LECTURE SHELL Ω: INTRODUCTION TO PROBABILITY THEORY ON THE METRIC SPACE OF INFINITE COIN-TOSSING SEQUENCES

MARK GAGARINE

ABSTRACT. The mechanisms regulating the rigorous laws of probability remain unknown to the budding student. In this lecture we introduce the foundational measure theory concepts which are used to develop the fundamental results of probability theory, with a focus on an application to problems involving coin tossing. Central towards the connection between measure theory and probability theory is the use of *Carathéodory's Extension Theorem*. We elucidate the theorem's critical role in bridging this transition by delving into its applications within probability, specifically through problems involving coin tossing. This examination clarifies the theoretical connections while simultaneously displaying the practicality of these foundational principles.

1. Introduction

In this lecture, we aim to provide a rigorous introduction to the formalization of probability and endow a mathematically sound framework for modelling probabilistic questions. We begin our analysis with introducing the measure theoretic definitions essential to the development of the framework under discussion. We clarify the notion of the "measure" of a set and explain the steps of defining a space which can assign such measures. We then transition our focus From Measure to Probability, starting in section 3, where we take a look at probability from a measure theoretic viewpoint. There we introduce the fundamental axioms of probability and establish some simple yet useful results in probability.

The crux of the lecture surrounds the discussion of Carathéodory's Extension Theorem. Despite the omitting the long and intricate proof, its significance is irrefutable, providing us with the necessary final steps in the assembly of our probability space. We conclude the section by working through a practical example, constructing a probability space that models tossing a coin infinitely many times.

2. Background and notation

In this section we will be describing the framework used to formally define what a probability is and introduce some mathematical tools used throughout the lecture. We start by diving into the concept of Measure Theory, the formalization of the "measure" of a set. That is, in $\mathbb{R}^n$ for example, how do we assign quantities such as lenth, volume, or surface area to a given subset A? What about an a different subset B? What about arbitrary subsets of a general metric space X! This is what we'll dive into.

Definition 2.1. Let X be an arbitrary non-empty space. Then the class $\mathcal{A}$ of subsets of X is called an *algebra* on X if the following properties hold.

(1) $X \in \mathcal{A}$
(2) $A \in \mathcal{A} \Rightarrow X \setminus A \in \mathcal{A}$
(3) $A_1, A_2, \cdots, A_n \in \mathcal{A} \Rightarrow A_1 \cup A_2 \cup \cdots \cup A_n \in \mathcal{A}$

In Definition (2.1), properties (1) and (2) both tell us an Algebra is closed under the compliment, but property (3) only ensures the space is closed under the union of **finitely** many subsets. This however is not good enough to describe probabilities, as we might want to work with infinitely many combinations of possibilities, for instance.

Definition 2.2. Let X be a non-empty space. The class $\mathcal{A}$ of subsets of X is called a σ-algebra if $\mathcal{A}$ is an *algebra* (in the sense of Definition (2.1)), which is also closed under the countable union of subsets. That is, for condition (3) we also have:

$$A_1, A_2, \cdots \in \mathcal{A} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{A}$$

Remark 2.3. It can be easily shown in by combining properties Definition (2.2) that $\emptyset \in \mathcal{A}$ in property (1), and $\cap_{i=1}^{\infty} A_i \in \mathcal{A}$ from combining (2) with DeMorgan's law.

Lemma 2.4. *Let $(\mathcal{A}_i)_{i \in I}$ be a family of σ-algebras on X. Then $\cap_{i \in I} \mathcal{A}_i$ is also a σ-algebra on X.*

Proof. We wish to show $\cap_{i \in I} A_i$ satisfies the 3 properties of a σ-algebra. That is,

(1) Since $\mathcal{A}_i$ is a σ-algebra, $X \in \mathcal{A}_i$ for every $i \in I$. So $X \in \cap_{i \in I} \mathcal{A}_i$.
(2) Let $A \in \mathcal{A}_i$, then $X \setminus A \in \mathcal{A}_i, \forall i \in I$. Hence $X \setminus A \in \cap_{i \in I} \mathcal{A}_i$
(3) Let $A_1, A_2, \cdots, \in \cap_{i \in I} \mathcal{A}_i$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}_i$ for every $i \in I$. So $\cup_{i=1}^{\infty} A_i \in \cap_{i \in I} \mathcal{A}_i$.

