

PROBABILISTIC MODELS AND PROBABILITY MEASURES**Contents**

1. Probabilistic experiments
2. Sample space
3. Discrete probability spaces
4. σ -fields (σ -algebras)
5. Probability measures
6. Continuity of probabilities
7. Monotone class theorem

1 PROBABILISTIC EXPERIMENTS

Probability theory is a mathematical framework that allows us to reason about phenomena or experiments whose outcome is uncertain. A probabilistic model is a mathematical model of a probabilistic experiment that satisfies certain mathematical properties (the axioms of probability theory), and which allows us to calculate probabilities and to reason about the likely outcomes of the experiment.

A probabilistic model is defined formally by a triple $(\Omega, \mathcal{F}, \mathbb{P})$, called a **probability space**, comprised of the following three elements:

- (a) Ω is the **sample space**, the set of possible outcomes of the experiment.
- (b) \mathcal{F} is a **σ -field**, a collection of subsets of Ω . (The term “ σ -algebra” is also commonly used, as a synonym.)
- (c) \mathbb{P} is a **probability measure**, a function that assigns a nonnegative probability to every set in the σ -field \mathcal{F} .

Our objective is to describe the three elements of a probability space, and explore some of their properties.

2 SAMPLE SPACE

The sample space is a set Ω comprised of all the possible outcomes of the experiment. Typical elements of Ω are often denoted by ω , and are called **elementary outcomes**, or simply outcomes. The sample space can be finite, e.g.,
 $= \{\omega_1, \dots, \omega_n\}$, countable, e.g., $\Omega = \mathbb{N}$, or uncountable, e.g., $\Omega = \mathbb{R}$ or
 $= \{0, 1\}^\infty$.

As a practical matter, the elements of Ω must be mutually exclusive and collectively exhaustive, in the sense that once the experiment is carried out, there is exactly one element of Ω that occurs.

Examples

- (a) If the experiment consists of a single roll of an ordinary die, the natural sample space is the set $\Omega = \{1, 2, \dots, 6\}$, consisting of 6 elements. The outcome $\omega = 2$ indicates that the result of the roll was 2.
- (b) If the experiment consists of five consecutive rolls of an ordinary die, the natural sample space is the set $\Omega = \{1, 2, \dots, 6\}^5$. The element $\omega = (3, 1, 1, 2, 5)$ is an example of a possible outcome.
- (c) If the experiment consists of an infinite number of consecutive rolls of an ordinary die, the natural sample space is the set $\Omega = \{1, 2, \dots, 6\}^\infty$. In this case, an elementary outcome is an infinite sequence, e.g., $\omega = (3, 1, 1, 5, \dots)$. Such a sample space would be appropriate if we intend to roll a die indefinitely and we are interested in studying, say, the number of rolls until a 4 is obtained for the first time.
- (d) If the experiment consists of measuring the velocity of a vehicle with infinite precision, a natural sample space is the set \mathbb{R} of real numbers.

Note that there is no discussion of probabilities so far. The set Ω simply specifies the possible outcomes.

3 DISCRETE PROBABILITY SPACES

Before continuing with the discussion of σ -fields and probability measures in their full generality, it is helpful to consider the simpler case where the sample space Ω is finite or countable.

Definition 1. A discrete probability space is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$ such that:

- (a) The sample space Ω is finite or countable: $\Omega = \{\omega_1, \omega_2, \dots\}$.
- (b) The σ -field \mathcal{F} is the set of all subsets of Ω .
- (c) The probability measure assigns a number in the set $[0, 1]$ to every subset of Ω . It is defined in terms of the probabilities $\mathbb{P}(\{\omega\})$ of the elementary outcomes, and satisfies

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}), \quad (1)$$

for every $A \subset \Omega$, and

$$\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1.$$

For simplicity, we will usually employ the notation $\mathbb{P}(\omega)$ instead of $\mathbb{P}(\{\omega\})$, and we will often denote $\mathbb{P}(\omega_i)$ by p_i .

The following are some examples of discrete probability spaces. Note that typically we do not provide an explicit expression for $\mathbb{P}(A)$ for every $A \subset \Omega$. It suffices to specify the probability of elementary outcomes, from which $\mathbb{P}(A)$ is readily obtained for any A .

Examples.

- (a) Consider a single toss of a coin. If we believe that heads (H) and tails (T) are equally likely, the following is an appropriate model. We set $\Omega = \{\omega_1, \omega_2\}$, where $\omega_1 = H$ and $\omega_2 = T$, and let $p_1 = p_2 = 1/2$. Here, $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$, and $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(H) = \mathbb{P}(T) = 1/2$, $\mathbb{P}(\{H, T\}) = 1$.
- (b) Consider a single roll of a die. if we believe that all six outcomes are equally likely, the following is an appropriate model. We set $\Omega = \{1, 2, \dots, 6\}$ and $p_1 = \dots = p_6 = 1/6$.
- (c) This example is not necessarily motivated by a meaningful experiment, yet it is a legitimate discrete probability space. Let $\Omega = \{1, 2, 5, a, v, aaa, *\}$, and $\mathbb{P}(1) = .1$, $\mathbb{P}(2) = .1$, $\mathbb{P}(5) = .3$, $\mathbb{P}(a) = .15$, $\mathbb{P}(v) = .15$, $\mathbb{P}(aaa) = .2$, $\mathbb{P}(*) = 0$.
- (d) Let $\Omega = \mathbb{N}$, and $p_k = (1/2)^k$, for $k = 1, 2, \dots$. More generally, given a parameter $p \in [0, 1)$, we can define $p_k = (1-p)p^{k-1}$, for $k = 1, 2, \dots$. This results in a legitimate probability space because $\sum_{k=1}^{\infty} (1-p)p^{k-1} = 1$.
- (e) Let $\Omega = \mathbb{N}$. We fix a parameter $\lambda > 0$, and let $p_k = e^{-\lambda} \lambda^k / k!$, for $k = 0, 1, \dots$. This results in a legitimate probability space because $\sum_{k=0}^{\infty} e^{-\lambda} \lambda^k / k! = 1$.
- (f) We toss an unbiased coin n times. We let $\Omega = \{0, 1\}^n$, and if we believe that all sequences of heads and tails are equally likely, we let $\mathbb{P}(\omega) = 1/2^n$ for every $\omega \in \Omega$.

- (g) We roll a die n times. We let $\Omega = \{1, 2, \dots, 6\}^n$, and if we believe that all elementary outcomes (6 -long sequences) are equally likely, we let $\mathbb{P}(\omega) = 1/6^n$ for every $\omega \in \Omega$.

Given the probabilities p_i , the problem of determining $\mathbb{P}(A)$ for some subset of Ω is conceptually straightforward. However, the calculations involved in determining the value of the sum $\sum_{\omega \in A} \mathbb{P}(\omega)$ can range from straightforward to daunting. Various methods that can simplify such calculations will be explored in future lectures.

4 σ -FIELDS

When the sample space Ω is uncountable, the idea of defining the probability of a general subset of Ω in terms of the probabilities of elementary outcomes runs into difficulties. Suppose, for example, that the experiment consists of drawing a number from the interval $[0, 1]$, and that we wish to model a situation where all elementary outcomes are “equally likely.” If we were to assign a probability of zero to every ω , this alone would not be of much help in determining the probability of a subset such as $[1/2, 3/4]$. If we were to assign the same positive value to every ω , we would obtain $\mathbb{P}(\{1, 1/2, 1/3, \dots\}) = \infty$, which is undesirable. A way out of this difficulty is to work directly with the probabilities of more general subsets of Ω (not just subsets consisting of a single element).

Ideally, we would like to specify the probability $\mathbb{P}(A)$ of every subset of Ω . However, if we wish our probabilities to have certain intuitive mathematical properties, we run into some insurmountable mathematical difficulties. A solution is provided by the following compromise: assign probabilities to only a partial collection of subsets of Ω . The sets in this collection are to be thought of as the “nice” subsets of Ω , or, alternatively, as the subsets of Ω of interest. Mathematically, we will require this collection to be a σ -field, a term that we define next.

Definition 2. Given a sample space Ω , a **σ -field** is a collection \mathcal{F} of subsets of Ω , with the following properties:

- (a) $\emptyset \in \mathcal{F}$.
- (b) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
- (c) If $A_i \in \mathcal{F}$ for every $i \in \mathbb{N}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A set A that belongs to \mathcal{F} is called an **event**, an **\mathcal{F} -measurable set**, or simply a **measurable set**. The pair (Ω, \mathcal{F}) is called a **measurable space**.

Remark. A σ -field is often called a **σ -algebra**, and these terms will be used interchangeably. If we relax condition (c) and require only finite unions to be in \mathcal{F} , we get a definition of field (or algebra) of sets – see Def. 4 below.

The term “event” is to be understood as follows. Once the experiment is concluded, the realized outcome ω either belongs to A , in which case we say that the event A has occurred, or it doesn’t, in which case we say that the event did not occur.

It turns out that if $A_i \in \mathcal{F}$ for every $i \in \mathbb{N}$, then $\cap_{i=1}^n A_i \in \mathcal{F}$, i.e., a σ -field is closed under countable intersections as well.

Exercise 1.

- (a) Let \mathcal{F} be a σ -field. Prove that if $A, B \in \mathcal{F}$, then $A \cap B \in \mathcal{F}$. More generally, given a countably infinite sequence of events $A_i \in \mathcal{F}$, prove that $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$. Hint: Use De Morgan’s law.
- (b) Prove that property (a) of σ -fields (that is, $\emptyset \in \mathcal{F}$) can be derived from properties (b) and (c), assuming that the σ -field \mathcal{F} is non-empty.

The following are some examples of σ -fields. (Check that this is indeed the case.)

Examples.

- (a) The trivial σ -field, $\mathcal{F} = \{\emptyset, \Omega\}$.
- (b) The collection $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$, where A is a fixed subset of Ω .
- (c) The set of all subsets of Ω : $\mathcal{F} = 2^{\Omega} = \{A \mid A \subset \Omega\}$.
- (d) Let $\Omega = \{1, 2, \dots, 6\}^n$, the sample space associated with n rolls of a die. Let $A = \{\omega = (\omega_1, \dots, \omega_n) \mid \omega_1 \leq 2\}$, $B = \{\omega = (\omega_1, \dots, \omega_n) \mid 3 \leq \omega_1 \leq 4\}$, and $C = \{\omega = (\omega_1, \dots, \omega_n) \mid \omega_1 \geq 5\}$, and $\mathcal{F} = \{\emptyset, A, B, C, A \cup B, A \cup C, B \cup C, \Omega\}$.

Example (d) above can be thought of as follows. We start with a number of subsets of Ω that we wish to have included in a σ -field (the sets A , B , and C , in this example). We then include more subsets, as needed, until a σ -field is constructed. More generally, given a collection of subsets of Ω , we can contemplate forming complements, countable unions, and countable intersections of these subsets, to form a new collection. We continue this process until no more sets are included in the collection, at which point we obtain a σ -field. This process is hard to formalize in a rigorous manner. An alternative way of defining this σ -field is provided below. We will need the following fact.

Proposition 1. *Let S be an index set (possibly infinite, or even uncountable), and suppose that for every s we have a σ -field \mathcal{F}_s of subsets of the same sample space. Let $\mathcal{F} = \cap_{s \in S} \mathcal{F}_s$, i.e., a set A belongs to \mathcal{F} if and only if $A \in \mathcal{F}_s$ for every $s \in S$. Then \mathcal{F} is a σ -field.*

Proof. We need to verify that \mathcal{F} has the three required properties. Since each \mathcal{F}_s is a σ -field, we have $\emptyset \in \mathcal{F}_s$, for every s , which implies that $\emptyset \in \mathcal{F}$. To establish the second property, suppose that $A \in \mathcal{F}$. Then, $A \in \mathcal{F}_s$, for every s . Since each \mathcal{F}_s is a σ -field, we have $A^c \in \mathcal{F}_s$, for every s . Therefore, $A^c \in \mathcal{F}$, as desired. Finally, to establish the third property, consider a sequence $\{A_i\}$ of elements of \mathcal{F} . In particular, for a given $s \in S$, every set A_i belongs to \mathcal{F}_s . Since \mathcal{F}_s is a σ -field, it follows that $\cup_{i=1}^{\infty} A_i \in \mathcal{F}_s$. Since this is true for every $s \in S$, it follows that $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$. This verifies the third property and establishes that \mathcal{F} is indeed a σ -field. \square

Suppose now that we start with a collection \mathcal{C} of subsets of Ω , which is not necessarily a σ -field. We wish to form a σ -field that contains \mathcal{C} . This is always possible, a simple choice being to just let $\mathcal{F} = 2^{\Omega}$. However, for technical reasons, we may wish the σ -field to contain no more sets than necessary. This leads us to define \mathcal{F} as the intersection of all σ -fields that contain \mathcal{C} . Note that if \mathcal{H} is any other σ -field that contains \mathcal{C} , then $\mathcal{F} \subset \mathcal{H}$. (This is because \mathcal{F} was defined as the intersection of various σ -fields, one of which is \mathcal{H} .) In this sense, \mathcal{F} is the **smallest** σ -field containing \mathcal{C} . The σ -field \mathcal{F} constructed in this manner is called the σ -field **generated** by \mathcal{C} , and is often denoted by $\sigma(\mathcal{C})$.

