A Friendly Introduction To Real Analysis I

(Chinese version first published on 2024.2.02)

O. Anqiao

Introduction

I plan to write a series of blogs on mathematical analysis next, and this will be the first installment focusing on real analysis.

My original intention in this series is to write articles suitable for beginners, so I will not pay too much attention to concepts that are too abstract or extremely difficult.After all, many people might associate terms like these with something very advanced, just like how we might have felt calculus was some kind of advanced stuff when we were in kindergarten. In reality, most people can naturally grasp these concepts after actually studying them. For beginners like myself, the biggest challenge is getting

> It's truly remarkable that humans can understand natural language. There's no need to feel confused or frustrated by the initial struggle with the dense formal language in textbooks. Everyone has gone through this phase at some point.

About Real Analysis

Real analysis is a broad concept that involves the study of analytic properties of real-valued functions and sequences (such as limits of sequences, differentiation and integration of functions, and various properties). Some sources equate real analysis with the study of "theory of real variable functions," while others may use different names for essentially the same content, such as "Theory of real variable functions" and "Real analysis." In some cases, courses titled "Real variable functions" may serve as prerequisites for courses in "*Real analysis*", indicating a foundational relationship between them. Additionally, in some graduate-level courses, the subject might be referred to simply as "*Analysis*". Thus, it seems that the name is not as important as the content, which is indeed the case.

（Feel free to skip over any mathematical terms that you find difficult to understand, or even proceed directly to the next chapter.）

（The following definition is limited to my articles on mathematical analysis. It is based on my personal understanding and serves only as a summary. I do not actively refute any other viewpoints or definitions）

> In Chinese, there are mainly two terms used to refer to areas primarily focused on real analysis: "实变函数论（theory of real variable functions）" and "实分析（real analysis）". This paragraph is mainly intended to clarify for Chinese readers.

In distinguishing them, theory of real variable functions belongs to real analysis and is one of its fundamental aspects. In discussions of theory of real variable functions, we cover some basic topics such as measure on Euclidean spaces, the construction of integrals, and various types of convergence. We primarily focus on functions defined on the real numbers. In a broader sense of real analysis, these topics include not only the aforementioned but also delve into more general concepts such as metric spaces and topological spaces. In theory of real variable functions, we mainly refer to Lebesgue measure spaces. Let $\mathcal{M}$ denote the sigma-algebra of Lebesgue measurable sets, m denote the Lebesgue measure, and $\mathbb{R}^n$ denote the real space. Thus, we use $(\mathbb{R}^n, \mathcal{M}, m)$ to represent this context. When discussing real analysis within this scope, I generally refer to it as "*theory of real variable functions*".

However, to extend to the category of real analysis, we will no longer be limited to *Lebesgue* measures. We Most of the theorems in the former *Lebesgue* measure space except *Lusin* will be generalized in general measures. Then we will discuss its integral and other properties based on this abstract measure space. We will further discuss some measures based on topological space, such as the Borel measure, and there are some things involving group theory like the Haar measure.

By studying real analysis, we delve deeper into some advanced mathematical concepts (here referring to their level of abstraction), such as integration over the real number field, linear transformations, convolution, H-L maximal functions, Fourier transforms, regular transformations, and so on. You might find that these "advanced concepts" seem somewhat familiar. Indeed, they are. However, from the perspective of a junior high school student following a typical curriculum, you wouldn't engage in rigorous mathematical proofs or formal derivations when performing a linear transformation on a vector/matrix, for example.

What we aim to do is to construct a solid framework of real analysis from the basics, providing a more rigorous and profound mathematical foundation for these concepts.

Real Numbers

Axiomatic System of Real Numbers

To ensure the rigor, consistency, and accuracy of the mathematical system, it's necessary to establish a set of basic properties and axioms defining real numbers at the outset of the article. This lays the groundwork for subsequent systematic and rigorous derivations and proofs.

$\mathbb{R}$ represents the set of real numbers. We use $\mathbb{R} \times \mathbb{R}$ to denote the Cartesian product of the real number set $\mathbb{R}$, which consists of all ordered pairs in the form (x, y) where x and y are real numbers. Building upon this, we define two operations as follows, considering elements x and y within the set:

- $+$ (Plus): $\mathbb{R} \times \mathbb{R} \to \mathbb{R}, (x, y) \mapsto x + y$
- $\cdot$ (Multiply): $\mathbb{R} \times \mathbb{R} \to \mathbb{R}, (x, y) \mapsto x \cdot y$

Therefore, in the field $(\mathbb{R}, +, \cdot)$, we consider for all $\forall x, y, z \in \mathbb{R}$:

- Addition/Multiplication Associativity: $x + (y + z) = (x + y) + z, x \cdot (y \cdot z) = (x \cdot y) \cdot z$
- Addition/Multiplication Commutativity: $x + y = y + x, x \cdot y = y \cdot x$
- Multiplication Distributivity: $x \cdot (y + z) = (x \cdot y) + (x \cdot z)$
- Additive Identity: $x + 0 = x$
- Multiplicative Identity: $x \cdot 1 = x$

