non-constant eigenfunctions.*

Proof. Suppose f is an eigenfunction of T, we should have $Tf = \eta f$ for some $\eta \in \mathbb{C}$ with $|\eta| = 1$. Then we have:

$$\langle f, \eta^n f\rangle = \langle f, T^n f\rangle.$$

For the left hand side, we have $\langle f, \eta^n f\rangle = (\eta^n)\langle f, f\rangle = (\eta^n)\|f\|^2$. For the right hand side, we have $\langle f, T^n f\rangle = \langle f, 1\rangle\langle 1, f\rangle = 0$. Therefore, we always have f are constant eigenfunctions. $\square$

3.9 Entropy

In the previous section, we have studied that the irrational rotation $x \to x + \alpha \mod 1$ and the doubling map $x \to 2x \mod 1$ are essentially different dynamical systems because the former one has lots of eigenfunctions while the latter has few. In this section, we will keep studying the topic of distinguishing dynamical systems by analyzing the shift map

$$(Tx)(n) = x(n+1)$$

on $X = \{0,1\}^{\mathbb{N}}$ wrt different coin measures ξ_p and ξ_q. Since both of them are mixing, both of them have no non-constant eigenfunctions. Therefore, we cannot distinguish them by properties of their Koopman operators. In this case, we will study the concept of entropy to distinguish them.

Definition 3.9.1 (Partition)**.** *Fix a measure space $(X, \mathscr{B}, \mu)$ with $\mu(X) = 1$. The partition of $(X, \mathscr{B}, \mu)$ is a finite tuple*

$$\eta = (A_1, \cdots, A_r)$$

of sets in $\mathscr{B}$ that are pairwise disjoint and cover X.

We can think the partition as classifying the points of X according to some rule, or as describing the distinct outcomes of an experiment that detects which member of the partition the outcome belongs.

For instance, the tuple

$$\eta = ([00], [01], [10], [11])$$

is a partition of $\{0,1\}^{\mathbb{N}}$ classifying points according to their first two terms. Also

$$\eta = ([0, \frac{1}{3}), [\frac{1}{3}, \frac{2}{3}), [\frac{2}{3}, 1))$$

is a partition of $[0,1)$ classifying points based on the first digits in their ternary expansions.

We can use the information function to describe 'how much' information we have gained from the experiment corresponding to the partition η.

Definition 3.9.2 (Information Function)**.** *Fix a measure space $(X, \mathscr{B}, \mu)$ with $\mu(X) = 1$. Given a partition $\eta = (A_1, \cdots, A_r)$ the function*

$$I(\eta) = -\sum_{j=1}^{r} 1_{A(j)} \log \mu(A(j))$$

is the information function of η. To make this function continues, we take $0 \log 0 = 0$.

Note: The information function is a simple function. One can think of $-\log \mu(A(j))$ as how surprised we are to learn that a point x belongs to $A(j)$. From the point of view of information theory it encodes how much storage is needed for the information obtained about a point x if it is found to lie in $A(j)$.

For instance, if $\eta = (X, \varnothing)$ then $I(\eta) = 0$. It means that we can not learn anything by determining a point x belongs to X or $\varnothing$. Also, if $\eta = (A, X \setminus A)$ then

$$I(\eta) = -1_A \log(\mu(A)) - 1_{X \setminus A} \log(1 - \mu(A)).$$

When A is very small, we are very surprised to learn that a point belongs to it.

The reason why we have such definition for the information function is that we also want to consider the appearance of independence.

Definition 3.9.3 (Independence)**.** *Fix a measure space $(X, \mathscr{B}, \mu)$ with $\mu(X) = 1$. Partitions $\eta = (A_1, \cdots, A_r)$ and $\theta = (B_1, \cdots, B_s)$ are independent if $\mu(A_i \cap B_j) = \mu(A_i)\mu(B_j)$ for all $1 \leq i \leq r$ and all $1 \leq j \leq s$*

Independent partitions do not affect each other experimentally. For example, when repeatedly tossing a coin the result of the previous toss does not influence the next one. The partitions $([0], [1])$ and $(T^{-1}[0], T^{-1}[1])$ are independent. We should expect that the information gained by performing both experiments is the sum of the information gained by performing the experiments consecutively. It is the join of two partitions that corresponds to performing both experiments at once.

Definition 3.9.4 (Join)**.** *Given partitions η and θ the partition*

$$\eta \vee \theta = (A \cap B : A \in \eta, B \in \theta)$$

is their join.