Example. Let $\Omega = [0, 1]$. The smallest σ -field that includes every interval $[a, b] \subset [0, 1]$ is hard to describe explicitly (it includes fairly complicated sets), but is still well-defined, by the above discussion. It is called the **Borel** σ -field, and is denoted by \mathcal{B} . A set $A \subset [0, 1]$ that belongs to this σ -field is called a **Borel set**.¹

¹The Borel σ -field is usually defined as the σ -field generated by the collection of open sets of

4.1 Other reasons for “small” σ -fields.

As we discussed earlier, one reason for using a σ -field which does not include all subsets of Ω is in order to avoid insurmountable mathematical difficulties. However, there is also another reason: we may want to capture the perspective of an observer who receives only partial information about the outcome of the experiment. In that case, it is convenient (loosely speaking) to let \mathcal{F} be just the set of events for which the observer will be able to tell whether they occurred or not.

With this perspective, a σ -field can be viewed as an abstract description of the information that an observer receives. In particular, if the information available to observers 1 and 2 is described by σ -fields \mathcal{F}_1 and \mathcal{F}_2 , respectively, and if $\mathcal{F}_2 \subset \mathcal{F}_1$, we have a situation in which observer 2 has less information.

Example. We flip a coin twice, and each flip results in Heads (H) or Tails (T). In this context, $\Omega = \{HH, HT, TH, TT\}$. The natural σ -field, \mathcal{F}_1 , is the collection of all subsets of Ω . Consider now an observer who sees only the result of the first coin flip. In this case, we describe the information available to that observer in terms of the smaller σ -field

$$\mathcal{F}_2 = \left\{ \emptyset, \Omega, \{HH, HT\}, \{TH, TT\} \right\}.$$

In particular, this observer can tell whether the event $\{HH, HT\}$ has occurred or not, but cannot tell whether the event $\{HH\}$ has occurred.

We will turn to this association of σ -fields to observers much later, when we consider conditional expectations given partial information.

5 PROBABILITY MEASURES

We are now ready to discuss the assignment of probabilities to events. We have already seen that when the sample space Ω is countable, this can be accomplished by assigning probabilities to individual elements $\omega \in \Omega$. However, as discussed before, this does not work when Ω is uncountable. We are then led to assign probabilities to certain subsets of Ω , specifically to the elements of a σ -field \mathcal{F} , and require that these probabilities have certain “natural” properties.

Besides probability measures, it is also convenient to define the notion of a measure more generally.

We will be using the following terminology. We say that a collection of sets $A_\alpha \subset \Omega$, where α ranges over some index set is **mutually exclusive** or that the sets are **disjoint** if $A_\alpha \cap A_{\alpha'} = \emptyset$, whenever $\alpha \neq \alpha'$. Also, the sets $A_\alpha \subset \Omega$ are called **collectively exhaustive** if $\cup_\alpha A_\alpha = \Omega$.

a topological space. But for the case of the unit interval our definition is an equivalent one.

Definition 3. Let (Ω, \mathcal{F}) be a measurable space. A **measure** is a function $\mu : \mathcal{F} \rightarrow [0, \infty]$, which assigns a nonnegative extended real number $\mu(A)$ to every set A in \mathcal{F} , and which satisfies the following two conditions:

- (a) $\mu(\emptyset) = 0$;
- (b) (**Countable additivity, or σ -additivity**) If $\{A_i\}$ is a sequence of disjoint sets that belong to \mathcal{F} , then $\mu(\cup_i A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

A **probability measure** is a measure \mathbb{P} with the additional property $\mathbb{P}(\Omega) = 1$. In that case, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

In short, a measure is a nonnegative extended real valued σ -additive set function with domain \mathcal{F} .

For any $A \in \mathcal{F}$, $\mathbb{P}(A)$ is called the probability of the event A . The assignment of unit probability to the event Ω expresses our certainty that the outcome of the experiment, no matter what it is, will be an element of Ω . Similarly, the outcome cannot be an element of the empty set; thus, the empty set cannot occur and is assigned zero probability. If an event $A \in \mathcal{F}$ satisfies $\mathbb{P}(A) = 1$, we say that A occurs **almost surely**. Note, however, that A happening almost surely is not the same as the condition $A = \Omega$. For a trivial example, let $\Omega = \{1, 2, 3\}$, $p_1 = .5$, $p_2 = .5$, $p_3 = 0$. Then the event $A = \{1, 2\}$ occurs almost surely, since $\mathbb{P}(A) = .5 + .5 = 1$, but $A \neq \Omega$. The outcome 3 has zero probability, but is still possible. We will study more interesting examples of almost sure events later on when we give examples of non-discrete probability spaces.

The countable additivity property is very important. Its intuitive meaning is the following. If we have several events A_1, A_2, \dots , out of which at most one can occur, then the probability that “one of them will occur” is equal to the sum of their individual probabilities. In this sense, probabilities (and more generally, measures) behave like the familiar notions of area or volume: the area or volume of a countable union of disjoint sets is the sum of their individual areas or volumes. Indeed, a measure is to be understood as some generalized notion of a volume. In this light, allowing the measure $\mu(A)$ of a set to be infinite is natural, since one can easily think of sets with infinite volume.

The properties of probability measures that are required by Definition 3 are often called the axioms of probability theory. Starting from these axioms, many other properties can be derived, as in the next proposition.

Proposition 2. Probability measures have the following properties.

- (a) **(Finite additivity)** If the events A_1, \dots, A_n are disjoint, then $\mathbb{P}(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i)$.
- (b) For any event A , we have $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- (c) If the events A and B satisfy $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- (d) **(Union bound)** For any sequence $\{A_i\}$ of events, we have

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

- (e) **(Inclusion-exclusion formula)** For any collection of events A_1, \dots, A_n ,

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{(i,j): i < j} \mathbb{P}(A_i \cap A_j) \\ &\quad + \sum_{(i,j,k): i < j < k} \mathbb{P}(A_i \cap A_j \cap A_k) + \dots + (-1)^{n-1} \mathbb{P}(A_1 \cap \dots \cap A_n). \end{aligned}$$

Proof.

- (a) This property is almost identical to condition (b) in the definition of a measure, except that it deals with a finite instead of a countably infinite collection of events. Given a finite collection of disjoint events A_1, \dots, A_n , let us define $A_k = \emptyset$ for $k > n$, to obtain an infinite sequence of disjoint events. Then,

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

Countable additivity was used in the second equality, and the fact $\mathbb{P}(\emptyset) = 0$ was used in the last equality.

- (b) The events A and A^c are disjoint. Using part (a), we have $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$. But $A \cup A^c = \Omega$, whose measure is equal to one, and the result follows.
- (c) The events A and $B \setminus A$ are disjoint. Also, $A \cup (B \setminus A) = B$. Therefore, using also part (a), we obtain $\mathbb{P}(A) \leq \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(B)$.
- (d) Left as an exercise.

- (e) Left as an exercise; a simple proof will be provided later, using random variables. \square

For the special case where $n = 2$, part (e) of Proposition 2 simplifies to

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Let us note that all properties (a), (c), and (d) in Proposition 2 are also valid for general measures (the proof is the same). Let us also note that for a probability measure, the property $\mathbb{P}(\emptyset) = 0$ need not be assumed, but can be derived from the other properties. Indeed, consider a sequence of sets A_i , each of which is equal to the empty set. These sets are disjoint, since $\emptyset \cap \emptyset = \emptyset$. Applying the countable additivity property, we obtain $\sum_{i=1}^{\infty} \mathbb{P}(\emptyset) = \mathbb{P}(\emptyset) \leq \mathbb{P}(\Omega) = 1$, which can only hold if $\mathbb{P}(\emptyset) = 0$.

Finite Additivity

Our definitions of σ -fields and of probability measures involve countable unions and a countable additivity property. A different mathematical structure is obtained if we replace countable unions and sums by finite ones. This leads us to the following definitions.

Definition 4. Let Ω be a sample space.

(a) A **field** is a collection \mathcal{F}_0 of subsets of Ω , with the following properties:

- (i) $\emptyset \in \mathcal{F}$.
- (ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
- (iii) If $A \in \mathcal{F}$ and $B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$.

(b) Let \mathcal{F}_0 be a field of subsets of Ω . A function $\mathbb{P} : \mathcal{F}_0 \rightarrow [0, 1]$ is said to be **finitely additive** if

$$A, B \in \mathcal{F}_0, A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

Remark. A field (of sets) is often called an **algebra** (of sets), and these terms will be used interchangeably.

We note that finite additivity, for the two case of two events, easily implies finite additivity for a general finite number n of events, namely, the property in

part (a) of Proposition 2. To see this, note that finite additivity for $n = 2$ allows us to write, for the case of three disjoint events,

$$\mathbb{P}(A_1 \cup A_2 \cup A_3) = \mathbb{P}(A_1) + \mathbb{P}(A_2 \cup A_3) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3),$$

and we can proceed inductively to generalize to the case of n events.

Finite additivity is strictly weaker than the countable additivity property of probability measures. In particular, finite additivity on a field, or even for the special case of a σ -field, does not, in general, imply countable additivity. The reason for introducing the stronger countable additivity property is that without it, we are severely limited in the types of probability calculations that are possible. On the other hand, finite additivity is often easier to verify.

6 CONTINUITY OF PROBABILITIES

Consider a probability space in which $\Omega = \mathbb{R}$. The sequence of events $A_n = [1, n]$ converges to the event $A = [1, \infty)$, and it is reasonable to expect that the probability of $[1, n]$ converges to the probability of $[1, \infty)$. Such a property is established in greater generality in the result that follows. This result provides us with a few alternative versions of such a continuity property, together with a converse which states that finite additivity together with continuity implies countable additivity. This last result is a useful tool that often simplifies the verification of the countable additivity property.

Theorem 1. (σ -additivity \iff continuity) Let \mathcal{F} be a field of subsets of Ω , and suppose that $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ satisfies $\mathbb{P}(\Omega) = 1$ as well as the finite additivity property. Then, the following are equivalent:

- (a) \mathbb{P} is σ -additive on \mathcal{F} . In other words, if $\{A_j\}_{j=1}^{\infty}$ is a disjoint sequence of sets in \mathcal{F} , $A_j \in \mathcal{F}$ and $A = \bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$ then $\mathbb{P}(A) = \sum_{j=1}^{\infty} \mathbb{P}(A_j)$.
- (b) If $\{A_i\}$ is an increasing sequence of sets in \mathcal{F} (i.e., $A_i \subset A_{i+1}$, for all i), and $A = \bigcup_{i=1}^{\infty} A_i$ belongs to \mathcal{F} , then $\lim_{i \rightarrow \infty} \mathbb{P}(A_i) = \mathbb{P}(A)$.
- (c) If $\{A_i\}$ is a decreasing sequence of sets in \mathcal{F} (i.e., $A_i \supset A_{i+1}$, for all i), and $A = \bigcap_{i=1}^{\infty} A_i$ belongs to \mathcal{F} , then $\lim_{i \rightarrow \infty} \mathbb{P}(A_i) = \mathbb{P}(A)$.
- (d) If $\{A_i\}$ is a decreasing sequence of sets in \mathcal{F} (i.e., $A_i \supset A_{i+1}$, for all i) and $\bigcap_{i=1}^{\infty} A_i$ is empty, then $\lim_{i \rightarrow \infty} \mathbb{P}(A_i) = 0$.

Notes:

- If \mathcal{F} is also a σ -algebra then A in (a), (b) and (c) is automatically in \mathcal{F} .

- Theorem extends to general (non-probability) measures provided $\mathbb{P}(\Omega) < \infty$.
- **Notation:** If $\{A_i\}$ is a decreasing sequence of sets (i.e., $A_i \supset A_{i+1}$, for all i) and $\cap_{i=1}^{\infty} A_i = A$, we write $A_i \downarrow A$. Thus, in part (d) above, we are assuming that $A_i \downarrow \emptyset$.

Proof. We first assume that (a) holds and establish (b). Observe that $A = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \dots$, and that the events $A_1, (A_2 \setminus A_1), (A_3 \setminus A_2), \dots$ are disjoint (check this). Therefore, using countable additivity,

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A_1) + \sum_{i=2}^{\infty} \mathbb{P}(A_i \setminus A_{i-1}) \\ &= \mathbb{P}(A_1) + \lim_{n \rightarrow \infty} \sum_{i=2}^n \mathbb{P}(A_i \setminus A_{i-1}) \\ &= \mathbb{P}(A_1) + \lim_{n \rightarrow \infty} \sum_{i=2}^n (\mathbb{P}(A_i) - \mathbb{P}(A_{i-1})) \\ &= \mathbb{P}(A_1) + \lim_{n \rightarrow \infty} (\mathbb{P}(A_n) - \mathbb{P}(A_1)) \\ &= \lim_{n \rightarrow \infty} \mathbb{P}(A_n).\end{aligned}$$

Suppose now that property (b) holds, let A_i be a decreasing sequence of sets, and let $A = \cap_{i=1}^{\infty} A_i \in \mathcal{F}$. Then, the sequence A_i^c is increasing, and De Morgan's law, together with property (b) imply that

$$\mathbb{P}(A^c) = \mathbb{P}\left(\left(\cap_{i=1}^{\infty} A_i\right)^c\right) = \mathbb{P}\left(\cup_{i=1}^{\infty} A_i^c\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_i^c),$$

and

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \lim_{n \rightarrow \infty} \mathbb{P}(A_i^c) = \lim_{n \rightarrow \infty} (1 - \mathbb{P}(A_i^c)) = \lim_{n \rightarrow \infty} \mathbb{P}(A_i).$$

Property (d) follows from property (c), because (d) is just the special case of (c) in which the set A is empty.