- Additive Inverse: $\exists -x \in \mathbb{R}$ such that $x + (-x) = 0$
- Multiplicative Inverse: $x \neq 0, \exists x^{-1} \in \mathbb{R}$ such that $x \cdot x^{-1} = 1$

We have defined the basic properties of addition and multiplication operations on the set of real numbers. This set of axioms is commonly referred to as the "field axioms." However, this alone is not sufficient. We know that the set of real numbers forms an ordered field, so we also need to define a linear order ≤ that satisfies the following properties for all $\forall x, y, z \in \mathbb{R}$:

- Antisymmetric relation: If $x \leq y$ and $y \leq x$, then $x = y$
- Transitive relation: If $x \leq y$ and $y \leq z$, then $x \leq z$
- Completeness: There are and only cases where $x \leq y$ and $y \leq x$
- Reflexive relation: $x \leq x$

However, even with this definition, what we have defined is only the rational numbers $\mathbb{Q}$, which are numbers that can be expressed as the ratio of two integers. This definition does not include irrational numbers such as $\sqrt{2}$, and from the perspective of the number line, there are gaps in it. Therefore, we need further definitions to encompass the entirety of real numbers.

Dedekind Cut

To ensure the completeness of real numbers, one well-known principle is the Dedekind Cut introduced by Richard Dedekind. The idea is to eliminate any "gaps" in the real number line by associating each cut with a point, providing an intuitive understanding that facilitates its definition.

For any non-empty set S of real numbers that is bounded above, there exists a real number α such that for all $x \in S$, we have $x \leq \alpha$. This α is called the supremum of S, also known as the least upper bound.

Dividing $\mathbb{R}$ arbitrarily into two non-empty, disjoint subsets A and B, representing the left and right subsets respectively, such that any real number α can be either the supremum of the left subset A or the infimum of the right subset B. In other words, for any $A|B$ cut in the real number field, there always exists a real number α corresponding to it. This ensures that each Dedekind cut corresponds to a real number, leaving no undefined regions on the number line.

Proof

We prove the existence of a Dedekind cut by considering three cases. Let's assume that A has an upper bound α, and B has a lower bound β.

The first case arises when A has a maximum value. If $\alpha \in A$, then for all $x \in A$, $x \leq \alpha$. Moreover, for all $x \in A$ and $y \in B$, we have $x \leq \alpha < y$. Therefore, we can choose the cut point to be α.

Similarly, the second case occurs when B has a minimum value. If $\beta \in B$, then for all $x \in A$ and $y \in B$, $x < \beta \leq y$. In this scenario, we choose the cut point to be β.

The remaining case is as follows: if $\alpha \notin A$ and $\beta \notin B$, then $\beta \in A$ and $\alpha \in B$. Under such conditions, we would have $\beta < \alpha$. However, from another perspective, since α is the least upper bound of A and $\beta < \alpha$, there should be at least one $x \in A$ such that $\beta < x$. At the same time, since β is the greatest lower bound of B, there exists $y \in B$ such that $\beta \leq y < x$. This contradicts the conditions that the set of real numbers is supposed to satisfy, as stated at the outset.

$$\forall x \in A, y \in B, x \leq \alpha < y \text{ or } x < \alpha \leq y$$

Hence, there does not exist a number that belongs to neither set A nor set B in the set of real numbers.

As a metric space, we can establish the completeness of the real number field through methods like Dedekind cuts. Similarly, we can use these methods to prove the completeness of other metric spaces.

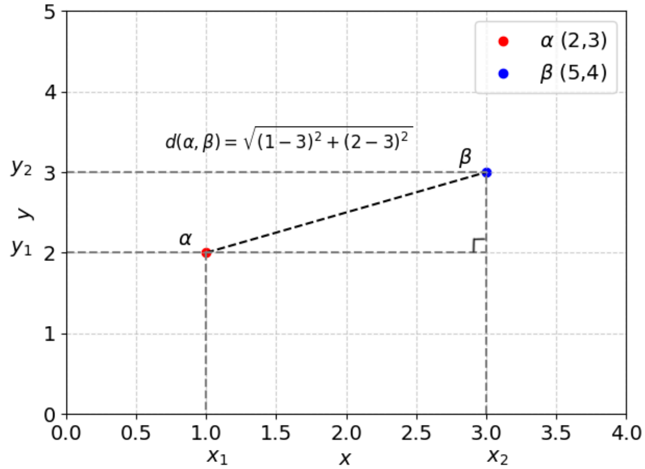
Although this may slightly exceed the scope and temporarily shift focus away from Real Analysis towards topics more centered on Functional Analysis, I believe it's valuable for readers to grasp these concepts.

Metric space

"Measurement" and "Metric space" are essentially mathematical concepts that we encounter from elementary school when we start learning basic plane geometry. However, at that stage, children often understand these concepts in an intuitive but not necessarily rigorous way. Teachers in elementary school tell us that in Euclidean space, the shortest distance between two points is a straight line. Then, we can use a ruler to measure the distance between two points. When we choose such a measuring tool for distance, the entire space becomes a metric space, where we can study various distance relationships.