Lemma 3.9.5. *If η and θ are independent partitions then $I(\eta \vee \theta) = I(\eta) + I(\theta)$.*

Proof. It can be proved by applying the formula of information function. □

The integral of the information function is the entropy of the partition.

Definition 3.9.6 (Entropy)**.** *Given a partition*

$$\eta = (A_1, \cdots, A_r)$$

of a probability space $(X, \mathscr{B}, \mu)$ *the quantity*

$$H(\eta) = -\sum_{i=1}^{r} \mu(A_i) \log \mu(A_i)$$

is its entropy.

We imagine that the entropy of a partition represents the average or expected surprise from conducting the η experiment. For instance

$$\eta = ([00], [01], [10], [11])$$

has, for the fair coin measure $\xi_{1/2}$ an entropy is

$$\begin{aligned} H(\eta) = &- \xi_{1/2}([00]) \log \xi_{1/2}([00]) - \xi_{1/2}([01]) \log \xi_{1/2}([01]) \\ &- \xi_{1/2}([10]) \log \xi_{1/2}([10]) - \xi_{1/2}([11]) \log \xi_{1/2}([11]) \\ = &- 4(\frac{1}{4} \log(1/4)) = \log(4). \end{aligned}$$

Due to the convexity, for any other partition η of $\{0,1\}^{\mathbb{N}}$ into four sets we have

$$0 \leq H(\eta) \leq \log 4.$$

Since we have studied the idea of partition and entropy, we will combine them with the dynamics.

Based on the partition η, we can use T to formulate a new partition

$$(T^{-1}A_1, \cdots, T^{-1}A_r).$$

The new partition $T^{-1}\eta$ corresponds to performing the experiment represented by η after one iteration of dynamics. Since T is measure-preserving, the partitions η and $T^{-1}\eta$ have the same entropy. We are curious about how much information do we gain from performing the experiment corresponding to η now and after one iteration of the dynamics, which can be defined as performing the experiment corresponding to the join

$$\eta \vee T^{-1}(\eta) = (A_i \cap T^{-1} : 1 \leq i, j \leq r)$$

of the two partitions. The entropy gained can be calculated by

$$H(\eta \vee T^{-1}\eta) - H(\eta).$$

For example, we have a partition $\eta = ([00], [10], [01], [11])$, then

$$\eta \vee T^{-1}\eta([000], [100], [010], [110], [001], [101], [011], [111])$$

and its entropy for the fair coin measure is $\log 8$. Note that

$$H(\eta \vee T^{-1}\eta) - H(\eta) = \log 8 - \log 4 = \log 2.$$

This difference as representing the extra bit needed to store the outcome of the experiment $\eta \vee T^{-1}\eta$ compared with the outcome of experiment η.

Definition 3.9.7 (Entropy of T for η)**.** *Fix a measure-preserving transformation T on a probability space $(X, \mathscr{B}, \mu)$. The quantity*

$$H(T, \eta) = \lim_{N\to\infty} \frac{1}{N} H\left(\bigvee_{n=0}^{N-1} T^{-n}\eta\right)$$

is the entropy of T for the partition η.

The entropy of T for the partition η is the exponential growth rate of the amount of information obtained by repeatedly performing the experiment corresponding to T after more and more iterations of the dynamics. The positivity of $H(T, \eta)$ tells us something about the randomness of T, and there is still information to be gained by performing again.

For example. we can calculate the entropy of the full shift T on $\{0, 1\}^{\mathbb{N}}$ for the partition $\eta = \{[0], [1]\}$ and the fair coin measure $\xi_{1/2}$. Since

$$\eta \vee T^{-1}\eta \vee \cdots \vee T^{-(N-1)}\eta$$

is equal to the partition of $\{0, 1\}^{\mathbb{N}}$ into cylinder sets of length N and each such cylinder has measure $1/2^N$ we see that

$$H\left(\bigvee_{n=0}^{N-1} T^{-n}\eta\right) = \log(2^N)$$

and therefore

$$H(T, \eta) = \lim_{N\to\infty} \frac{1}{N} \log(2^N) = \log 2 \tag{17}$$

is the entropy of T wrt η.

However the entropies may be different wrt different partitions of X. Therefore we have a more formalized definition for the entropy of a measure-preserving map T.