To complete the proof, we now assume that property (d) holds and establish that property (a) holds as well. Let $B_i \in \mathcal{F}$ be disjoint events. Let $A_n = \cup_{i=n}^{\infty} B_i$. Note that $\{A_n\}$ is a decreasing sequence of events. We claim that $\cap_{n=1}^{\infty} A_n = \emptyset$. Indeed, if $\omega \in A_1$, then $\omega \in B_n$ for some n , which implies that $\omega \notin \cup_{i=n+1}^{\infty} B_i = A_{n+1}$. Therefore, no element of A_1 can belong to all of the sets A_n , which means that the intersection of the sets A_n is empty. Property (d) then implies that $\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 0$.

Applying finite additivity to the n disjoint sets $B_1, B_2, \dots, B_{n-1}, \cup_{i=n}^{\infty} B_i$, we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{n-1} \mathbb{P}(B_i) + \mathbb{P}\left(\bigcup_{i=n}^{\infty} B_i\right).$$

This equality holds for any n , and we can take the limit as $n \rightarrow \infty$. The first term on the right-hand side converges to $\sum_{i=1}^{\infty} \mathbb{P}(B_i)$. The second term is $\mathbb{P}(A_n)$, and as observed before, converges to zero. We conclude that

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i),$$

and property (a) holds. \square

6.1 Discrete probability spaces revisited

In Section 3, we defined $\mathbb{P}(A)$ for every $A \subset \Omega$ in terms of the probabilities of individual outcomes. We actually need to verify that this formula results in probabilities that satisfy countable additivity.

To this effect, we can use Theorem 1. We only need to verify (i) finite additivity and (ii) the continuity property in part (d).

Regarding finite additivity, it suffices to consider the case of two sets; the general case is obtained by induction on the number of sets. Suppose that the sets $A = \{\omega_1, \omega_2, \dots\}$ and $B = \{\omega'_1, \omega'_2, \dots\}$ are disjoint. Let $a_i = \mathbb{P}(\omega_i)$ and $b = \mathbb{P}(\omega'_i)$. We then have $A \cup B = \{\omega_1, \omega'_1, \omega_2, \omega'_2, \dots\}$ and

$$\mathbb{P}(A \cup B) = a_1 + b_1 + a_2 + b_2 + \dots = \sum_{i=1}^{\infty} (a_i + b_i) = \sum_{i=1}^{\infty} a_i + \sum_{i=1}^{\infty} b_i = \mathbb{P}(A) + \mathbb{P}(B).$$

The second and third equalities above are elementary properties of infinite series involving nonnegative numbers (more generally of absolutely convergent infinite series); namely, the order of summation or the grouping of the summands does not matter.

Regarding continuity, we need to show that

$$A_n \downarrow \emptyset \quad \Rightarrow \quad \mathbb{P}(A_n) \rightarrow 0.$$

Indeed, without loss of generality, we may assume $\Omega = \{1, 2, \dots\}$ is the set of natural numbers (to be denoted in this course by either \mathbb{N} or \mathbb{Z}_+). Fix some $\epsilon > 0$. Since $\sum_{i=1}^{\infty} \mathbb{P}(i) = \sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1$ is a convergent series, it follows that there exists some $m \in \mathbb{N}$ for which

$$\sum_{i \geq m} \mathbb{P}(i) \leq \epsilon.$$

On the other hand, since $A_n \downarrow \emptyset$, it follows that for every i , there exists some n_i such that $i \notin A_n$, for $n \geq n_i$. By using this property for $i = 1, \dots, m - 1$, we see that

$$A_n \subseteq \{m, m + 1, \dots\},$$

when n is large enough. Thus, for all large enough n ,

$$\mathbb{P}(A_n) \leq \mathbb{P}(\{m, m + 1, \dots\}) = \sum_{i \geq m} \mathbb{P}(i) \leq \epsilon.$$

It follows that $\lim_{n \rightarrow \infty} \mathbb{P}(A_n) \leq \epsilon$. Since ϵ can be an arbitrarily small positive number, we conclude that $\lim_{n \rightarrow \infty} \mathbb{P}(A_n) = 0$.

7 MONOTONE CLASS THEOREM

We will soon find that one often needs to prove that a certain collection of sets is a σ -algebra. Such verifications are facilitated by the following theorem.

Definition 5. A collection of sets \mathcal{M} is a **monotone class** if all increasing and decreasing sequences of sets from \mathcal{M} have limits belonging to \mathcal{M} . Formally, let $A_n \in \mathcal{M}$ for all n

$$\begin{aligned} A_n \nearrow A &\Rightarrow A \in \mathcal{M} \\ A_n \searrow A &\Rightarrow A \in \mathcal{M}. \end{aligned}$$

The minimal monotone class containing a collection \mathcal{C} is denoted $\mu(\mathcal{C})$.

Note that $\mu(\mathcal{C})$ is well-defined by an analog of Proposition 1 for intersections of monotone classes.

Theorem 2. If \mathcal{A} is an algebra (field) of sets, then

$$\mu(\mathcal{A}) = \sigma(\mathcal{A}).$$

Proof. First, note that any σ -algebra is necessarily a monotone class. Thus

$$\mu(\mathcal{A}) \subseteq \sigma(\mathcal{A}).$$

Second, any collection \mathcal{F} of sets which is simultaneously a monotone class and

an algebra is a σ -algebra. To see this, suppose $A_k \in \mathcal{F}$ and notice that

$$\bigcup_{k=1}^{\infty} A_k = \lim_{n \rightarrow \infty} B_n, \quad B_n \triangleq \bigcup_{k=1}^n A_k.$$

Then, $B_n \in \mathcal{F}$ because \mathcal{F} is an algebra and $\lim_n B_n \in \mathcal{F}$ because \mathcal{F} is a monotone class.

It remains to prove that $\mathcal{M} \triangleq \mu(\mathcal{A})$ is an algebra. To that end, define another collection of sets

$$\mathcal{L}_1 \triangleq \{A \in \mathcal{M} \mid A^c \in \mathcal{M}\}.$$

Note that clearly \mathcal{L}_1 contains \mathcal{A} . Furthermore, for any increasing sequence $E_n \nearrow E$ of subsets of \mathcal{L}_1 we have $E \in \mathcal{L}_1$ since

$$(E_n)^c \searrow E^c$$

and \mathcal{M} is a monotone class. Similarly, \mathcal{L}_1 is closed under decreasing limits. Thus \mathcal{L}_1 is a monotone class. By minimality of $\mathcal{M} = \mu(\mathcal{A})$ we conclude

$$\mathcal{L}_1 = \mathcal{M},$$

and hence \mathcal{M} is closed under taking complements.

Proceeding in the same way, fix a set $B \in \mathcal{A}$ and define

$$\mathcal{L}_B \triangleq \{A \in \mathcal{M} \mid A \cap B \in \mathcal{M}\}.$$

Clearly, \mathcal{L}_B contains \mathcal{A} and is a monotone class (since $A_n \nearrow A \Rightarrow A_n \cap B \nearrow A \cap B$). Again, $\mathcal{L}_B = \mathcal{M}$. Hence, as B was arbitrary, \mathcal{M} is closed under taking intersections and (by taking complements) unions with sets from \mathcal{A} .

Finally, let

$$\mathcal{L}_2 \triangleq \{A \in \mathcal{M} \mid A \cap M \in \mathcal{M} \text{ and } A \cup M \in \mathcal{M} \text{ for all sets } M \in \mathcal{M}\}.$$

As we have shown above, \mathcal{L}_2 contains \mathcal{A} . If $A_n \in \mathcal{L}_2$ and $A_n \nearrow A$, then for any M we have

$$A_n \cap M \in \mathcal{M}$$

by the definition of \mathcal{L}_2 and, on the other hand,

$$A_n \cap M \nearrow A \cap M \in \mathcal{M}$$

since \mathcal{M} is a monotone class. Thus $A \cap M \in \mathcal{M}$. Applying this argument to $(A_n^c \cap M^c) \searrow A^c \cap M^c$ and noticing that \mathcal{M} is closed under complements we obtain $A \cup M \in \mathcal{M}$. Hence $A \in \mathcal{L}_2$ and \mathcal{L}_2 is a monotone class implying

$$\mathcal{L}_2 = \mathcal{M}$$

or that \mathcal{M} is an algebra. \square

Remark (Caution: real analysis). *The importance of the monotone class theorem is that it allows one to avoid the use of transfinite induction when proving properties of σ -algebras. However, if you understand transfinite induction many of the tricky constructions involving monotone classes become much less mysterious. For example, constructing $\mu(\mathcal{A})$ involves taking \mathcal{A} , then adding all the limits of increasing and decreasing sets (thus forming new sets “tier 2”), then adding the limits of increasing and decreasing sets in tier 2 (forming “tier 3”), etc. Transfinite induction gives a rigorous sense to the definition, “let $\mu(\mathcal{A})$ be the first tier at which this procedure stabilizes”. Intuitively, then, $\mu(\mathcal{A})$ is closed under the operation of taking limits. Now if E is a set in any tier then $E \cap A$ is also a set in the same tier (assuming $A \in \mathcal{A}$). Consequently, $\mu(\mathcal{A})$ is automatically closed under intersections with sets from \mathcal{A} . Similarly, one may replace $A \in \mathcal{A}$ with any A in tier 2, 3, etc – eventually proving $\mu(\mathcal{A})$ is closed under intersections.*

References

- (a) E. Cinlar, Chapter I, Sections 1-4.
- (b) Grimmett and Stirzaker, Chapter 1, Sections 1.1-1.3.
- (c) Williams, Chapter 1, Sections 1.0-1.5, 1.9-1.10.
- (d) Florescu, Tudor: Chapter 1, Sections 2.1–2.3.5.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

TWO FUNDAMENTAL PROBABILISTIC MODELS

Contents

1. Extending measures from algebras to σ -algebras (Carathéodory)
2. Coin tosses: a “uniform” measure on $\{0, 1\}^\infty$
3. Lebesgue measure on $[0, 1]$ and \mathbb{R}
4. Completion of a probability space
5. Further remarks
6. Appendix: strange sets

The following are two fundamental probabilistic models that can serve as building blocks for more complex models:

- (a) A model of an infinite sequence of fair coin tosses that assigns equal probability, $1/2^n$, to every possible sequence of length n .
- (b) The **uniform distribution** on $[0, 1]$, which assigns probability $b - a$ to every interval $[a, b] \subset [0, 1]$.

These two models are often encountered in elementary probability and used without further discussion. Strictly speaking, however, we need to make sure that these two models are well-posed, that is, consistent with the axioms of probability. To this effect, we need to define appropriate σ -algebras and probability measures on the corresponding sample spaces. In what follows, we describe the required construction, while omitting the proofs of the more technical steps.

1 EXTENDING MEASURES FROM ALGEBRAS TO σ -ALGEBRAS

The general outline of the construction we will use is as follows. We are interested in defining a probability measure with certain properties on a given

measurable space (Ω, \mathcal{F}) . We consider a smaller collection, $\mathcal{F}_0 \subset \mathcal{F}$, of subsets of Ω , which is an algebra, and on which the desired probabilities are easy to define.¹ Furthermore, we make sure that \mathcal{F}_0 is rich enough, so that the σ -algebra it generates is the same as the desired σ -algebra \mathcal{F} . We then extend the definition of the probability measure from \mathcal{F}_0 to the entire σ -algebra \mathcal{F} . This is possible, under a few conditions, by virtue of the following fundamental result from measure theory due to Carathéodory.

We begin by asking the following question: Is probability measure on a σ -algebra $\mathcal{F} = \sigma(\mathcal{C})$ completely determined by its values on the generating collection \mathcal{C} ? The answer is no as the next exercise demonstrates.

Exercise 1. Let $\Omega = \{H, T\}^2$ (two coin tosses). Consider two probability measures: under \mathbb{P}_1 two fair coins are tossed independently, while under the second coin toss is just taken to be equal to the first. Let $\{\{HH, HT\}, \{HH, TH\}\}$. Show that $\sigma(\mathcal{C}) = 2^\Omega$ and that $\mathbb{P}_1 = \mathbb{P}_2$ on \mathcal{C} .

However, it turns out that probability measures coinciding on an *algebra* of sets must necessarily coincide on the σ -algebra generated by it.

Proposition 1. Let \mathcal{A} be an algebra of subsets of Ω and probability measures on $\mathcal{F} = \sigma(\mathcal{A})$. If λ and μ agree on \mathcal{A} then they agree on all \mathcal{F} .