Therefore, what we need to pay attention to is the essence of "measurement". A metric space consists of a non-empty set X, which we call the point set, equipped with a metric function $d: X \times X \to \mathbb{R}$, satisfying:

- Non-negativity: $\forall x, y \in X, d(x, y) \geq 0$, and equality holds if and only if $x = y$.
- Symmetry: $\forall x, y \in X, d(x, y) = d(y, x)$.
- Triangle inequality: $\forall x, y, z \in X, d(x, z) \leq d(x, y) + d(y, z)$.



Such a metric function measures the distance between two points in the space, allowing us to discuss concepts like closeness and convergence of points in the space. Then, we can represent this metric space with an ordered pair (X, d). Taking Euclidean space as an example, in Euclidean space, the set is typically the collection of n-dimensional real number vectors, represented as $\mathbb{R}^n$, so each point can be represented as $(x_1, x_2, \cdots, x_n)$.

So, when we have two points, $\alpha = (x_1, x_2, \cdots, x_n)$ and $\beta = (y_1, y_2, \cdots, y_n)$, the distance function d in Euclidean space is defined as:

$$d(\alpha, \beta) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Limits and Convergence

We say that a sequence converges if the terms of the sequence get closer and closer to a specific value.

Rigorously speaking, a sequence can be seen as a mapping from a set of natural numbers (or its subset) to the set of real numbers. When we represent a ordered set of numbers $a_1, a_2, \cdots, a_n$ with $\{a_n\}$ $(n \geq 1)$ and refer to it as a sequence, its essence is a mapping:

$$f : \mathbb{N} \to \mathbb{R}, \; n \mapsto f(n) = a_n$$

When we talk about the limit of a sequence, we are concerned with the specific value that the numbers in the sequence tend towards as the number of terms increases. However, this intuitive explanation clearly cannot rigorously define a limit. Yet, the foundation of calculus is built upon this concept. If conclusions drawn from here are unreliable due to lack of rigor, calculus would become merely a superficial transformation based on intuition. Therefore, early mathematicians began to seek something that could rigorously formalize the concept of a limit.

Consider a positive real number ϵ. For a sequence $\{a_n\}$, if there exists $L \in \mathbb{R}$ such that for any positive real number ϵ, there exists a positive integer N such that for all $n > N$, $|a_n - L| < \epsilon$, then the sequence $\{a_n\}$ is said to converge to L. When the limit exists, we say that the sequence converges, denoted as $\lim_{n\to\infty} a_n = L$.

Here, ϵ represents the precision required to approach the limit. We hope that the distance between each term in the sequence and the limit is less than this precision. There always exists an N such that starting from the Nth term of the sequence, the error between each term of the sequence and the limit does not exceed ϵ, indicating that the term is approaching the limit with the required precision. Intuitively, we usually assume ϵ to be a small number and N to be a large number. This rigorous definition of a limit is called the "$\epsilon - N$ language". We can represent the convergence of a sequence $\{a_n\}$ to a limit L as n tends to infinity:

$$\forall \epsilon > 0 \ \exists N_\epsilon \in \mathbb{N} \ \forall n \in \mathbb{N} \ \left[n > N_\epsilon \implies |a_n - L| < \epsilon \right]$$

We often hear about Cauchy sequences, which are sequences satisfying the Cauchy condition. In a metric space, consider a positive real number ϵ. A sequence $\{a_n\}$ is called a Cauchy sequence if for any $\epsilon > 0$, there exists an integer N such that for all $m, n > N$, $|a_m - a_n| < \epsilon$. In the real number space, any Cauchy sequence always converges within the real number field . Therefore, we call the real number space a complete metric space. From the perspective of the Cauchy condition, completeness of a metric space means that any Cauchy sequence in that space converges within the space itself. Any closed and bounded subset in Euclidean space $\mathbb{R}^n$ is called a compact space. Any compact space is complete.

For convenience, let's define a function here.

$$f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}, \ n \mapsto d(n, m) = |a_n - a_m|$$

Property: Cauchy sequences on $\mathbb{R}$ are all bounded.

Proof

According to the definition of Cauchy sequences, considering a Cauchy sequence a_n, and a positive real number ϵ, there exists a corresponding positive integer N such that for all $m, n > N$, we have $d(m, n) < \epsilon$. Thus, we can choose

$$r = \max\{\epsilon, d(a_1, a_{N+1}), d(a_2, a_{N+1}), \cdots, d(a_N, a_{N+1})\}$$

Since we have taken r as the maximum of all possible distances, for all $n > N$, we have $d(a_n, a_{N+1}) < \epsilon$. For all $n \leq i \leq N$, we have $d(a_i, a_{N+1}) \leq r$.

Considering a_{N+1} as the center and r as the radius, the entire Cauchy sequence is contained within a finite-radius open ball. Therefore:

$$B_r(a_{N+1}) = \{x \in \mathbb{R} | d(x, a_{N+1}) < r\}$$

This open ball includes all real numbers that are within a distance less than r from a_{N+1}, forming the bounded range of the Cauchy sequence.

Additionally, as shown above, we will typically specifically define the distance between two elements, because using the absolute value $|a_n - a_m|$ directly might hinder us from generalizing the properties of Cauchy sequences to more general metric spaces.