$\square$

Definition 2.5. Consider a family of subsets $\mathcal{M} \subseteq \mathcal{P}(X)$. Then the unique smallest σ-algebra containing every subset in $\mathcal{M}$, is the *σ-algebra generated* by $\mathcal{M}$, defined as

$$\sigma(\mathcal{M}) = \cap_{i \in I} \mathcal{A}_i, \forall \mathcal{A}_i \supseteq \mathcal{M}, \text{ where } \mathcal{A}_i \text{ is a } \sigma\text{-algebra.}$$

Example 2.6. Let $X = \{0, 1, 2\}$, and consider the family of subsets $\mathcal{M} = \{\{0\}, \{1\}\}$. Then the σ-algebra generated by $\mathcal{M}$ is, $\sigma(\mathcal{M}) = \{\emptyset, X, \{0\}, \{1\}, \{0, 1\}, \{1, 2\}, \{0, 2\}, \{2\}\}$. It is quite easy to verify that the conditions of a σ-algebra are met by inspection.

When working with finite spaces, it can be easy to find the smallest σ-algebra by construction. It is often the case the power set is the σ-algebra generated by a finite space and by some countable spaces. But if a space is countable and very large, or when we start looking at uncountable spaces, it quickly gets tricky and tedious to define the σ-algebra generated in an ad-hoc manner. In the context of metric spaces however, we have the following definition which we will later see becomes quite handy.

Definition 2.7. Let (X, d) be a metric space, and consider the topology $\mathcal{T}$ induced by d. Then the *Borel σ-algebra* is the σ-albegra generated by $\mathcal{T}$.

$$\mathcal{B}(X) = \sigma(\mathcal{T})$$

Definition and Remark 2.8. Let X be a set, and let $\mathcal{A}$ be a σ-algebra on X. Then the pair $(X, \mathcal{A})$ is a *measurable space.*

Definition 2.8 now gives us the notion of *measurable sets.* That is, we can now say the subset $A \subseteq \mathcal{A}$ is *measurable* if $(X, \mathcal{A})$ is a *measurable space*. So we introduce the function which assigns a measure to a set.

Definition and Remark 2.9. Let $(X, \mathcal{A})$ be a measurable space. Then the mapping $\mu(\mathcal{A}) : \mathcal{A} \to [0, \infty) \cup \{\infty\}$ is a *measure* if the following properties hold.

1. $\mu(\emptyset) = 0$
2. $\mu\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu\left(A_i\right)$, $\forall A_i \in \mathcal{A}, A_i \cap A_j = \emptyset, i \neq j$, called *countable additivity.*

The triple $(X, \mathcal{A}, \mu)$ is called a *measure space.* We now have enough tools to start developing the framework of measure theoretic probability theory.

3. From Measure to Probability

As we start to develop a probabilistic viewpoint on measure theory, we change up the notation slightly. Instead of referring to a space X, we now call it the sample space Ω, representing the set of all possible outcomes which can occur. The σ-algebra $\mathcal{A}$ is now referred to as the *event space*, representing a collection of events A - a set of possible outcomes. Finally, we also adjust our measure theory definitions slightly, as seen below.

Definition 3.1. The set function P on a σ-algebra $\mathcal{A}$ is a *probability measure* if the following properties hold.

1. $P(\emptyset) = 0,\ P(\Omega) = 1$
2. $0 \leq P(A) \leq 1, \forall A \in \mathcal{A}$
3. P satisfies *countable additivity* (as in Definition (2.9))

Notation 3.2. In probability theory, we further develop the concept of a measure space by restricting our focus to probability measures. This specialization transforms the general measure space into a more specific framework known as a *probability space*, denoted by the triple $(\Omega, \mathcal{A}, P)$.

It only becomes natural to work with finitely many events, and the following lemma provides us with an easy way to do so.

Lemma 3.3. *Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let sets $A_1, \cdots, A_n \in \mathcal{A}$ be disjoint. Then $\cup_{i=1}^{n} A_i$ is also in $\mathcal{A}$, and*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right).$$

Proof. Since $\mathcal{A}$ is a σ-algebra, $\emptyset \in \mathcal{A}$. Let $\emptyset = A_i,\ \forall i > n$. So $A_1, \cdots, A_i, \cdots \in \mathcal{A}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$ (by Def.(2.2.3)). Then $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right) = \sum_{i=1}^{n} P\left(A_i\right) + 0 + \cdots$ since $P(A_{n+1}) = P(A_{n+2}) = \cdots = P(\emptyset) = 0$. Notice that $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{n} A_i \cup \emptyset \cup \cdots = \cup_{i=1}^{n} A_i$.

So it follows that $P\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right)$, as required. □

This useful result is called *finite additivity*, which easily follows from *countable additivity.* It can give us a shortcut to prove some more interesting facts about probability measures.