Definition 3.9.8 (Entropy of a Measure-preserving Map T)**.** *The entropy of a measure-preserving map T on a probability space $(X, \mathscr{B}, \mu)$ is the supermum*

$$H(T) = \sup\{H(T, \eta) : \eta \text{ a partition of } (X, \mathscr{B}) \text{ with finite entropy}\} \tag{18}$$

of all possible entropies of T with respect to partitions that themselves have finite entropy.

Then we can go back to our original aim.

Theorem 3.9.9. *If If $(X, \mathscr{B}, \mu, T)$ and $(Y, \mathbb{C}, v, S)$ are isomorphic then they have the same entropy.*

Here we have an important theorem to compute the entropy of a measure-preserving transformations.

Theorem 3.9.10 (Kolmogorov-Sinai)**.** *If η is a partition such that*

$$\sigma\left(\bigvee_{n=0}^{\infty} T^{-n}\eta\right) = \mathscr{B}$$

then $H(T) = H(T, \eta)$.

Since the cylinder sets generate the Borel $\sigma-$ algebra on $\{0,1\}^{\mathbb{N}}$ we can calculate the entropy of a measure-preserving transformation T on $\{0,1\}^{\mathbb{N}}$ using the partition $\{[0],[1]\}$. If μ_p is the coin measure, we have

$$H(\eta) = -(1-p)\log(1-p) - p\log(p).$$

Therefore if $p+q=1$, T with μ_p and T with μ_q isomorphic measure-preserving transformations.

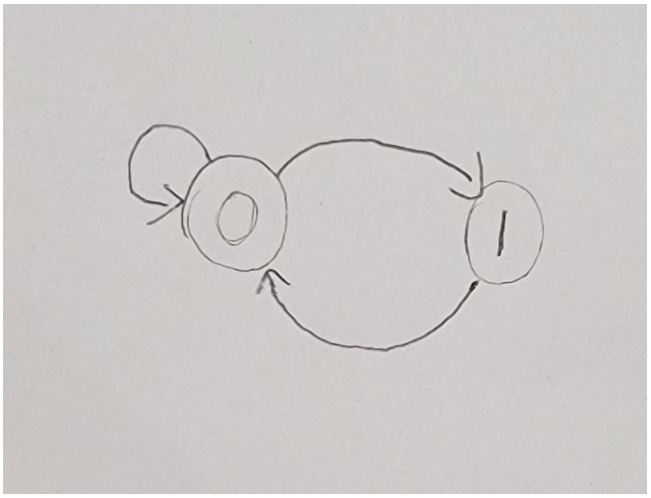

Figure 3: This is the visualization of the set Y (19). It contains two vertices 0 and 1, and three edges that are $0 \to 0$, $0 \to 1$, and $1 \to 0$.

3.10 Markov Chain

In this section, we will study the Markov chain under the view of ergodic theory.

Consider the set

$$Y = \{x \in \{0,1\}^{\mathbb{N}} : x(n) \Rightarrow x(n+1) = 0\}, \tag{19}$$

which is certainly invariant for the shift map: if $y \in Y$ then $T(y)$ belongs to Y as well. To study the dynamics of T on Y using ergodic theory we want a measure on Y that is T invariant. How can we equip Y with a probability measure μ that is T invariant?

We can visualize Y as Figure 3.To any endless journey on the graph one associates a sequence in Y by recording the labels associated with the visited vertices. As the vertex labeled 1 cannot be visited consecutively, and as there are not other restrictions, we get all sequences in Y this way.

We can assign probabilities to each traversal, and they can be encoded in a matrix:

$$\begin{bmatrix} q(0,0) & q(0,1) \\ q(1,0) & q(1,1) \end{bmatrix}$$

For each column of this matrix we have

$$q(0,0) + q(0,1) = 1$$
$$q(1,0) + q(1,1) = 1$$

wherer $q(i,j)$ refers to the probability for moving i to j in one step. In our case $q(1,1) = 1$. To entirely determine the measure we also need the probability of the starting location, which

can be written as

$$p(0) + p(1) = 1.$$

where $p(i)$ is the probability that one begins at vertex i. With the definition of $p(i)$ and $q(i,j)$, we can define

$$v([\epsilon_1 \cdots \epsilon_r]) = p(\epsilon_1) \prod_{i=1}^{r-1} q(\epsilon_i, \epsilon_{i+1})$$

on all cylinder sets $[\epsilon_1 \cdots \epsilon_r]$. For instance, we have

$$v([01]) = p(0)q(0,1).$$

Note that if μ is to be an invariant measure then we must have

$$p(i) = \mu([i]) = \mu(T^{-1}[i]) = p(0)q(0,i) + p(1)q(1,i)$$

which is to say

$$\begin{bmatrix} p(0) & p(1) \end{bmatrix} \begin{bmatrix} q(0,0) & q(0,1) \\ q(1,0) & q(1,1) \end{bmatrix} = \begin{bmatrix} p(0) & p(1) \end{bmatrix}$$

holds.