Common Algebraic structures

An algebraic structure is a mathematical object that consists of a set and some operations defined on it, which satisfy certain properties. This is a fundamental concept in universal algebra and will be crucial in many discussions later on. Therefore, it is important to introduce this concept early on to provide readers with at least a basic understanding of some simpler algebraic structures.

Group

A group is an algebraic structure consisting of a set equipped with a binary operation that satisfies the associativity, identity element, and inverse element properties. It is a simple algebraic structure. Consider a set G, with $\cdot$ representing its binary operation. The combination of two elements x and y to form an element belonging to G is denoted as $x \cdot y$. For all $x, y, z \in G$, if the following properties hold:

- Closure: $a \cdot b \in G$
- Associativity: $(x \cdot y) \cdot z = x \cdot (y \cdot z)$
- Identity element: There exists e in G such that $xe = ex = x$ holds.
- Inverse element: Each element x has an inverse $x^{-1} \in G$, satisfying $xx^{-1} = x^{-1}x = e$.

Then we say $(G, \cdot)$ is a group. This is straightforward, but there are some important conclusions we should know. Firstly, the identity element is unique. Secondly, the inverse of each element is unique, and for each element, its inverse is also the inverse of its inverse.

Additionally, there are some structures similar to groups in different scenarios, which may also be special groups. It's worth noting them for possible future reference. If this set R satisfies the closure property and associativity as described above, it forms a semigroup. If a semigroup satisfies the identity element property, it becomes a monoid. If a group satisfies the commutative law, i.e., $\forall x, y \in G,\ x \cdot y = y \cdot x$, it is called an Abelian group or a commutative group. A group with only one element is called a trivial group, and every group contains a trivial group.

Ring

A ring is a more complex algebraic structure than a group. It consists of a set equipped with two binary operations, denoted by $+$ and $\cdot$, which we call addition and multiplication (though these operations differ from the usual arithmetic addition and multiplication). Consider a set R. The pair $(R, +)$ forms an abelian group, with the identity element denoted as 0, and the inverse element of an element x represented as $-x$. This is called the additive group. Moreover:

- $(R, \cdot)$ forms a semigroup.
- Multiplication distributes over addition, meaning for all $x, y, z \in R$, the following equalities hold:
- $x \cdot (y + z) = (x \cdot y) + (x \cdot z)$
- $(x + y) \cdot z = (x \cdot z) + (y \cdot z)$

Then we call $(R, +, \cdot)$ a ring.

If multiplication on R is commutative, then it is called a commutative ring. If there exists an identity element for multiplication, denoted as 1, then it is called a unital ring. If every nonzero element x in a unital ring R has a multiplicative inverse x^{-1}, then R is a division ring.

When $R = \{0\}$, where 0 is the identity element for both addition and multiplication, it is called the zero ring. Additionally, any set starting from integers $\mathbb{Z}$ and onwards, with the usual addition and multiplication operations, forms a commutative ring.

Field

The structure of a field is similar to that of a ring, consisting of a set and two binary operations. However, its properties are stronger, satisfying both the commutative ring and division ring properties, in other words, an extension of the four basic arithmetic operations of addition, subtraction, multiplication, and division.

> More literature often uses K to denote the set as a field, derived from the German term "Körper" for field, perhaps due to historical reasons or to avoid conflicts with the word "Field". I'll use it this way, without delving too much into the reasons.

Consider a set K equipped with binary operations $+$ and $\cdot$, satisfying:

- $(K, +)$ is an abelian group with 0 as its identity element.
- $(K \setminus \{0\}, \cdot)$ is an abelian group.
- Multiplication $\cdot$ distributes over addition $+$.

Then we call $(K, +, \cdot)$ a field. When we talk about the field of rational numbers, real numbers, or complex numbers, it becomes clear (and I won't mix up these terms later).

Basics of Measure theory

Introduction

When it comes to measuring the size of sets, readers with mathematics learning progress at the level of junior high school or below may first think of the number of elements in a set, known as cardinality. This concept is simple for countable sets, but it fails in many cases. For example, we know that any interval on the real number line can contain infinitely many points. Additionally, there are issues related to dimensionality. Therefore, we need to consider a more rigorous and advanced method of description.

Typically, our operations such as integration are performed on intervals, which are generally represented as sets. To extend these operations to more general sets, the concept of measure is developed. Intuitively, measure is a mathematical concept used to measure the size or length of a set, often represented by a non-negative real number. Thus, measure theory can be regarded as part of real analysis. From the perspective of real analysis, measure theory provides a more general way to handle the size of sets, introducing the concept of abstract measure spaces.

Measure space and σ-*algebra*

Before discussing measures and metric spaces, we should first have a basic understanding of sigma algebras. Essentially, a sigma algebra is a family of sets that satisfies certain properties. We use $\mathcal{P}(X)$ to denote the set of all subsets of a set X, which is called the power set of X. Another notation for this is 2^X. Formally, it is defined as:

$$\mathcal{P}(X) = 2^X = \{A | A \subseteq X\}$$

Given a non-empty set X, let $\mathcal{F} \subseteq \mathcal{P}(X)$ be a subset of the power set of X. Here, $\mathcal{F}$ is a collection of subsets satisfying:

- $X \in \mathcal{F}$
- If $A \in \mathcal{F}$, then $X \backslash A \in \mathcal{F}$, meaning that for every set in $\mathcal{F}$, its complement must also be in $\mathcal{F}$.
- If countably many sets $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$, indicating that $\mathcal{F}$ is closed under countable unions of sets.