Proposition 3.4. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $A \subseteq \mathcal{A}$ be an event. Define $A^c = \Omega \setminus A$ to be the compliment of A. Then,*

$$P(A) = 1 - P(A^c)$$

Proof. Notice A and A^c are disjoint. So by lemma 3.3, $P(A \cup A^c) = P(A) + P(A^c)$. We also note that $A \cup A^c = \Omega$, thus $P(A \cup A^c) = P(\Omega) = 1$. Hence we get the congruence

$$1 = P(A) + P(A^c) \Leftrightarrow P(A) = 1 - P(A^c)$$

□

Theorem 3.5. *(P is a continuous set function)*
Let $(\Omega, \mathcal{A}, P)$ be a probability space.

(1) *Suppose $A_1, A_2, \cdots \in \mathcal{A}$ is an increasing sequence of events such that $A_1 \subseteq A_2 \subseteq \cdots$ with $\lim_{i\to\infty} A_i = \cup_{i=1}^{\infty} A_i = A$. Then $P(A) = \lim_{i\to\infty} P(A_i)$.*
(2) *Suppose $B_1, B_2, \cdots \in \mathcal{A}$ is a decreasing sequence of events such that $B_1 \supseteq B_2 \supseteq \cdots$ with $\lim_{i\to\infty} B_i = \cap_{i=1}^{\infty} B_i = B$. Then $P(B) = \lim_{i\to\infty} P(B_i)$.*

Proof. For (1), choose A to be written as the disjoint unions where $A_0 = \emptyset$ and

$$A = A_1 \cup (A_2 \setminus A_1)) \cup (A_3 \setminus A_2) \cup \cdots$$

Then by countable additivity, $P(A) = \sum_{k=1}^{\infty} P(A_k \setminus A_{k-1}) = \lim_{i\to\infty} \sum_{k=1}^{i} \{P(A_k) - P(A_{k-1})\}$. So it follows that $P(A) = \lim_{i\to\infty} P(A_i)$ as required.

For (2), choose $A_i = B_i^c$ in the preceding proof. They by applying proposition 3.4, we obtain $1 - P(A) = 1 - \lim_{i\to\infty} P(A_i)$. □

The preceding properties are called *continuity from above* and *continuity from below.*

4. Carathéodory's Extension Theorem

Continuing our decent into the framework of probability theory, we now encounter a theorem which enables us to define a probability measure on any set we are interested in working with. However, the natural question arises: what if our set of choice doesn't form a σ-algebra? How then would we be allowed to apply the principals of probability we found in section 3? This is where Carathéodory's Extension Theorem plays a critical role. In essence, the theorem provides us with a general method to find a sensible σ-algebra to work with. Moreover, it facilitates the unique *extension* of a probability measure to this

σ-algebra, bequeathing us with a new measure which coincides with the original measure on our set. Despite the significance of this theorem, its proof is delves deeper into measure and integration theory, which beyond the scope of this lecture.

Theorem 4.1. *(Carathéodory's Extension Theorem)* [1] *A probability measure P defined on an algebra $\mathcal{A}$ has a unique extension to a probability measure on $\sigma(\mathcal{A})$.*

Theorem 4.1 is remarkable since it requires us to initially define a probability measure on a rather small algebra of sets, while ensuring a unique extension to a σ-algebra. What's particularly interesting is the case when the algebra on our metric space X is such that is relates to or generates the open sets. We then have an extension of our measure to the σ-algebra generated by these open sets - the Borel σ-algebra of X, $\mathcal{B}(X)$ (as in Definition(2.7)).

We examine how to apply theorem 4.1 to find a suitable probability space modelling infinite sequences of coin tosses. Suppose we let S represent the finite set of all possible outcomes from a simple experiment - in the case of coin tossing, $S = \{0, 1\}$, with 0 and 1 representing a coin landing on tails or heads, respectively. Example 4.2 will demonstrate how theorem 4.1 allows us to find a suitable probability space that broadly model sequences resulting from repeating **any** simple experiment infinitely many times.

Example 4.2. Let $S = \{0, 1\}$ denote the possible outcomes of a coin toss as above, and let $\Omega = S^\infty$ be the space of all infinite binary strings

$$\omega = \left(X^{(1)}(\omega), X^{(2)}(\omega), \cdots\right) \in \Omega, \text{ where } X^{(n)}(\omega) \in S$$

The sample space Ω is precisely the space used to define the compact metric space with weighted hamming distance - as discussed in the Lecture Shell 15 from [3]. In fact, Ω is the countable cartesian product of copies of S, and each $X^{(n)}(\cdot) : \Omega \to S$ is called the *natural projection of* Ω *onto* S.