Proof. Define

$$\mathcal{L} = \{A \in \mathcal{F} : \mu(A) = \lambda(A)\}.$$

Take a sequence of sets $A_n \in \mathcal{L}$ with $A_n \nearrow A$. By Theorem 1 from Lecture 1 (continuity of σ -additive measures) we have $\mu(A) = \lambda(A)$ and hence $A \in \mathcal{L}$. Same argument applies to decreasing sequences $A_n \searrow A$. Therefore \mathcal{L} is a monotone class, containing the algebra \mathcal{A} . By the monotone class theorem (Theorem 2 from Lecture 1) $\mathcal{L} = \sigma(\mathcal{A})$. \square

Remark. One frequently constructs σ -algebras from p -systems. A collection of subsets \mathcal{C} is called a **p -system** if it is closed under finite intersections. In the next lecture we will show that measures coinciding on a p -system also coincide on the algebra generated by it. By virtue of Proposition 1 they will then also coincide on the σ -algebra generated by the p -system. In short: it is sufficient to verify agreement of measures on any *generating p -system*.

Remark. Proposition remains true if we replace probability measures with finite or even σ -finite measures (A measure μ is called σ -finite if the set Ω can be partitioned into a countable union of sets, each of which has finite measure.).

¹An algebra (or a field) is a collection of subsets of the sample space that includes the empty set closed under taking complements and under *finite* unions.

“Wilder” measures, however, may violate the proposition: E.g. consider an algebra of finite unions $\bigcup_{i=1}^n (a_i, b_i]$ on $(0, 1]$. The σ -algebra generated by this algebra is the Borel one on $(0, 1]$. Let $\lambda(A)$ and $\mu(A)$ be equal to cardinality and twice the cardinality of A , respectively. The on the algebra of finite unions they coincide (giving infinite measure to any non-empty set), while being clearly different.

Theorem 1. (Carathéodory’s extension theorem) *Let \mathcal{F}_0 be an algebra of subsets of a sample space Ω , and let $\mathcal{F} = \sigma(\mathcal{F}_0)$ be the σ -algebra that it generates. Suppose that \mathbb{P}_0 is a mapping from \mathcal{F}_0 to $[0, 1]$ that satisfies $\mathbb{P}_0(\Omega) = 1$, as well as countable additivity on \mathcal{F}_0 .*

Then, \mathbb{P}_0 can be extended uniquely to a probability measure on (Ω, \mathcal{F}) . That is, there exists a unique probability measure \mathbb{P} on (Ω, \mathcal{F}) such that $\mathbb{P}(A) = \mathbb{P}_0(A)$ for all $A \in \mathcal{F}_0$.

Remarks:

- (a) The proof of the extension theorem is not too long; see, e.g., Appendix A of [Williams]. The key steps are:

- Define a σ -subadditive set-function $\lambda^* : 2^\Omega \rightarrow [0, 1]$ (called an outer measure)

$$\lambda^*(E) \triangleq \inf \left\{ \sum_{j=1}^{\infty} \mathbb{P}(A_j) : A_j \in \mathcal{F}, E \subset \bigcup_j A_j \right\}.$$

- Define a collection of sets

$$\bar{\mathcal{F}} \triangleq \{E : \lambda^*(F) = \lambda^*(F \cap E) + \lambda^*(F \cap E^c) \quad \forall F \subset \Omega\}.$$

- Show that $\bar{\mathcal{F}}$ is a σ -algebra containing \mathcal{F}_0 and that λ^* is a probability measure on it, which coincides with \mathbb{P} on \mathcal{F}_0 . Then restrict from $\bar{\mathcal{F}}$ to \mathcal{F} .

- (b) Although the extension theorem is a powerful result, the key step in constructing probability measures is verification of the countable additivity property of \mathbb{P}_0 on \mathcal{F}_0 . By Theorem 1 from Lecture 1, it suffices to verify that if $\{A_i\}$ is a decreasing sequence of sets in \mathcal{F}_0 and if $\bigcap_{i=1}^{\infty} A_i$ is empty, then $\lim_{n \rightarrow \infty} \mathbb{P}_0(A_i) = 0$. We will soon see how such a verification is done.

In the next two sections, we consider the two models of interest. We define appropriate algebras, define probabilities for the events in those algebras, and then use the extension theorem to obtain a probability measure.

2 COIN TOSSES: A “UNIFORM” MEASURE ON $\{0, 1\}^\infty$

Consider an infinite sequence of fair coin tosses. We wish to construct a probabilistic model of this experiment under which every possible sequence of results of the first n tosses has the same probability, $1/2^n$.

The sample space for this experiment is the set $\{0, 1\}^\infty$ of all infinite sequences $\omega = (\omega_1, \omega_2, \dots)$ of zeroes and ones (we use zeroes and ones instead of heads and tails).

First, we want to argue that it is not possible to define a good “uniform” measure on the collection of all subsets 2^Ω . This will justify the whole idea of introducing the concept of a σ -algebra. Let us try to understand what exactly we mean by “uniform”. Fix an infinite string $b \in \{0, 1\}^\infty$. Let us introduce the modulo-2 addition (XOR) as:

$$\omega \oplus b = (\omega_1 + b_1, \dots, \omega_n + b_n, \dots) \text{mod} 2,$$

and similarly for sets

$$A \oplus b = \{\omega \oplus b : \omega \in A\}.$$

Informally, a realization of coin tosses is in the set $A \oplus b$ iff it is in A after we invert every coordinate j for which $b_j = 1$. It is natural to require that our measure be such that

$$\mathbb{P}[A \oplus b] = \mathbb{P}[A]. \quad (1)$$

(In mathematical terms, we want \mathbb{P} to be translation invariant.)

Let us show that it is not possible to define a σ -additive \mathbb{P} on all of 2^Ω so that (1) holds. Indeed, suppose such a \mathbb{P} existed. Then define an equivalence relation on Ω : $\omega \sim \omega'$ if these binary sequences disagree in at most finitely many places. Let A be a set of representatives, one for each equivalence class; and let B be the equivalence class of the 0-sequence. It is clear that

$$\Omega = \bigcup_{\omega \in B} \omega \oplus A.$$

On the other hand since B is countable and the sets in the union are disjoint:

$$1 = \mathbb{P}[\Omega] = \sum_{\omega \in B} \mathbb{P}[\omega \oplus A] = \sum_{\omega \in B} \mathbb{P}[A],$$

which is impossible for any choice of $\mathbb{P}[A]$.

In conclusion, we showed that “uniform” (in the sense of (1)) probability measure on Ω must necessarily be defined on a σ -algebra that may not include A and hence be strictly smaller than 2^Ω . We construct such a σ -algebra next.

2.1 An algebra and a σ -algebra of subsets of $\{0, 1\}^n$

Let \mathcal{F}_n be the collection of events whose occurrence can be decided by looking at the results of the first n tosses. For example, the event $\{\omega \mid \omega_1 = 1 \text{ and } \omega_2 \neq \omega_4\}$ belongs to \mathcal{F}_4 (as well as to \mathcal{F}_k for every $k \geq 4$).

Let B be an arbitrary (possibly empty) subset of $\{0, 1\}^n$. Consider the set

$$A = \{\omega \in \{0, 1\}^\infty \mid (\omega_1, \omega_2, \dots, \omega_n) \in B\}.$$

We can express $A \subset \{0, 1\}^\infty$ in the form $A = B \times \{0, 1\}^\infty$. This is simply saying that any sequence in A can be viewed as a pair consisting of a n -long sequence that belongs to B , followed by an arbitrary infinite sequence. The event A belongs to \mathcal{F}_n , and all elements of \mathcal{F}_n are of this form, for some A . It is easily verified that \mathcal{F}_n is a σ -algebra.

Exercise 2. Provide a formal proof that \mathcal{F}_n is a σ -algebra.

The σ -algebra \mathcal{F}_n , for any fixed n , is too small; it can only serve to model the first n coin tosses. We are interested instead in sets that belong to \mathcal{F}_n , for arbitrary n , and this leads us to our main definition:

$$\mathcal{F}_0 = \bigcup_{n=1}^{\infty} \mathcal{F}_n = \{A : \omega \in A \iff (\omega_1, \dots, \omega_n) \in B, n \geq 0, B \in \{0, 1\}^n\},$$

i.e. \mathcal{F}_0 is the collection of all those sets A for which membership $\omega \stackrel{?}{\in} A$ can be determined on the basis of inspecting only finitely many coordinates $(\omega_1, \dots, \omega_n)$ for some $n \geq 0$.²

Example. Let $A_n = \{\omega \mid \omega_n = 1\}$, the event that the n th toss results in a “1”. Note that $A_n \in \mathcal{F}_n$. Let $A = \bigcup_{i=1}^{\infty} A_n$, which is the event that there is at least one “1” in the infinite toss sequence. The event A does not belong to \mathcal{F}_n , for any n . (Intuitively, having observed a sequence of n zeroes does not allow us to decide whether there will be a subsequent “1” or not.) Consider also the complement of A , which is the event that the outcome of the experiment is an infinite string of zeroes. Once more, we see that A^c does not belong to \mathcal{F}_0 .

The preceding example shows that \mathcal{F}_0 is not a σ -algebra. On the other hand, it can be verified that \mathcal{F}_0 is an algebra.

Exercise 3. Prove that \mathcal{F}_0 is an algebra.

²**Warning:** the union $\bigcup_{i=1}^{\infty} \mathcal{F}_i = \mathcal{F}_0$ is not the same as the collection of sets of the form $\bigcup_{i=1}^{\infty} A_i$, for $A_i \in \mathcal{F}_i$. For an illustration, if $\mathcal{F}_1 = \{\{a\}, \{b, c\}\}$ and $\mathcal{F}_2 = \{\{d\}\}$, then $\mathcal{F}_1 \cup \mathcal{F}_2 = \{\{a\}, \{b, c\}, \{d\}\}$. Note that $\{b, c\} \cup \{d\} = \{b, c, d\}$ is not in $\mathcal{F}_1 \cup \mathcal{F}_2$.

We would like to have a probability model that assigns probabilities to all of the events in \mathcal{F}_n , for every n . This means that we need a σ -algebra that includes \mathcal{F}_0 . On the other hand, we would like our σ -algebra to be as small as possible, i.e., contain as few subsets of $\{0, 1\}^n$ as possible, to minimize the possibility that it includes pathological sets to which probabilities cannot be assigned. This leads us to define \mathcal{F} as the sigma-algebra $\sigma(\mathcal{F}_0)$ generated by \mathcal{F}_0 .

2.2 A probability measure on $(\{0, 1\}^\infty, \mathcal{F})$

We start by defining a finitely additive function \mathbb{P}_0 on the algebra \mathcal{F}_0 that also satisfies $\mathbb{P}_0(\{0, 1\}^\infty) = 1$. This is accomplished as follows. Every set A in \mathcal{F}_0 is of the form $B \times \{0, 1\}^\infty$, for some n and some $B \subset \{0, 1\}^n$. We then let $\mathbb{P}_0(A) = |B|/2^n$.³ Note that the event $\{\omega_1, \omega_2, \dots, \omega_n\} \times \{0, 1\}^\infty$, which is the event that the first n tosses resulted in a particular sequence $\{\omega_1, \omega_2, \dots, \omega_n\}$, is assigned probability $1/2^n$. In particular, all possible sequences of length n are assigned equal probability, as desired.

Before proceeding further, we need to verify that the above definition is *consistent*. Note that same set A can belong to \mathcal{F}_n for several values of n . We therefore need to check that when we apply the definition of $\mathbb{P}_0(A)$ for different choices of n , we obtain the same value. Indeed, suppose that $A \in \mathcal{F}_m$, which implies that $A \in \mathcal{F}_n$, for $n > m$. In this case, $A = B \times \{0, 1\}^\infty = C \times \{0, 1\}^\infty$, where $B \subset \{0, 1\}^n$ and $C \subset \{0, 1\}^m$. Thus, $B = C \times \{0, 1\}^{n-m}$, and $|B| = |C| \cdot 2^{n-m}$. One application of the definition yields $\mathbb{P}_0(A) = |B|/2^n$, and another yields $\mathbb{P}_0(A) = |C|/2^m$. Since $|B| = |C| \cdot 2^{n-m}$, they both yield the same value.

It is easily verified that $\mathbb{P}_0(\Omega) = 1$, and that \mathbb{P}_0 is finitely additive: if $A_1, A_2 \subset \mathcal{F}_n$ are disjoint, then $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$.

It turns out that \mathbb{P}_0 is also countably additive on \mathcal{F}_0 .

Lemma 1. \mathbb{P}_0 is σ -additive on \mathcal{F}_0

Proof. According to Theorem 1 of Lecture 1 it is sufficient to show that

$$A_n \searrow \emptyset \quad \Rightarrow \quad \mathbb{P}(A_n) \rightarrow 0.$$

In fact, we will show that

$$A_n \searrow \emptyset \quad \Rightarrow \quad \exists N \geq 1 \forall n \geq N : A_n = \emptyset. \quad (2)$$

³For any set A , $|A|$ denotes its cardinality, the number of elements it contains.

Let us call a $b \in \{0, 1\}^m$ a “great prefix” if there exists infinitely many $n \geq 1$ with the property that A_n contains some ω with $\omega_i = b_i$ for all $i = 1, \dots, m$.