Then we call $\mathcal{F}$ a sigma algebra. Additionally, a sigma algebra includes the empty set $\varnothing$ because $X \in \mathcal{F}$, and when $A = X, X \backslash A = X \backslash X = \varnothing \in \mathcal{F}$. Therefore, the smallest sigma algebra on the set X is $\{\varnothing, X\}$, and the most general sigma algebra is $\mathcal{P}(X)$.

By introducing a sigma-algebra, we can define a set space containing the sigma-algebra as a measurable space, denoted as $(X, \mathcal{F})$. For all $A \in \mathcal{F}$, we call A a measurable set. However, this is not yet a measure space. Here, we need to distinguish that a measure space also requires a measure function, defined as:

$$\mu : \mathcal{F} \to [0, +\infty]$$

This function maps the sigma algebra $\mathcal{F}$ to non-negative real numbers to satisfy the properties of a measure space. Thus, a measure space can be represented as $(X, \mathcal{F}, \mu)$. A measure space has the following two important properties:

- Null set measure is zero: $\mu(\varnothing) = 0$.
- It is countably additive; for disjoint countable sets $A_1, A_2, \cdots \in \mathcal{F}$, we have $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

In general measure spaces, subsets of null sets may not necessarily be measurable, which often poses a significant challenge. Therefore, based on this, if a measure space $(X, \mathcal{F}, \mu)$

satisfies that any subset of a null set measured by μ remains an element of $\mathcal{F}$, then we call it a complete measure space.

Borel Set

We haven't discussed the concept of "topological space" yet because introducing it here may confuse readers. Therefore, a brief mention of the general definition of Borel sets is sufficient (we will revisit the definition of Borel sets in $\mathbb{R}^n$ later).

Consider a topological space (X, τ). The smallest σ-algebra generated by the topology τ is called the Borel σ-algebra of X, denoted by $\mathcal{B}(X)$. The elements of the Borel σ-algebra are called Borel sets on X. Borel sets are closed under complementation and countable unions, meaning:

- Closure under complementation: $E \in \mathcal{B}(X) \Longleftrightarrow E^c \in \mathcal{B}(X)$
- Closure under countable unions: If there exist countable sequences $A_1, A_2, \cdots \in \mathcal{B}(X)$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}(X)$.

The fundamental elements of the topological structure are open sets. Borel sets form the smallest σ-algebra containing all open sets. Measures defined on Borel sets are called Borel measures. It's worth noting that Borel sets are typically not complete. In a general metric space, the Borel σ-algebra may not include all subsets of zero measure.

External measures

There exist sets that are immeasurable, meaning they do not fit into the standard notion of measure. For such non-measurable sets, the introduction of outer measure allows us to study measure theory on a broader class of sets. Through appropriate constructions, outer measures can be extended to measurable measures, providing a better description of the properties of sets. This is an important concept.

Consider a set X and a family of sets $\mathcal{F} = 2^X$, along with a function μ^* mapping to

$[0, \infty]$:

$$\mu^* : \mathcal{F} \to [0, \infty]$$

Similarly, the outer measure of the empty set is zero, i.e., $\mu^*(\varnothing) = 0$, and it satisfies the following properties:

- Monotonicity: If $A \subseteq B$, then $\mu^*(A) \leq \mu^*(B)$.
- Countable subadditivity: For any sequence of subsets A_i of X, satisfying $\mu^*(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$

We then call μ^* an outer measure defined on the family of sets $\mathcal{F}$.
From these properties, we can also conclude that the outer measure is non-negative, i.e., for all $E \in \mathcal{F}$, $\mu^*(E) \geq 0$.

When constructing the outer measure, we need to consider how to take the infimum of all possible estimates among all possible coverings to ensure that the outer measure is, in a broad sense, the "smallest measure."

Consider the outer measure μ^* of a measurable set X. Let $\mathcal{E}$ be the subset family of X containing the empty set, representing the infimum of measures of all possible coverings $A_1, A_2, \cdots$, and let $\mu(A_k)$ denote the measure of A_k. Then, for any $E \subseteq X$, its outer measure is defined as follows:

$$\mu^*(E) = \inf \left\{ \sum_{k=1}^{\infty} \mu(A_k) | E \subseteq \bigcup_{k=1}^{\infty} A_k, A_k \in \mathcal{E} \right\}$$

A more intuitive way to understand this is to think of the "size" of a set in terms of "range". Imagine building a larger range from several smaller ranges (the "external" part for the set). This larger range covers all possible locations the set might occupy. Then, shrink it down (take the infimum) to the smallest range that still covers the entire set.