Let $S^n = S \times \cdots \times S$ be the finite cartesian product of n copies of S, consisting of the sequences of outcomes $\left(x^{(1)}, \cdots, x^{(n)}\right)$ of n many elements of S. These sequence of outcomes then represent the first n repetitions of the experiment.

Next, we define the *cylinder of rank* n to be the set

$$A = \left\{\omega : \left(X^{(1)}(\omega), \cdots, X^{(n)}(\omega)\right) \in H\right\}, \text{ where } H \subseteq S^n$$

Now let $\mathcal{C}_0$ be the class consisting of cylinders of all ranks. We claim $\mathcal{C}_0$ is an algebra.

Verification of Claim. For the rest of the verification, fix a cylinder $A \in \mathcal{C}_0$. If $H = \emptyset$, then $A = \left\{\omega : \left(X^{(1)}(\omega), \cdots, X^{(n)}(\omega)\right) \in \emptyset\right\} = \emptyset$. If we replace H by $S^n \setminus H$, then

$$\Omega \setminus A = \left\{\omega : \left(X^{(1)}(\omega), \cdots, X^{(n)}(\omega)\right) \notin H\right\} \in \mathcal{C}_0.$$

[1] This version of the extension theorem is sometimes referred to as the *Hahn-Kolmogrov Theorem.* It could be seen as the "little brother" to Carathéodory due to it having a more restricted application to probability measures. The general extension theorem, as originally proposed by Carathéodory, stipulates that a *pre-measure* defined on a *semi-ring of sets* on X, $\mathcal{R}$, extends to a measure on the σ-algebra generated by $\mathcal{R}$, and the extension is unique if the pre-measure is *σ-finite.*

Finally, define $B = \{\omega : (X^{(1)}(\omega), \cdots, X^{(n)}(\omega)) \in I\}$ to be the cylinder of rank m. Without loss of generality, take $n \leq m$. $m < n$ follows by symmetry. Suppose H' is a collection of sequences $(x^{(1)}, \cdots x^{(m)})$ in S^m, where the truncated sequence $(x^{(1)}, \cdots x^{(n)})$ lies in H in S^n. So rewrite $A = \{\omega : (X^{(1)}(\omega), \cdots, X^{(n)}(\omega)) \in H'\}$. It is clear that we get the cylinder

$$A \cup B = \Big\{\omega : \Big(X^{(1)}(\omega), \cdots, X^{(n)}(\omega)\Big) \in H' \cup I\Big\} \in \mathcal{C}_0$$

So it follows tha $\mathcal{C}_0$ is closed under finite unions, and hence it is an indeed an algebra. $\square$

Now let p_x for $x \in S$ be *probabilities.* That is, $p_x \geq 0, \forall x \in S$ and $\sum_{x \in S} p_x = 1$.

Define the set function[2] P on $\mathcal{C}_0$ where $\mathcal{C}$ is the σ-algebra of Ω generated by $\mathcal{C}_0$.

$$P(A) = \sum_H (p_{x^{(1)}} \cdots p_{x^{(n)}})$$

Since we said S represents the outcomes of tossing a *fair*[3] coin, set $p_0 = p_1 = \frac{1}{2}$. So for a cylinder $A \in \mathcal{C}_0$, take $P(A) = \sum_H \frac{1}{2^n}$. We now claim P defines a *countably additive probability measure on the algebra* $\mathcal{C}_0$.

Verification of Claim. We begin by showing P defines a finitely additive probability measure on the algebra $\mathcal{C}_0$.

So, without loss of generality suppose $n \leq m$. Take A and B to be the cylinders of rank m defined in the preceding claim. Then A and B are disjoint, hence H' and I are disjoint as well. So by lemma 3.3 we get

$$P(A \cup B) = \sum_{H' \cup I} (p_{x^{(1)}} \cdots p_{x^{(m)}}) = P(A) + P(B)$$

Now if we take $H = S^n$ for every $n \geq 1$, we get $A = \Omega$. Thus we can confirm $P(\Omega) = 1$. We have now verified that P in fact does define a finitely additive probability measure on C_0. By using the fact[4] that every finitely additive probability measure on the algebra C_0 of cylinders in Ω is countably additive, we are now done with the claim and can proceed to the remaining piece of the example. $\square$

Let $\mathcal{C}$ be the σ-algebra of Ω generated by $\mathcal{C}_0$. Then by Theorem 4.1, P defined on $\mathcal{C}_0$ uniquely extends to a probability measure on $\mathcal{C}$. Hence, $(\Omega, \mathcal{C}, P)$ is a probability space which models coin tossing, with P also being the product measure on $\mathcal{C}$.