Note that $A_N = \emptyset$ for some $N \geq 1$ is equivalent to stating that $b = \emptyset$ (zero-length prefix) is not great. So assume, to arrive at a contradiction, that $b = \emptyset$ is great. Notice that if b is a great prefix than either $b0$ or $b1$ (juxtaposition) must be great too. Indeed, we can split the infinitely many A_n ’s which contain ω with b as prefix in two groups depending on whether ω_{m+1} equals zero or one. One of these groups must be infinite.

In other words, any great prefix can be extended by one more digit. In this way by induction we can construct an infinite sequence $b = (b_1, \dots, b_n, \dots)$ with the property that any initial segment if it is a great prefix. Now we show that $b \in A_m$ for any m . Indeed, by definition of \mathcal{F}_0 there must exist n such that $A_m \in \mathcal{F}_n$. Consider prefix (b_1, \dots, b_n) . It is great by construction and thus there must exist $\omega \in A_\ell$ such that $\omega_i = b_i$ for all $i \geq \ell$. In fact there are infinitely many such ℓ ’s and so in particular there is an $\ell \geq m$. Hence, from $A_\ell \subset A_m$ we conclude $\omega \in A_m$. But now recall that $A_m \in \mathcal{F}_n$ and since ω and b share the initial n values, we must also have $b \in A_m$. In all, we have shown $b \in A_m$ for all m and thus $\cap A_m \neq \emptyset$ contradicting the assumption.

(Here is a “fancy” proof for analysts: The space $\{0, 1\}^\infty$ is compact in the product topology (Tikhonov’s theorem) and thus A_n is a decreasing sequence of compacts.) □

We can now invoke the Extension Theorem and conclude that there exists a unique probability measure on \mathcal{F} , the σ -algebra generated by \mathcal{F}_0 , that agrees with \mathbb{P}_0 on \mathcal{F}_0 . This probability measure assigns equal probability, $1/2^n$, to every possible sequence of length n , as desired. This confirms that the intuitive process of an infinite sequence of coin flips can be captured rigorously within the framework of probability theory.

Exercise 4. Consider the probability space $(\{0, 1\}^\infty, \mathcal{F}, \mathbb{P})$. Let A be the set of all infinite sequences ω for which $\omega_n = 0$ for every odd n .

- (a) Establish that $A \notin \mathcal{F}_0$, but $A \in \mathcal{F}$.
- (b) Compute $\mathbb{P}(A)$.

Exercise 5. Show that \mathbb{P} is translation invariant (1) for all ω and $A \in \mathcal{F}$. (Hint: the monotone class theorem may be helpful.)

3 LEBESGUE MEASURE ON $[0, 1]$ AND ON \mathbb{R}

In this section, we construct the **uniform** probability measure on $[0, 1]$, also known as **Lebesgue** measure. Under the Lebesgue measure, the measure as-

sized to any subset of $[0, 1]$ is meant to be equal to its length. While the definition of length is immediate for simple sets (e.g., the set $[a, b]$ has length $b - a$), more general sets present more of a challenge.

We start by considering the sample space $\Omega = (0, 1]$, which is slightly more convenient than the sample space $[0, 1]$, though in the end it will result in essentially the same probability space. Similarly to the case of coin-tosses, a translation-invariant probability measure defined on all subsets of Ω does not exist, see Section 6. Thus, our first goal is to define a proper σ -algebra.

3.1 A σ -algebra and an algebra of subsets of $(0, 1]$

Consider the collection \mathcal{C} of all intervals $[a, b]$ contained in $(0, 1]$, and let \mathcal{F} be the σ -algebra generated by \mathcal{C} . As mentioned in the Lecture 1 notes, this is called the **Borel** σ -algebra, and is denoted by \mathcal{B} . Sets in this σ -algebra are called **Borel sets** or **Borel measurable sets**.

Any set that can be formed by starting with intervals $[a, b]$ and using a countable number of set-theoretic operations (taking complements, or forming countable unions and intersections of previously formed sets) is a Borel set. For example, it can be verified that single-element sets, $\{a\}$, are Borel sets. Furthermore, intervals $(a, b]$ are also Borel sets since they are of the form $[a, b] \setminus \{a\}$. Every countable set is also a Borel set, since it is the union of countably many single-element sets. In particular, the set of rational numbers in $(0, 1]$, as well as its complement, the set of irrational numbers in $(0, 1]$, is a Borel set. While Borel sets can be fairly complicated, not every set is a Borel set; see Sections 5-6.

As usual, directly defining a probability measure for all Borel sets is difficult, so we start by considering a smaller collection, \mathcal{F}_0 , of subsets of $(0, 1]$. We let \mathcal{F}_0 consist of the empty set and all sets that are finite unions of intervals of the form $(a, b]$. In more detail, if a set $A \in \mathcal{F}_0$ is nonempty, it is of the form

$$A = (a_1, b_1] \cup \cdots \cup (a_n, b_n],$$

where $0 \leq a_1 < b_1 \leq a_2 < b_2 \leq \cdots \leq a_n < b_n \leq 1$ and $n \in \mathbb{N}$.

Lemma 1. We have $\sigma(\mathcal{F}_0) = \sigma(\mathcal{C}) = \mathcal{B}$.

Proof. We have already argued that every interval of the form $(a, b]$ is a Borel set. Hence, a typical element of \mathcal{F}_0 (a finite union of such intervals) is also a Borel set. Therefore, $\mathcal{F}_0 \subset \mathcal{B}$, which implies that $\sigma(\mathcal{F}_0) \subset \sigma(\mathcal{B}) = \mathcal{B}$. (The last equality holds because \mathcal{B} is already a σ -algebra and is therefore equal to the smallest σ -algebra that contains \mathcal{B} .)

Consider $0 < a < b \leq 1$ and take a sequence of rationals $a_n \nearrow a$ and $b_n \searrow b$. Then

$$[a, b] = \cap_{n=1}^{\infty} (a_n, b_n]$$

Since $(a_n, b_n] \in \mathcal{F}_0$ it follows that $[a, b] \in \sigma(\mathcal{F}_0)$. Thus, $\mathcal{C} \subset \sigma(\mathcal{F}_0)$, which implies that

$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma(\sigma(\mathcal{F}_0)) = \sigma(\mathcal{F}_0) \subset \mathcal{B}.$$

(The second equality holds because the smallest σ -algebra containing $\sigma(\mathcal{F}_0)$ is $\sigma(\mathcal{F}_0)$ itself.) The first equality in the statement of the proposition follows. Finally, the equality $\sigma(\mathcal{C}) = \mathcal{B}$ is just the definition of \mathcal{B} . \square

Lemma 2.

- (a) *The collection \mathcal{F}_0 is an algebra.*
- (b) *The collection \mathcal{F}_0 is not a σ -algebra.*

Proof.

- (a) By definition, $\emptyset \in \mathcal{F}_0$. Note that $\emptyset^c = (0, 1] \in \mathcal{F}_0$. More generally, if A is of the form $A = (a_1, b_1] \cup \dots \cup (a_n, b_n]$, its complement is $(0, a_1] \cup (b_1, a_2] \cup \dots \cup (b_n, 1]$, which is also in \mathcal{F}_0 . Furthermore, the union of two sets that are unions of finitely many intervals of the form $(a, b]$ is also a union of finitely many such intervals. For example, if $A = (1/8, 2/8] \cup (4/8, 7/8]$ and $B = (3/8, 5/8]$, then $A \cup B = (1/8, 2/8] \cup (3/8, 7/8]$.
- (b) To see that \mathcal{F}_0 is not a σ -algebra, note that $(0, 1 - 2^{-n}] \in \mathcal{F}_0$, for every $n \in \mathbb{N}$, but the union of these sets, which is $(0, 1)$, does not belong to \mathcal{F}_0 . \square

3.2 The uniform measure on $(0, 1]$

For every $A \in \mathcal{F}_0$ of the form

$$A = (a_1, b_1] \cup \dots \cup (a_n, b_n],$$

we define

$$\mathbb{P}_0(A) = (b_1 - a_1) + \dots + (b_n - a_n),$$

which is its total length. Note that $\mathbb{P}_0(\) = \mathbb{P}((0, 1]) = 1$. Also \mathbb{P}_0 is *finitely additive*. Indeed if A_1, \dots, A_n are disjoint finite unions of intervals of the form $(a, b]$, then $A = \cup_{1 \leq i \leq n} A_i$ is also a finite union of such intervals and its total length is the sum of the lengths of the sets A_i .

Lemma 2. \mathbb{P}_0 is σ -additive on \mathcal{F}_0

Proof. (Optional) First notice the following: For $A \in \mathcal{F}_0$ and any $\epsilon > 0$ there exists a closed subset $C \subset A$ such that

$$A = C \cup E,$$

where $E \in \mathcal{F}_0$ and $\mathbb{P}_0(E) \leq \epsilon$. For the basic interval $(a, b]$ this follows by writing

$$(a, b] = [a - \epsilon, b] \cup (a, a - \epsilon]$$

and for other sets in \mathcal{F}_0 similarly.

Next consider $A_n \searrow \emptyset$ with $A_n \in \mathcal{F}_0$. Fix arbitrary $\epsilon_0 > 0$ and let $\epsilon_n = \epsilon_0 2^{-n}$. Select decompositions as above

$$A_n = C_n \cup E_n, \quad \mathbb{P}_0(E_n) \leq \epsilon_n.$$

Sets C_n are not necessarily nested, so define $F_n = \bigcap_{k=1}^n C_k$ and notice that

$$A_n \setminus F_n \subseteq \bigcup_{k=1}^n E_k, \tag{3}$$

which is shown by induction. Suppose that all F_n are non-empty and select in each F_n an arbitrary element x_n . Note that $x_n \in F_m$ for all $m \leq n$. By compactness of F_1 the sequence x_n must contain a subsequence x_{n_i} converging to some point x^* . By the preceding observation for every fixed m and all sufficiently large i we have $x_{n_i} \in F_m$, and thus $x^* \in F_m$ for all m . This contradicts the fact that $\bigcap F_m = \emptyset$. We conclude that there must exist some N such that $F_N = \emptyset$.

Consequently, from (3) we get that

$$A_N \subseteq \bigcup_{k=1}^N E_k$$

implying by the union bound that

$$\mathbb{P}_0(A_N) \leq \sum_{k=1}^N \mathbb{P}_0(E_k) = \epsilon_0 \sum_{k=1}^N 2^{-k} \leq \epsilon_0$$

Thus, for any $\epsilon_0 > 0$ we have

$$\lim_{n \rightarrow \infty} \mathbb{P}_0(A_n) \leq \epsilon_0$$

implying the limit is actually zero. \square

We can now apply the Extension Theorem and conclude that there exists a probability measure \mathbb{P} , called the **Lebesgue** or **uniform** measure, defined on the entire Borel σ -algebra \mathcal{B} , that agrees with \mathbb{P}_0 on \mathcal{F}_0 . In particular, $\mathbb{P}((a, b]) = b - a$, for every interval $(a, b] \subset (0, 1]$.

By augmenting the sample space Ω to include 0, and assigning zero probability to it, we obtain a new probability model with sample space $\Omega = [0, 1]$. (Exercise: define formally the sigma-algebra on $[0, 1]$, starting from the σ -algebra on $(0, 1]$.)

Exercise 6. Let A be the set of irrational numbers in $[0, 1]$. Show that $\mathbb{P}(A) = 1$.

Example. Let A be the set of points in $[0, 1]$ whose decimal representation contains only odd digits. (We disallow decimal representations that end with an infinite string of nines. Under this condition, every number has a unique decimal representation.) What is the Lebesgue measure of this set?

Observe that $A = \bigcap_{n=1}^{\infty} A_n$, where A_n is the set of points whose first n digits are odd. It can be checked that A_n is a disjoint union of 5^n intervals, each with length $1/10^n$. Thus, $\mathbb{P}(A_n) = 5^n/10^n = 1/2^n$. Since $A \subset A_n$, we obtain $\mathbb{P}(A) \leq \mathbb{P}(A_n) = 1/2^n$. Since this is true for every n , we conclude that $\mathbb{P}(A) = 0$.

Exercise 7. Let A be the set of points in $[0, 1]$ whose decimal representation contains at least one digit equal to 9. Find the Lebesgue measure of that set.

Note that there is nothing special about the interval $(0, 1]$. For example, if we let $\Omega = (c, d]$, where $c < d$, and if $(a, b] \subset (c, d]$, we can define $\mathbb{P}_0((a, b]) = (b - a)/(d - c)$ and proceed as above to obtain a uniform probability measure on the set $(c, d]$, as well as on the set $[c, d]$.

On the other hand, a “uniform” probability measure on the entire real line, \mathbb{R} , that assigns equal probability to intervals of equal length, is incompatible with the requirement $\mathbb{P}(\Omega) = 1$. What we obtain instead, in the next section, is a notion of length which becomes infinite for certain sets.

3.3 The Lebesgue measure on \mathbb{R}

Let $\Omega = \mathbb{R}$. We first define a σ -algebra of subsets of \mathbb{R} . This can be done in several ways. It can be verified that the three alternatives below are equivalent.