In this case, according to the definition of infimum, given $\epsilon > 0$, countable sets A_k satisfy $E \subseteq \bigcup_{k=1}^{\infty} A_k$, where:

$$\mu^*(E) \leq \sum_{k=1}^{\infty} \mu(A_k) \leq \mu^*(E) + \epsilon$$

Carathéodory's Extension Theorem

In measure theory, there is a famous theorem called the Carathéodory extension theorem, which describes how to extend an outer measure to a complete measure space.

Carathéodory's extension theorem can be stated in various ways, but here is a more straightforward one: We can construct a complete measure space based on the original measurable set family, such that the definition of the measure holds for a more general set system while preserving the properties of outer measure.

Consider a measurable set X and its outer measure μ^*. If for all $A \subseteq X$, we have:

$$\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^c)$$

Then we call E a measurable set, where E^c is the complement of E, and in this case, $\mu^*(E) = \mu(E)$. Then we consider a collection $\mathcal{M}$ of measurable sets in X to form a sigma algebra:

$$\mathcal{M} = \{E \subseteq X | \forall A \subseteq X, \mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^c)\}$$

This set $\mathcal{M}$ already satisfies the properties of a sigma algebra.

Proof

What we need to prove?
Given a non-empty set X, a collection $\mathcal{M}$ satisfying the following conditions is called a sigma algebra on the set X:

1. $X \in \mathcal{M}$.
2. If $A \in \mathcal{M}$, then $X \backslash A \in \mathcal{M}$, which is equivalent to $A^c \in \mathcal{M}$.

3. If there exists a countable sequence $A_1, A_2, \cdots \in \mathcal{M}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{M}$.

(1) X itself is measurable, which obviously implies that $X \in \mathcal{M}$, satisfying $X \in \mathcal{M}$.

(2) When $E \in \mathcal{M}$, E is measurable. According to the properties of measurable sets, we know that the complement of E, denoted as E^c, is also measurable. Therefore, it is evident that:

$$E \in \mathcal{M} \Longleftrightarrow E^c \in \mathcal{M}$$

(3) When countable sets $A_1, A_2, \cdots \in \mathcal{M}$, where each A_i is measurable, this implies that it holds for an infinite count as well. Let's first select two sets $M, N \in \mathcal{M}$ to prove that $\mathcal{M}$ **is closed under countable finite unions**. Although not comprehensive, this is relatively simple and serves as an initial step. Consider $E \subseteq X$, we need to prove:

$$\mu^*(E) = \mu^*(E \cap (M \cup N)) + \mu^*(E \cap (M \cup N)^c)$$

Since both M and N are measurable, we can derive:

$$\begin{aligned}
&\mu^*(E \cap (M \cup N)) + \mu^*(E \cap (M \cup N)^c) \\
&= \mu^*(E \cap M) + \mu^*(E \cap N \cap M^c) + \mu^*(E \cap (M \cup N)^c) \\
&= \mu^*(E \cap M) + \mu^*(E \cap N \cap M^c) + \mu^*(E \cap M^c \cap N^c) \\
&= \mu^*(E \cap M) + \mu^*(E \cap M^c) \\
&= \mu^*(E)
\end{aligned}$$

Now we have proven the closure of $\mathcal{M}$ under the union of finite sets, but that's not enough. We need to prove it for countable cases as well.
Consider a positive integer n and countably many measurable sets $A_1, A_2, \cdots \in \mathcal{M}$. Define the set $E_n = \bigcup_{i=1}^{n} A_i$. E_n itself is increasing ($E_1 \subseteq E_2 \subseteq \cdots$). Due to the subadditivity of outer measure, we can conclude:

$$\mu^*(E_n) = \mu^*\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mu^*(A_i)$$

We can consider the limit case:

$$E = \lim_{n \to \infty} E_n = \bigcup_{i=1}^{\infty} A_i$$

$$\mu^*(E) = \mu^* \left(\bigcup_{i=1}^{\infty} A_i \right) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$$

By taking n to infinity, we obtain $\mu^*(E) \leq \sum_{i=1}^{\infty} \mu^*(A_i)$. This process is similar to using a limiting process to prove properties of measurable sets, extending from the case of finitely many sets to countably many sets. Consequently, we can conclude that $\mathcal{M}$ is closed under countable unions, and thus $\mathcal{M}$ is a sigma algebra.

Now, we have an outer measure μ^*. We hope to find a measure function μ such that for any measurable set $E \in \mathcal{M}$, we have $\mu(E) = \mu^*(E)$. We can consider restricting the outer measure μ^* to the measurable sets $\mathcal{M}$, naturally introducing a measure function $\mu : \mathcal{M} \to [0, \infty]$. Thus, we construct a measure space $(X, \mathcal{M}, \mu)$.

Metric outer measure

The Metric outer measure is an introduced measure concept in metric spaces. Essentially, the outer measure is a more general concept compared to the measure, as it can be defined on any set family, while the metric outer measure is defined on metric spaces, typically measuring the size of sets using the diameter of sets.