Remark 4.3. As a result of Carathéodory's Extension Theorem, we have now defined a probability measure for our sample space Ω. This lets us ask some interesting questions about specific combinations of coin tosses.

[2] This measure is sometimes called the *product measure.*

[3] Fair as in meaning all outcomes are equally likely.

[4] This follows from Theorem 2.3 in section 2 of chapter 1 from [1]

5. Working with Infinite Coin-Tosses

Now for a more practical approach to probability, we can use the theory we have developed to practice the computational questions one may encounter in an introductory probability course. For the rest of the section, fix $(\Omega, \mathcal{A}, P)$ to be the probability space modelling infinite sequences of coin tosses as we found in example 4.2.

Definition 5.1. Let $(Y, \mathcal{B})$ be a measurable space, and define the function $X : \Omega \to Y$ which for any $B \in \mathcal{B}$, the pre-image of B is contained in $\mathcal{A}$[5]. Then X is a *random element*.

A random element generally describes characteristics of outcomes that may be of special interest to us. This allows us to formulate a question into notation that integrates nicely with the theory and formulas we have developed.

Example 5.2. What is the probability that a random sequence $X \in \Omega$ starts with 100 occurrences of the digit 1?

Solution. Define the sequence of random variables $X^{(i)} : \Omega \to \mathbb{R}$, with each $X^{(i)} = 1$ if the i^{th} coin toss in the sequence ω is heads (1), and 0 otherwise[6].

Let $A = \{\omega \in \Omega : X^{(1)} = \cdots = X^{(100)=1}\}$ represent the event that the first 100 elements of the sequence are 1. That is, A is the set of random sequences in Ω whose first 100 observed outcomes are all 1, with the form $(1, \cdots, 1, x^{(101)}, \cdots)$.

Plugging A into P, we get

$$P(A) = \sum_{A} (p_{x^{(1)}} \cdot p_{x^{(2)}} \cdots) = \frac{1}{2^{100}}$$

□

We can verify this result intuitively as well. It also follows from the independence of tossing fair coins that, we get the probability of observing 100 consecutive 1's from the start given by the product of the probabilities for all 100 tosses.

[5] A function which satisfies this condition is said to be $(\mathcal{A}, \mathcal{B})$ *measurable*.

[6] This is often referred to as an *indicator random variable*.

Example 5.3. Next, we ask what is the probability that a random sequence $X \in \Omega$ contains 100 occurrences of the digit 1 starting from anywhere in the sequence?

Solution. Define E to be the event where 100 consecutive occurrences of the digit 1 occur in the sequence. Further, lets define F to be the random sequence of independent blocks consisting of 100 coin tosses. We can then use property 3.4 to simplify the calculation and instead find $1-P(F^c)$. That is, we are looking for 1 minus the probability a block consisting of 100 consecutive occurrences of the digit 1 does not occur.

Notice, we can represent F as the cylinder $F = \left\{\omega : \left(X^{(1)}(\omega), X^{(2)}(\omega), \cdots\right) \in I\right\}$ where each $X^{(i)}(\omega)$ represents the outcome of 100 coin tosses. Then each probability of the compliment is $p_{x^{(i)}} = 1 - \frac{1}{2^{100}}$, which follows from 5.2. So

$$P(F^c) = \sum_I (p_{x^{(1)}} \cdot p_{x^{(2)}} \cdots)$$

Since $E^c \subseteq F^c$ and using the fact that $A \subseteq B \Rightarrow P(A) \leq P(B)$, we see

$$P(E^c) \leq P(F^c) = \left(1 - \frac{1}{2^{100}}\right)^{\infty} = 0$$

Thus, $0 \leq P(E^c) \leq 0$. So by the squeeze theorem, we get $P(E^c) = 0$. It follows that

$$P(E) = 1 - 0 = 1$$

. $\square$

References

[1] Patrick P. Billingsley. *Probability and Measure*, 3rd edition, Wiley Publishers, 1995.

[2] R.M. Dudley. *Real analysis and probability*, 2nd edition, Cambridge University Press, 2003.

[3] A. Nica. Lecture shells for PMath 351 lectures in Winter Term 2024, available on the Learn web-site of the course.

Faculty of Mathematics, University of Waterloo

Email address: magagarine@uwaterloo.ca