- (a) Let \mathcal{C} be the collection of all intervals of the form $[a, b]$, and let $\mathcal{B} = \sigma(\mathcal{C})$ be the σ -algebra that it generates.
- (b) Let \mathcal{D} be the collection of all intervals of the form $(-\infty, b]$, and let $\mathcal{B} = \sigma(\mathcal{D})$ be the σ -algebra that it generates.

- (c) For any n , we define the Borel σ -algebra of $(n, n + 1]$ as the σ -algebra generated by sets of the form $[a, b] \subset (n, n + 1]$. We then say that A is a Borel subset of \mathbb{R} if $A \cap (n, n + 1]$ is a Borel subset of $(n, n + 1]$, for every n .

Exercise 8. Show that the above three definitions of \mathcal{B} are equivalent.

Let \mathbb{P}_n be the uniform measure on $(n, n + 1]$ (defined on the Borel sets in $(n, n + 1]$). Given a set $A \subset \mathbb{R}$, we decompose it into countably many pieces, each piece contained in some interval $(n, n + 1]$, and define its “length” $\mu(A)$ using countable additivity:

$$\mu(A) = \sum_{n=-\infty}^{\infty} \mathbb{P}_n(A \cap (n, n + 1]).$$

It turns out that μ is a measure on $(\mathbb{R}, \mathcal{B})$, called again **Lebesgue measure**. However, it is not a probability measure because $\mu(\mathbb{R}) = \infty$.

Exercise 9. Show that μ is a measure on $(\mathbb{R}, \mathcal{B})$. Hint: Use the countable additivity of the measures \mathbb{P}_n to establish the countable additivity of μ . You can also use the fact that if the numbers a_{ij} are nonnegative, then $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}$.

Similar to the case of $\{0, 1\}^{\infty}$, there exist subsets of $[0, 1]$ that are not Borel sets. In fact the similarities between the models of Sections 2 and 3 are much deeper; the two models are essentially equivalent, although we will not elaborate on the meaning of this. Let us only say that the equivalence relies on the one-to-one correspondence of the sets $[0, 1]$ and $\{0, 1\}^{\infty}$ obtained through the binary representation of real numbers. Intuitively, generating a real number at random, according to the uniform distribution (Lebesgue measure) on $[0, 1]$, is probabilistically equivalent to generating each bit in its binary expansion at random.

4 COMPLETION OF A PROBABILITY SPACE

Starting with an algebra \mathcal{F}_0 and a countably additive function \mathbb{P}_0 on that algebra, the Extension Theorem leads to a probability measure on the smallest σ -algebra containing \mathcal{F}_0 . Can we extend the measure further, to a larger σ -algebra? If so, is the extension unique, or will there have to be some arbitrary choices? We describe here a generic extension that assigns probabilities to certain additional sets A for which there is little choice.

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $B \in \mathcal{F}$, and $\mathbb{P}(B) = 0$. Any set B with this property is called a **null** set. (Note that in this context,

“null” is not the same as “empty.”) Suppose now that $A \subset B$. If the set A is not in \mathcal{F} , it is not assigned a probability; were it to be assigned one, the only choice that would not lead to a contradiction is a value of zero.

The first step is to augment the σ -algebra \mathcal{F} so that it includes all subsets of null sets. This is accomplished as follows:

- (a) Let \mathcal{N} be the collection of all subsets of null sets;
- (b) Define $\mathcal{F}^* = \sigma(\mathcal{F} \cup \mathcal{N})$, the smallest σ -algebra that contains \mathcal{F} as well as all subsets of null sets.
- (c) Extend \mathbb{P} in some natural manner to obtain a new probability measure \mathbb{P}^* on (Ω, \mathcal{F}^*) . In particular, we let $\mathbb{P}^*(A) = 0$ for every subset $A \subset B$ of every null set $B \in \mathcal{F}$. It turns out that such an extension is always possible and unique.

Details of this construction will be worked out in an exercise.

The resulting probability space is said to be **complete**. It has the property that all subsets of null sets are included in the σ -algebra and are also null sets.

When $\Omega = [0, 1]$ (or $\Omega = \mathbb{R}$), \mathcal{F} is the Borel σ -algebra, and \mathbb{P} is Lebesgue measure, we obtain an augmented σ -algebra \mathcal{F}^* and a measure \mathbb{P}^* . The sets in \mathcal{F}^* are called **Lebesgue measurable** sets. The new measure \mathbb{P}^* is referred to by the same name as the measure \mathbb{P} (“Lebesgue measure”).

5 FURTHER REMARKS

We record here a few interesting facts related to Borel σ -algebras and the Lebesgue measure. Their proofs tend to be fairly involved.

- (a) There exist sets that are Lebesgue measurable but not Borel measurable, i.e., \mathcal{F} is a proper subset of \mathcal{F}^* .
- (b) There are as many Borel measurable sets as there are points on the real line (this is the “cardinality of the continuum”), but there are as many Lebesgue measurable sets as there are subsets of the real line (which is a higher cardinality) [Billingsley]

Note: In Lecture 4 we will introduce a Cantor set, which has cardinality of the continuum, while being of measure 0. Since all subsets of a Cantor set (being null-sets) are measurable, it is clear that $|\mathcal{F}^*| \geq |2^{\mathbb{R}}|$. Showing that $|\mathcal{F}| = |\mathbb{R}|$ is a lot more involved. Note that this difference in cardinalities automatically implies there is a “wealth” of Lebesgue-measurable sets which are not Borel.

- (c) There exist subsets of $[0, 1]$ that are not Lebesgue measurable; see Section 6 below and [Williams, p. 192].
- (d) It is not possible to construct a probability space in which the σ -algebra includes all subsets of $[0, 1]$, with the property that $\mathbb{P}(\{x\}) = 0$ for every $x \in (0, 1]$ [Billingsley, pp. 45-46].

6 APPENDIX: ON STRANGE SETS (optional reading)

In this appendix, we provide some evidence that not every subset of $(0, 1]$ is Lebesgue measurable, and, furthermore, that Lebesgue measure cannot be extended to a measure defined for all subsets of $(0, 1]$.

Let “+” stand for addition modulo 1 in $(0, 1]$. For example, $0.5 + 0.7 = 0.2$, instead of 1.2. You may want to visualize $(0, 1]$ as a circle that wraps around so that after 1, one starts again at 0. If $A \subset (0, 1]$, and x is a number, then $A + x$ stands for the set of all numbers of the form $y + x$ where $y \in A$.

Define x and y to be *equivalent* if $x + r = y$ for some rational number r . Then, $(0, 1]$ can be partitioned into equivalence classes. (That is, all elements in the same equivalence class are equivalent, elements belonging to different equivalence classes are not equivalent, and every $x \in (0, 1]$ belongs to exactly one equivalence class.) Let us pick exactly one element from each equivalence class, and let H be the set of the elements picked this way. (This fact that a set H can be legitimately formed this way involves the Axiom of Choice, a generally accepted axiom of set theory.) We will now consider the sets of the form $H + r$, where r ranges over the rational numbers in $(0, 1]$. Note that there are countably many such sets.

The sets $H + r$ are disjoint. (Indeed, if $r_1 \neq r_2$, and if the two sets $H + r_1$, $H + r_2$ share the point $h_1 + r_1 = h_2 + r_2$, with $h_1, h_2 \in H$, then h_1 and h_2 differ by a rational number and are equivalent. If $h_1 \neq h_2$, this contradicts the construction of H , which contains exactly one element from each equivalence class. If $h_1 = h_2$, then $r_1 = r_2$, which is again a contradiction.) Therefore, $(0, 1]$ is the union of the countably many disjoint sets $H + r$.

The sets $H + r$, for different r , are “translations” of each other (they are all formed by starting from the set H and adding a number, modulo 1). Let us say that a measure is *translation-invariant* if it has the following property: if A and $A + x$ are measurable sets, then $\mathbb{P}(A) = \mathbb{P}(A + x)$. Suppose that \mathbb{P} is a translation invariant probability measure, defined on all subsets of $(0, 1]$. Then,

$$1 = \mathbb{P}((0, 1]) = \sum_r \mathbb{P}(H + r) = \sum_r \mathbb{P}(H),$$

where the sum is taken over all rational numbers in $(0, 1]$. But this impossible. We conclude that a translation-invariant measure, defined on all subsets of $(0, 1]$ does not exist.

On the other hand, it can be verified that the Lebesgue measure is translation-invariant on the Borel σ -algebra, as well as its extension, the Lebesgue σ -algebra. This implies that the Lebesgue σ -algebra does not include all subsets of $(0, 1]$.

An even stronger, and more counterintuitive example is the following. It indicates, that the ordinary notion of area or volume cannot be applied to arbitrary sets.

The Banach-Tarski Paradox. Let S be the two-dimensional surface of the unit sphere in three dimensions. There exists a subset F of S such that for any $k \geq 3$,

$$S = (\tau_1 F) \cup \dots \cup (\tau_k F),$$

where each τ_i is a rigid rotation and the sets $\tau_i F$ are disjoint. For example, S can be made up by three rotated copies of F (suggesting probability equal to $1/3$, but also by four rotated copies of F , suggesting probability equal to $1/4$). Ordinary geometric intuition clearly fails when dealing with arbitrary sets.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

CONDITIONING AND INDEPENDENCE

Most of the material in this lecture is covered in [Bertsekas & Tsitsiklis] Sections 1.3-1.5 and Problem 48 (or problem 43, in the 1st edition), available at <http://athenasc.com/Prob-2nd-Ch1.pdf>. Solutions to the end of chapter problems are available at: http://athenasc.com/prob-solved_2ndedition.pdf. These lecture notes provide some additional details and twists.

Contents

1. Conditional probability
2. Independence
3. The Borel-Cantelli lemma

1 CONDITIONAL PROBABILITY

Definition 1. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and an event $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. For every event $A \in \mathcal{F}$, the conditional probability that A occurs given that B occurs is denoted by $\mathbb{P}(A | B)$ and is defined by

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Theorem 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

- (a) If B is an event with $\mathbb{P}(B) > 0$, then $\mathbb{P}(\Omega | B) = 1$, and for any sequence $\{A_i\}$ of disjoint events, we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i | B\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i | B).$$

As a result, suppose $\mathbb{P}_B : \mathcal{F} \rightarrow [0, 1]$ is defined by $\mathbb{P}_B(A) = \mathbb{P}(A | B)$. Then, \mathbb{P}_B is a probability measure on (Ω, \mathcal{F}) .

- (b) Let A be an event. If the events B_i , $i \in \mathbb{N}$, form a partition of Ω , and $\mathbb{P}(B_i) > 0$ for every i , then

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A | B_i) \mathbb{P}(B_i).$$

In particular, if B is an event with $\mathbb{P}(B) > 0$ and $\mathbb{P}(B^c) > 0$, then

$$\mathbb{P}(A) = \mathbb{P}(A | B)\mathbb{P}(B) + \mathbb{P}(A | B^c)\mathbb{P}(B^c).$$

- (c) **(Bayes' rule)** Let A be an event with $\mathbb{P}(A) > 0$. If the events B_i , $i \in \mathbb{N}$, form a partition of Ω , and $\mathbb{P}(B_i) > 0$ for every i , then for every i

$$\mathbb{P}(B_i | A) = \frac{\mathbb{P}(B_i)\mathbb{P}(A | B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B_i)\mathbb{P}(A | B_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B_j)\mathbb{P}(A | B_j)}.$$

- (d) For any sequence $\{A_i\}$ of events, we have

$$\mathbb{P}(\bigcap_{i=1}^{\infty} A_i) = \mathbb{P}(A_1) \prod_{i=2}^{\infty} \mathbb{P}(A_i | A_1 \cap \dots \cap A_{i-1}),$$

as long as all conditional probabilities are well defined.

Proof.

- (a) We have $\mathbb{P}(\Omega | B) = \mathbb{P}(\Omega \cap B) / \mathbb{P}(B) = \mathbb{P}(B) / \mathbb{P}(B) = 1$. Also

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i | B\right) = \frac{\mathbb{P}(B \cap (\bigcup_{i=1}^{\infty} A_i))}{\mathbb{P}(B)} = \frac{\mathbb{P}(\bigcup_{i=1}^{\infty} (B \cap A_i))}{\mathbb{P}(B)}.$$

Since the sets $B \cap A_i$, $i \in \mathbb{N}$ are disjoint, countable additivity, applied to the

right-hand side, yields

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i | B) = \frac{\sum_{i=1}^{\infty} \mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i | B),$$

as claimed.

(b) We have

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap \Omega) = \mathbb{P}\left(A \cap (\cup_{i=1}^{\infty} B_i)\right) = \mathbb{P}\left(\cup_{i=1}^{\infty} (A \cap B_i)\right) \\ &= \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A | B_i) \mathbb{P}(B_i).\end{aligned}$$

In the second equality, we used the fact that the sets B_i form a partition of Ω . In the next to last equality, we used the fact that the sets B_i are disjoint and the countable additivity property.

(c) This follows from the fact

$$\mathbb{P}(B_i | A) = \mathbb{P}(B_i \cap A) / \mathbb{P}(A) = \mathbb{P}(A | B_i) \mathbb{P}(B_i) / \mathbb{P}(A),$$

and the result from part (c).