We will take this for granted to define a measure v on $\{0,1\}^{\mathbb{N}}$. We will verify it is T invariant. For $C = [\epsilon_1 \cdots \epsilon_n]$, we can calculate

$$\begin{aligned} v(T^{-1}C) &= v([0\epsilon_1 \cdots \epsilon_n]) + v([1\epsilon_1 \cdots \epsilon_n]) \\ &= \big(p(0)q(0,\epsilon_1) + p(1)q(1,\epsilon_1)\big) \prod_{i=1}^{r-1} q(\epsilon_i, \epsilon_{i+1}) \\ &= p(\epsilon_1) \prod_{i=1}^{r-1} q(\epsilon_i, \epsilon_{i+1}) = v(C) \end{aligned}$$

so v is T invariant.

Proposition 3.10.1. *For the above measure the quantity*

$$-p(0)q(0,0)\log q(0,0) - p(0)q(0,1)\log q(0,1)$$

is the entropy of T.

In more general case, we have

$$-p(0)q(0,0)\log q(0,0) - p(0)q(0,1)\log q(0,1) - p(1)q(1,0)\log q(1,0) - p(1)q(1,1)\log q(1,1)$$

is the entropy of T.

Proof. We can use Kolmogorov-Sinai Theorem 3.9.10 to compute it. □

We would like to maximize

$$-p(0)q(0,0)\log q(0,0) - p(0)q(0,1)\log q(0,1)$$

Because we have total freedom in the parameters $p(0)$ and $q(0,0)$. Fixing their values then determines $p(1)$ and $q(0,1)$ by the laws of total probability. What is the best way to choose their values? Absent any other information about the dynamics, or any other quantity that we might be interested in, it is often reasonable to choose the values that maximize the entropy. To do the optimaztion, we will use the Parry measure.

Theorem 3.10.2 (parry)**.** *Let B be a $k \times k$ matrix with entries from $\{0,1\}$. Suppose there is $r \in \mathbb{N}$ with all entries of B^r positive. Let λ be the largest positive eigenvalue of B. Fix left and right eigenvectors u and v of B respectively with*

$$u(1)v(1) + \cdots + u(k)v(k) = 1$$

such that both have an eigenvalue of λ. With

$$p(i) = u(i)v(i)$$

and

$$q(i,j) = \frac{B(i,j)}{\lambda}\frac{v(j)}{v(i)}$$

the corresponding Markov measure maximizes the entropy for T on the set

$$Y = \{x \in \{1,\cdots,k\}^{\mathbb{N}} : B(x(n),x(n+1)) = 1 \text{ for all } n \in \mathbb{N}\} \tag{20}$$

where transitions are determined by B.

In our case

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

with eigenvalues

$$\frac{1-\sqrt{5}}{2} \quad \frac{1+\sqrt{5}}{2}$$

and the latter must be λ. The vectors

$$U = [1+\sqrt{5} \quad 2]$$

$$V = \begin{bmatrix} 1+\sqrt{5} \\ 2 \end{bmatrix}$$

are left and right eigenvectors respectively of B with eigenvalue λ. Since the ratio of the eigenvectors is unchanged by scaling we conclude that

$$q(0,0) = \frac{B(0,0)}{\lambda}\frac{v(0)}{v(0)} = \frac{1}{\lambda}$$

and

$$q(0,1) = \frac{B(0,1)}{\lambda}\frac{v(1)}{v(0)} = \frac{1}{\lambda^2}$$

are the values that will maximize the entropy. As

$$p(0)q(0,1) = p(1) = 1 - p(0)$$

we conclude that

$$p(0) = \frac{\lambda^2}{1+\lambda^2}$$

and

$$p(1) = \frac{1}{1+\lambda^2}.$$

which gives the entropy value $H(T) = \log \lambda$.

References

[1] J.-F. Le Gall. *Measure theory, probability, and stochastic processes*. Springer, 2022.