Consider a metric space (X, d), where X is the set and $d : X \times X \to \mathbb{R}$ is the metric function as before. Consider a family of open balls $\{B\}$, where each open ball B_k is contained in the metric space, composed of a center x_k and a radius r_k, defined as:

$$B_k(x_k, r_k) = \{\alpha \in X : d(\alpha, x_k) < r_k\}$$

Then, I'll use $\mathcal{D}(B_k)$ to represent the diameter of the open ball B_k, and I'll continue to use this notation without further emphasis. The diameter of an open ball B_k is defined as the maximum distance between points it contains, that is:

$$\mathcal{D}(B_k) = \sup\{d(\alpha, \beta) : \alpha, \beta \in B_k\}$$

We define the metric outer measure of a subset $E \subseteq X$ as:

$$\psi(E) = \inf\left\{\sum_{k=1}^{\infty} \mathcal{D}(B_k) | E \subseteq \bigcup_{k=1}^{\infty} B_k\right\}$$

The metric outer measure, in all possible open ball coverings of E, i.e., $E \subseteq \bigcup_{k=1}^{\infty} B_k$, is the infimum of the diameters of all these open balls. Open balls are fundamental open sets in metric spaces, so essentially the construction is the same as before, just a change in notation. For example, in one-dimensional space, open balls degenerate into intervals, which are intuitively understood as lines. When the radius is 0, it degenerates into a point.

The metric outer measure possesses properties of general outer measures, along with some additional properties.

- If the distance between sets E_1 and E_2 in X is $d(E_1, E_2) > 0$, then $\psi(E_1 \cup E_2) = \psi(E_1) + \psi(E_2)$

Here, we will later refer to sets with a distance greater than zero as positively separated sets, or simply positively separated. Intuitively, this implies that the outer measure of the union of two disjoint sets equals the sum of their individual outer measures.

- For $A_n \in X$, if we define $A_1 \subseteq A_2 \subseteq \cdots \subseteq A = \bigcup_{n=1}^{\infty} A_n$ and ensure that A_n and A_{n+1} are positively separated, then:

$$\psi(A) = \sup \psi(A_n)$$

In some cases, by gradually increasing and positively separating the sequence of sets, the measure of the entire set tends to the least upper bound of the measure of any set in the sequence. This can be understood as the limit behavior of the measure on the sequence of gradually expanding sets. (There is not much explanation on this property in the Wiki.)

- Given a set $A \subseteq B$, where B is an open set, if $\psi(A) < \infty$, consider the positive integer n such that:

$$A_n = \{x \in A : d(x, B^c) \geq \frac{1}{n}\}$$

Then：

$$\lim_{n \to \infty} \psi(A) = \psi(A^n)$$

> The proof of this property in "Foundations of Modern Analysis" is very concise, my cat can understand it. There's no point in repeating it, and the specific page reference will be provided at the end of the article.
> Furthermore, this property is succinctly used in the book to prove the following theorem (1):

- (1) In the context of outer measure in metric spaces, every closed set is measurable.
- (2) In the context of outer measure in metric spaces, every Borel set is measurable.

Lebesgue measure

Lebesgue measure is the standard measure on Euclidean space, representing concepts like "length," "area," and "volume" in finite-dimensional Euclidean spaces (and even higher dimensions).

However, reality is often less perfect. For instance, with strict Lebesgue measurable sets, the concept is defined on a specific class of measurable sets. For some non-measurable sets, this concept does not apply. Therefore, as described earlier regarding outer measures, to make our constructed Lebesgue measure more general, we first need to establish a Lebesgue outer measure. Then, similar to the "shrinkage" process mentioned earlier, we construct a Lebesgue measure that can be defined on a broader class of sets.

In the one-dimensional Euclidean space $\mathbb{R}$ that we have been familiar with since childhood, we have the concept of "intervals." For example, an open interval can be represented as $I = (a, b)$, and we know that its length is $\ell(I) = b - a$. We also allow $\ell(I) = +\infty$, which is intuitive. However, in the n-dimensional Euclidean space $\mathbb{R}^n$ we cannot refer to intervals but instead use rectangles to represent regions in space. An open rectangle, representing an open region in space, is composed of the Cartesian product of n intervals:

$$R = I_1 \times I_2 \times \cdots \times I_n = (a, b) \times (a_1, b_1) \times \cdots \times (a_n, b_n) \subseteq \mathbb{R}^n$$

In the one-dimensional case, we already have the measure of intervals, which is the so-called length $\ell(I) = b - a$. Therefore, we just need to extend it to higher dimensions to represent the measure of open rectangles:

$$\lambda(R) = \prod_{i=1}^{n} (b_i - a_i)$$

Therefore, we can define the outer measure for a set $E \subseteq \mathbb{R}^n$. Consider a sequence of open rectangles $\{R\}$, then the outer measure (note that our R itself is considered in the real number field, so no additional restrictions are needed):

$$m^*(E) = \inf \left\{ \sum_{k=1}^{\infty} \lambda(R_k) : E \subseteq \bigcup_{k=1}^{\infty} R_k \right\}$$

This is when we call $\{R\}$ an L-covering of E. Simply put, it means that the union of elements in this set of open rectangle sets contains E, i.e., $E \subseteq \bigcup_{k=1}^{\infty} R_k$. At this point, we have a $\sum_{k=1}^{\infty} \lambda(R_k) \in \mathbb{R}$, which is the measure of the covering. Similar to the definition of outer measure we discussed earlier, due to the definition of infimum, for any $\epsilon > 0$, we have:

$$\sum_{k=1}^{\infty} \lambda(R_k) < m^*(E) + \epsilon$$

The idea is to take the infimum to ensure that our approximation measure does not exceed the actual outer measure $m^*(E)$, by introducing a small positive number ϵ to ensure that our approximation does not deviate too much from the actual measure.