(d) Note that the sequence of events $\cap_{i=1}^n A_i$ is decreasing and converges to $\cap_{i=1}^{\infty} A_i$. By the continuity property of probability measures, we have $\mathbb{P}(\cap_{i=1}^{\infty} A_i) = \lim_{n \rightarrow \infty} \mathbb{P}(\cap_{i=1}^n A_i)$. Note that

$$\begin{aligned}\mathbb{P}(\cap_{i=1}^n A_i) &= \mathbb{P}(A_1) \cdot \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} \cdot \frac{\mathbb{P}(A_1 \cap A_2 \cap A_3)}{\mathbb{P}(A_1 \cap A_2)} \cdots \frac{\mathbb{P}(A_1 \cap \cdots \cap A_n)}{\mathbb{P}(A_1 \cap \cdots \cap A_{n-1})} \\ &= \mathbb{P}(A_1) \prod_{i=2}^n \mathbb{P}(A_i | A_1 \cap \cdots \cap A_{i-1}).\end{aligned}$$

Taking the limit, as $n \rightarrow \infty$, we obtain the claimed result. \square

2 INDEPENDENCE

Intuitively we call two events A, B independent if the occurrence or nonoccurrence of one does not affect the probability assigned to the other. The following definition formalizes and generalizes the notion of independence.

Definition 2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

- (a) Two events, A and B , are said to be **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
 Notation: $A \perp\!\!\!\perp B$. If $\mathbb{P}(B) > 0$, an equivalent condition is $\mathbb{P}(A) = \mathbb{P}(A | B)$.

- (b) Let S be an index set (possibly infinite, or even uncountable), and let $\{A_s \mid s \in S\}$ be a family (set) of events. The events in this family are said to be independent if for every finite subset S_0 of S , we have

$$\mathbb{P}(\cap_{s \in S_0} A_s) = \prod_{s \in S_0} \mathbb{P}(A_s).$$

- (c) Let $\mathcal{F}_1 \subset \mathcal{F}$ and $\mathcal{F}_2 \subset \mathcal{F}$ be two σ -fields. We say that \mathcal{F}_1 and \mathcal{F}_2 are independent (write $\mathcal{F}_1 \perp\!\!\!\perp \mathcal{F}_2$) if any two events $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$ are independent.
- (d) More generally, let S be an index set, and for every $s \in S$, let \mathcal{F}_s be a σ -field contained in \mathcal{F} . We say that the σ -fields \mathcal{F}_s are **independent** if the following holds. If we pick one event A_s from each \mathcal{F}_s , the events in the resulting family $\{A_s \mid s \in S\}$ are independent.

Example. Consider an infinite sequence of fair coin tosses, under the model constructed in the Lecture 2 notes. The following statements are intuitively obvious (although a formal proof would require a few steps).

- (a) Let A_i be the event that the i th toss resulted in a “1”. If $i \neq j$, the events A_i and A_j are independent.
- (b) The events in the (infinite) family $\{A_i \mid i \in \mathbb{N}\}$ are independent. This statement captures the intuitive idea of “independent” coin tosses.
- (c) Let \mathcal{F}_n be the collection of all events whose occurrence can be decided by looking at the results of tosses $2n$ and $2n + 1$. (Note that each \mathcal{F}_n is a σ -field comprised of finitely many events.) Then, the families \mathcal{F}_n , $n \in \mathbb{N}$, are independent.
- (d) Let \mathcal{F}_1 (respectively, \mathcal{F}_2) be the collection of all events whose occurrence can be decided by looking at the results of the coin tosses at odd (respectively, even) times n . More formally, Let H_i be the event that the i th toss resulted in a 1. Let \mathcal{C} be the collection of events $\mathcal{C} = \{H_i \mid i \text{ is odd}\}$, and finally let $\mathcal{F}_1 = \sigma(\mathcal{C})$, so that \mathcal{F}_1 is the smallest σ -field that contains all the events H_i , for odd i . We define \mathcal{F}_2 similarly, using even times instead of odd times. Then, the two σ -fields \mathcal{F}_1 and \mathcal{F}_2 turn out to be independent. This statement captures the intuitive idea that knowing the results of the tosses at odd times provides no information on the results of the tosses at even times.

2.1 How to check independence of σ -algebras? p -systems.

How can one establish that two complicated σ -fields (e.g., as in the last example above) are independent? It turns out that one only needs to check independence for smaller collections of sets – see the theorem below. This is similar to the question of uniqueness of extension that we discussed in previous Lecture. There we have seen an example of a collection of sets \mathcal{C} and a pair of distinct probability measures that coincide on \mathcal{C} but differ on $\sigma(\mathcal{C})$. At the same time we have shown that measures coinciding on an algebra \mathcal{A} must necessarily coincide on $\sigma(\mathcal{A})$. Similarly, one can show that checking independence between σ -algebras can be reduced to checking independence between any two generating algebras. In fact, in both of these questions we can reduce to checking collections that are even smaller than algebras.

Definition 3. A collection of sets Π closed under finite intersections (that is, $A, B \in \Pi \Rightarrow A \cap B \in \Pi$) is called a p -system.

Examples of p -systems are intervals (a, b) on \mathbb{R} , rectangles $(a, b) \times (c, d)$ on \mathbb{R}^2 , convex sets in \mathbb{R}^d , etc.

Theorem 2. Let Π_1 and Π_2 be p -systems and $\mathcal{F}_i = \sigma(\Pi_i)$, $i = 1, 2$. If

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \quad (1)$$

for every $A \in \Pi_1$, $B \in \Pi_2$, then \mathcal{F}_1 and \mathcal{F}_2 are independent.

Proof. Fix an arbitrary $B \in \Pi_2$ and define a collection of sets

$$\mathcal{L}_B \triangleq \{E \in \mathcal{F}_1 : \mathbb{P}(E \cap B) = \mathbb{P}(E)\mathbb{P}(B)\}.$$

By assumption $\Pi_1 \subseteq \mathcal{L}_B$. We also have:

1. Clearly $\Omega \in \mathcal{L}_B$.
2. If $A_1 \subset A_2$ and both belong to \mathcal{L}_B then from

$$\mathbb{P}((A_2 \setminus A_1) \cap B) + \mathbb{P}(A_1 \cap B) = \mathbb{P}(A_2 \cap B)$$

we conclude that $A_2 \setminus A_1 \in \mathcal{L}_B$.

3. \mathcal{L}_B is a monotone class. Indeed, if $A_n \nearrow A$ and $A_n \in \mathcal{L}_B$ then $A_n \cap B \nearrow A \cap B$ and by continuity of \mathbb{P} we have

$$\mathbb{P}(A \cap B) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n \cap B) = \mathbb{P}(B) \cdot \lim_{n \rightarrow \infty} \mathbb{P}(A_n) = \mathbb{P}(B)\mathbb{P}(A),$$

implying $A \in \mathcal{L}_B$. Similar argument holds for $A_n \searrow A$.

It turns out that 1 and 2 imply that \mathcal{L}_B contains $\alpha(\Pi_1)$ (see Proposition to follow). Thus by the monotone class theorem $\mathcal{L}_B = \mathcal{F}_1$. Thus (1) holds for all $A \in \mathcal{F}_1$ and $B \in \Pi_2$. By symmetry it also holds for all $A \in \Pi_1$ and $B \in \mathcal{F}_2$. And applying the above argument again (with Π_2 replaced by \mathcal{F}_2) for all of \mathcal{F}_1 and \mathcal{F}_2 . \square

Proposition 1. *Let Π be a p-system on Ω . Let \mathcal{D} be a collection containing Π satisfying the following:*

1. $\Omega \in \mathcal{D}$
2. *For all $A, B \in \mathcal{D}$ such that $A \subset B$ we have $B \setminus A \in \mathcal{D}$ (i.e. \mathcal{D} closed under “punching holes”).*

Then $\mathcal{D} \supset \alpha(\Pi)$. Thus $\alpha(\Pi)$ is the smallest collection of sets containing Π and closed under punching holes.

Proof. Let \mathcal{D}_0 be the smallest collection of sets containing Π and satisfying conditions 1 and 2. We will show $\mathcal{D}_0 \supset \alpha(\Pi)$. Note that any p-system satisfying 1 and 2 is automatically an algebra. Thus it is sufficient to prove \mathcal{D}_0 is a p-system. Fix $C \in \Pi$ and let

$$\mathcal{L}_C = \{A \in \mathcal{D}_0 : A \cap C \in \mathcal{D}_0\}$$

On one hand, \mathcal{L}_C contains Π and Ω . On the other hand, \mathcal{L}_C is closed under punching holes: For $A \subset B$ we have $(B \setminus A) \cap C = (B \cap C) \setminus (A \cap C)$. Thus $\mathcal{L}_C = \mathcal{D}_0$ by minimality of \mathcal{D}_0 . Hence \mathcal{D}_0 is closed under intersections with elements of Π .

Next take an arbitrary $D \in \mathcal{D}_0$. We have

$$\mathcal{L}_D = \{A \in \mathcal{D}_0 : A \cap D \in \mathcal{D}_0\}$$

containing Π (and Ω) by the previous argument and closed under punching holes (same reasoning). Thus $\mathcal{L}_D = \mathcal{D}_0$ and \mathcal{D}_0 is closed under intersections. \square

Exercise 1. *Let Π be a p-system and $\mathcal{A} = \alpha(\Pi)$ the algebra generated by it. Suppose P and Q are two finitely additive non-negative set-functions with $P(\Omega) = Q(\Omega) = 1$. Show that if P and Q agree on Π then they agree on \mathcal{A} . (Hint: As usual let $\mathcal{D} = \{E \in \mathcal{A} : P(E) = Q(E)\}$ and use the proposition above).*

Exercise 2. *Show that $\alpha(\mathcal{C})$ consists of \emptyset, Ω and all sets that can be written in the sum-of-products form (this should remind you of disjunctive normal form).*

$$(A_{1,1} \cap \dots \cap A_{1,n_1}) \cup \dots \cup (A_{m,1} \cap \dots \cap A_{m,n_m}),$$

with $A_{i,j} \in \mathcal{C}$ or $A_{i,j}^c \in \mathcal{C}$.

3 THE BOREL-CANTELLI LEMMA

The Borel-Cantelli lemma is a tool that is often used to establish that a certain event has probability zero or one. Given a sequence of events A_n , $n \in \mathbb{N}$, the event $\{A_n \text{ i.o.}\}$ (read as “ A_n occurs infinitely often”) is defined to be the event consisting of all $\omega \in \Omega$ that belong to infinitely many A_n . Show that equivalently

$$\{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i.$$

This event is also sometimes denoted by $\limsup_{n \rightarrow \infty} A_n$.

Theorem 3. (Borel-Cantelli lemma) Let $\{A_n\}$ be a sequence of events and let $A = \{A_n \text{ i.o.}\}$.

- (a) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A) = 0$.
- (b) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and the events A_n , $n \in \mathbb{N}$, are independent, then $\mathbb{P}(A) = 1$.

Remark: The result in part (b) is not true without the independence assumption. Indeed, consider an arbitrary event C such that $0 < \mathbb{P}(C) < 1$ and let $A_n = C$ for all n . Then $\mathbb{P}(\{A_n \text{ i.o.}\}) = \mathbb{P}(C) < 1$, even though $\sum_n \mathbb{P}(A_n) = \infty$.

The following lemma is useful here and in many other contexts.

Lemma 1. Suppose that $0 \leq p_i \leq 1$ for every $i \in \mathbb{N}$. Then:

$$\sum_{i=1}^{\infty} p_i = \infty \quad \Rightarrow \quad \prod_{i=1}^{\infty} (1 - p_i) = 0 \tag{2}$$

$$\sum_{i=1}^{\infty} p_i = \infty \quad \Leftarrow \quad \prod_{i=1}^{\infty} (1 - p_i) = 0, p_i < 1 \tag{3}$$

Proof. Note that $\log(1 - x)$ is a concave function of its argument, and its derivative at $x = 0$ is -1 . It follows that $\log(1 - x) \leq -x$, for $x \in [0, 1]$. We then

have

$$\begin{aligned}
\log \prod_{i=1}^{\infty} (1 - p_i) &= \log \left(\lim_{k \rightarrow \infty} \prod_{i=1}^k (1 - p_i) \right) \\
&\leq \log \prod_{i=1}^k (1 - p_i) \\
&= \sum_{i=1}^k \log(1 - p_i) \\
&\leq \sum_{i=1}^k (-p_i).
\end{aligned}$$

This is true for every k . By taking the limit as $k \rightarrow \infty$, we obtain $\log \prod_{i=1}^{\infty} (1 - p_i) = -\infty$, and $\prod_{i=1}^{\infty} (1 - p_i) = 0$.

For the converse statement, note that under $p_i < 1$ we have

$$\prod_{i=1}^{\infty} (1 - p_i) = 0 \iff \forall n \prod_{i=n}^{\infty} (1 - p_i) = 0$$

If $p_i \not\rightarrow 0$ the result is automatic. Hence, we may also assume $p_i \rightarrow 0$. Then taking n so large that $p_i \leq 1 - e^{-1}$ for all $i \geq n$ we may apply the lower bound

$$\log(1 - x) \geq -\frac{e}{e-1}x \quad \forall 0 \leq x \leq 1 - e^{-1}.$$

Then, for arbitrary large C we have for all sufficiently large n

$$-C \geq \log \prod_{i=1}^n (1 - p_i) \geq -\frac{e}{e-1} \sum_{i=1}^n p_i,$$

implying $\sum_{i=1}^{\infty} p_i \geq C'$ for all $C' > 0$. □

Proof of Theorem 3.