Reflecting on the conditions mentioned in the previous section regarding the Carathéodory extension theorem, consider a set $A \subseteq \mathbb{R}$ that satisfies:

$$m^*(A) = m^*(A \cap E) + m(A \cap E^c)$$

Then we call E a Lebesgue measurable set. As mentioned earlier, the collection of these measurable sets forms a σ-algebra. For any Lebesgue measurable set E, its Lebesgue measure is defined as $m(E) = m^*(E)$, following the same logic as before.

Linear properties

Regarding the properties of Lebesgue measure, there's no need for me to elaborate here as Wikipedia has provided a comprehensive coverage.

> Wiki: https://en.wikipedia.org/wiki/Lebesgue_measure

Lebesgue measure has two important properties:

- Translational invariance: Translating a set does not change its measurability. Consider a Lebesgue measurable set $E \subseteq \mathbb{R}^n$ and shift its points by a vector $\mathbf{x}$, denoted as $E + \mathbf{x} = \{a + \mathbf{x} : a \in E\}$. Then, we have $m(E + \mathbf{x}) = m(E)$.
- Scaling property: Scaling a set does not change its measurability. Consider a Lebesgue measurable set $E \subseteq \mathbb{R}^n$ and scale its points by a scaling factor $c > 0$, denoted as $cE = \{ca : a \in E\}$. Then, we have $m(cE) = c^n m(E)$.

Both of these properties naturally lead to the idea of linear transformations. Indeed, we can directly generalize these properties using general linear transformations. Consider a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ and a Lebesgue measurable set $E \subseteq \mathbb{R}^n$. Then, $T(E)$ is also Lebesgue measurable, and satisfies:

$$m(T(E)) = |\det(T)| \cdot m(E)$$

- Here, $|\det(T)|$ represents the absolute value of the determinant.

$\mathbb{R}^n$ and Borel Set

Lebesgue outer measure is a metric outer measure, so all Borel sets in $\mathbb{R}^n$ are Lebesgue measurable. Conversely, Borel measurable sets are generated by the smallest σ-algebra generated by all open sets, typically incomplete, while Lebesgue measurable sets are generated by the σ-algebra generated by all open sets and null sets. Lebesgue measurable sets are complete Borel sets.

Borel sets have regularity in terms of outer measure, meaning they can be approximated from inside/outside by open/closed sets. This is divided into inner regularity and outer regularity.

- Inner regularity: For a Borel measurable set $E \subset \mathbb{R}^n$, there exists an open set $\mathcal{O}$ such that $E \subseteq \mathcal{O}$ and $m(\mathcal{O} \backslash E)$ can be arbitrarily small. This means we can approximate the interior of E with this open set $\mathcal{O}$.
- Outer regularity: For a Borel measurable set $E \subset \mathbb{R}^n$, there exists a closed set $\mathcal{C}$ such that $\mathcal{C} \subseteq E$ and $m(E \backslash \mathcal{C})$ can be arbitrarily small. This means we can approximate the exterior of E with this closed set $\mathcal{C}$.

Borel sets are considered representative of "well-behaved" sets. $\mathbb{R}^n$ and its open sets, closed sets, empty set, countable intersections of open sets, and countable unions of closed sets are all Borel sets. The intersection, union, and complement of Borel sets are also Borel sets, as well as the limit superior and limit inferior sets of Borel sets.

As a counterexample to illustrate the completeness of Lebesgue measure with respect to Borel sets, consider the Cantor set. The Cantor set is Borel measurable because it is constructed by finite intersections and complements of closed sets. However, there exist subsets of the Cantor set that are measurable but not in the Borel algebra. Under Lebesgue measure, the Cantor set has measure zero, making it measurable. This implies that Lebesgue measure can extend to sets beyond the Borel algebra... However, I would prefer to delve into this topic further in a discussion of abstract spaces or perhaps in a separate short article.

Summary

When I publish this English version, *A friendly introduction to real analysis II* Chinese version has already been published.

I've built a Github repository for mathematical analysis articles, where I will publish unified notation conventions, supplements, citation notes, errata, and anything related to my mathematical analysis blog:

https://github.com/MAKIROR/Mathematical-Analysis

References

1. Terence Tao. *陶哲轩实分析(第3版)*
2. Avner Friedman. *Foundations of Modern Analysis*
3. Halsey Royden, Patrick Fitzpatrick. *Real Analysis (4th Edition)*

Wiki:
https://en.wikipedia.org/wiki/Real_analysis
https://en.wikipedia.org/wiki/Construction_of_the_real_numbers
https://en.wikipedia.org/wiki/Measure_(mathematics)