- (a) The assumption $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ implies that $\lim_{n \rightarrow \infty} \sum_{i=n}^{\infty} \mathbb{P}(A_i) = 0$. Note that for every n , we have $A \subset \cup_{i=n}^{\infty} A_i$. Then, the union bound implies that

$$\mathbb{P}(A) \leq \mathbb{P}(\cup_{i=n}^{\infty} A_i) \leq \sum_{i=n}^{\infty} \mathbb{P}(A_i).$$

We take the limit of both sides as $n \rightarrow \infty$. Since the right-hand side converges to zero, $\mathbb{P}(A)$ must be equal to zero.

- (b) Let $B_n = \cup_{i=n}^{\infty} A_i$, and note that $A = \cap_{n=1}^{\infty} B_n$. We claim that $\mathbb{P}(B_n^c) = 0$. This will imply the desired result because

$$\mathbb{P}(A^c) = \mathbb{P}\left(\cup_{n=1}^{\infty} B_n^c\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(B_n^c) = 0.$$

Let us fix some n and some $m \geq n$. We have, using independence (show that independence of $\{A_n\}$ implies independence of $\{A_n^c\}$)

$$\mathbb{P}(\cap_{i=n}^m A_i^c) = \prod_{i=n}^m \mathbb{P}(A_i^c) = \prod_{i=n}^m (1 - \mathbb{P}(A_i)).$$

The assumption $\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \infty$ implies that $\sum_{i=n}^{\infty} \mathbb{P}(A_i) = \infty$. Using Lemma 1, with the sequence $\{\mathbb{P}(A_i) \mid i \geq n\}$ replacing the sequence $\{p_i\}$, we obtain

$$\mathbb{P}(B_n^c) = \mathbb{P}(\cap_{i=n}^{\infty} A_i^c) = \lim_{m \rightarrow \infty} \mathbb{P}(\cap_{i=n}^m A_i^c) = \lim_{m \rightarrow \infty} \prod_{i=n}^m (1 - \mathbb{P}(A_i)) = 0,$$

where the second equality made use of the continuity property of probability measures. \square

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

RANDOM VARIABLES

Contents

1. Random variables and measurable functions
2. Cumulative distribution functions
3. Discrete random variables
4. Continuous random variables
 - A. Continuity
 - B. The Cantor set, an unusual random variable, and a singular measure

1 RANDOM VARIABLES AND MEASURABLE FUNCTIONS

Loosely speaking a random variable provides us with a numerical value, depending on the outcome of an experiment. More precisely, a random variable can be viewed as a function from the sample space to the real numbers, and we will use the notation $X(\omega)$ to denote the numerical value of a random variable X , when the outcome of the experiment is some particular ω . We may be interested in the probability that the outcome of the experiment is such that X is no larger than some c , i.e., that the outcome belongs to the set $\{\omega \mid X(\omega) \leq c\}$. Of course, in order to have a probability assigned to that set, we need to make sure that it is \mathcal{F} -measurable. This motivates Definition 1 below.

Example 1. Consider a sequence of five consecutive coin tosses. An appropriate sample space is $\Omega = \{0, 1\}^n$, where “1” stands for heads and “0” for tails. Let \mathcal{F} be the collection of all subsets of Ω , and suppose that a probability measure \mathbb{P} has been assigned to (Ω, \mathcal{F}) . We are interested in the number of heads obtained in this experiment. This quantity can be described by the function $X : \Omega \rightarrow \mathbb{R}$, defined by

$$X(\omega_1, \dots, \omega_n) = \omega_1 + \dots + \omega_n.$$

With this definition, the set $\{\omega \mid X(\omega) < 4\}$ is just the event that there were fewer than 4 heads overall, belongs to the σ -field \mathcal{F} , and therefore has a well-defined probability.

Consider the real line, and let \mathcal{B} be the associated Borel σ -field. Sometimes, we will also allow random variables that take values in the extended real line, $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. We define the Borel σ -field on $\overline{\mathbb{R}}$, also denoted by \mathcal{B} , as the smallest σ -field that contains all Borel subsets of \mathbb{R} and the sets $\{-\infty\}$ and $\{\infty\}$.

1.1 Random variables

Definition 1. (Random variables) Let (Ω, \mathcal{F}) be a measurable space.

- (a) A function $X : \Omega \rightarrow \mathbb{R}$ is a **random variable** if the set $\{\omega \mid X(\omega) \leq c\}$ is \mathcal{F} -measurable for every $c \in \mathbb{R}$.
- (b) A function $X : \Omega \rightarrow \overline{\mathbb{R}}$ is an **extended-valued random variable** if the set $\{\omega \mid X(\omega) \leq c\}$ is \mathcal{F} -measurable for every $c \in \overline{\mathbb{R}}$.

We note here a convention that will be followed throughout: we will always use upper case letters to denote random variables and lower case letters to denote numerical values (elements of $\overline{\mathbb{R}}$). Thus, a statement such as “ $X(\omega) = x = 5$ ” means that when the outcome happens to be ω , then the realized value of the random variable is a particular number x , equal to 5.

Example 2. (Indicator functions) Suppose that $A \subset \Omega$, and let $I_A : \Omega \rightarrow \{0, 1\}$ be the indicator function of that set; i.e., $I_A(\omega) = 1$, if $\omega \in A$, and $I_A(\omega) = 0$, otherwise. If $A \in \mathcal{F}$, then I_A is a random variable. But if $A \notin \mathcal{F}$, then I_A is not a random variable.

Example 3. (A function of a random variable) Suppose that X is a random variable, and let us define a function $Y : \Omega \rightarrow \mathbb{R}$ by letting $Y(\omega) = X^3(\omega)$, for every $\omega \in \Omega$, or $Y = X^3$ for short. We claim that Y is also a random variable. Indeed, for any $c \in \mathbb{R}$, the set $\{\omega \mid Y(\omega) \leq c\}$ is the same as the set $\{\omega \mid X(\omega) \leq c^{1/3}\}$, which is in \mathcal{F} , since X is a random variable.

1.2 The law of a random variable

For a random variable X , the event $\{\omega \mid X(\omega) \leq c\}$ is often written as $\{X \leq c\}$, and is sometimes just called “the event that $X \leq c$.” The probability of this event is well defined, since this event belongs to \mathcal{F} . Let now B be a more general subset of the real line. We use the notation $X^{-1}(B)$ or $\{X \in B\}$ to denote the set $\{\omega \mid X(\omega) \in B\}$.

Because the collection of intervals of the form $(-\infty, c]$ generates the Borel σ -field in \mathbb{R} , it can be shown that if X is a random variable, then for any Borel set B , the set $X^{-1}(B)$ is \mathcal{F} -measurable. It follows that the probability $\mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \mid X(\omega) \in B\})$ is well-defined. It is often denoted by $\mathbb{P}(X \in B)$.

Exercise 1. Suppose X is a random variable. Show that for every Borel subset $B \subset \mathbb{R}$, the set $X^{-1}(B)$ is \mathcal{F} -measurable. (Hint: Define the collection $\mathcal{L} = \{B : X^{-1}(B) \in \mathcal{F}\}.$)

Definition 2. (The probability law of a random variable) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \rightarrow \mathbb{R}$ be a random variable.

- (a) For every Borel subset B of the real line (i.e., $B \in \mathcal{B}$), we define $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$.
- (b) The resulting function $\mathbb{P}_X : \mathcal{B} \rightarrow [0, 1]$ is called the **probability law** of X .

Sometimes, \mathbb{P}_X is also called the distribution of X , not to be confused with the cumulative distribution function defined in the next section.

According to the next result, the law \mathbb{P}_X of X is also a probability measure. Notice here that \mathbb{P}_X is a measure on $(\mathbb{R}, \mathcal{B})$, as opposed to the original measure \mathbb{P} , which is a measure on (Ω, \mathcal{F}) . In many instances, the original probability space $(\Omega, \mathcal{F}, \mathbb{P})$ remains in the background, hidden or unused, and one works directly with the much more tangible probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$. Indeed, if we are only interested in the statistical properties of the random variable X , the latter space will do.

Proposition 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let X be a random variable. Then, the law \mathbb{P}_X of X is a probability measure on $(\mathbb{R}, \mathcal{B})$.

Proof: Clearly, $\mathbb{P}_X(B) \geq 0$, for every Borel set B . Also, $\mathbb{P}_X(\mathbb{R}) = \mathbb{P}(X \in \mathbb{R}) = \mathbb{P}(\Omega) = 1$. We now verify countable additivity. Let $\{B_i\}$ be a countable sequence of disjoint Borel subsets of \mathbb{R} . Note that the sets $X^{-1}(B_i)$ are also disjoint, and that

$$X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right) = \bigcup_{i=1}^{\infty} X^{-1}(B_i),$$

or, in different notation,

$$\{X \in \bigcup_{i=1}^{\infty} B_i\} = \bigcup_{i=1}^{\infty} \{X \in B_i\}.$$

Therefore, using countable additivity on the original probability space, we have

$$\mathbb{P}_X(\cup_{i=1}^{\infty} B_i) = \mathbb{P}(X \in \cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mathbb{P}(X \in B_i) = \sum_{i=1}^{\infty} \mathbb{P}_X(B_i).$$

□

1.3 Technical digression: measurable functions

The following generalizes the definition of a random variable.

Definition 3. Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be two measurable spaces. A function $f : \Omega_1 \rightarrow \Omega_2$ is called **$(\mathcal{F}_1, \mathcal{F}_2)$ -measurable** (or just measurable, if the relevant σ -fields are clear from the context) if $f^{-1}(B) \in \mathcal{F}_1$ for every $B \in \mathcal{F}_2$.

According to the above definition, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and taking into account the discussion in Section 1.1, a random variable on a probability space is a function $X : \Omega \rightarrow \mathbb{R}$ that is $(\mathcal{F}, \mathcal{B})$ -measurable.

As a general rule, functions constructed from other measurable functions using certain simple operations are measurable. We collect, without proof, a number of relevant facts below.

Theorem 1. Let (Ω, \mathcal{F}) be a measurable space.

- (a) **(Simple random variables)** If $A \in \mathcal{F}$, the corresponding indicator function I_A is measurable (more, precisely, it is $(\mathcal{F}, \mathcal{B})$ -measurable).
 - (b) If A_1, \dots, A_n are \mathcal{F} -measurable sets, and x_1, \dots, x_n are real numbers, the function $X = \sum_{i=1}^n x_i I_{A_i}$, or in more detail,
- $$X(\omega) = \sum_{i=1}^n x_i I_{A_i}(\omega), \quad \forall \omega \in \Omega,$$
- is a random variable (and is called a **simple** random variable).
- (c) Suppose that $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$, and that $X : \mathbb{R} \rightarrow \mathbb{R}$ is continuous. Then, X is a random variable.
 - (d) **(Functions of a random variable)** Let X be a random variable, and suppose that $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous (or more generally, $(\mathcal{B}, \mathcal{B})$ -measurable). Then, $f(X)$ is a random variable.
 - (e) **(Functions of multiple random variables)** Let X_1, \dots, X_n be random variables, and suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous. Then, $f(X_1, \dots, X_n)$ is a random variable. In particular, $X_1 + X_2$ and $X_1 X_2$ are random variables.

Another way that we can form a random variable is by taking the limit of a sequence of random variables. Let us first introduce some terminology. Let each f_n be a function from some set Ω into \mathbb{R} . Consider a new function $f = \inf_n f_n$ defined by $f(\omega) = \inf_n f_n(\omega)$, for every $\omega \in \Omega$. The functions $\sup_n f_n$, $\liminf_{n \rightarrow \infty} f_n$, and $\limsup_{n \rightarrow \infty} f_n$ are defined similarly. (Note that even if the f_n are everywhere finite, the above defined functions may turn out to be extended-valued.) If the limit $\lim_{n \rightarrow \infty} f_n(\omega)$ exists for every ω , we say that the sequence of functions $\{f_n\}$ **converges pointwise**, and define its **pointwise limit** to be the function f defined by $f(\omega) = \lim_{n \rightarrow \infty} f_n(\omega)$. For example, suppose that $\Omega = [0, 1]$ and that $f_n(\omega) = \omega^n$. Then, the pointwise limit $f = \lim_{n \rightarrow \infty} f_n$ exists, and satisfies $f(1) = 1$, and $f(\omega) = 0$ for $\omega \in [0, 1)$.

Theorem 2. Let (Ω, \mathcal{F}) be a measurable space. If X_n is a random variable for every n , then $\inf_n X_n$, $\sup_n X_n$, $\liminf_{n \rightarrow \infty} X_n$, and $\limsup_{n \rightarrow \infty} X_n$ are random variables. Furthermore, if the sequence $\{X_n\}$ converges pointwise, and $X = \lim_{n \rightarrow \infty} X_n$, then X is also a random variable.

As a special case of Theorem 2, we have that a pointwise limit of a sequence of simple random variables (with “simple” defined in the statement of Theorem 1) is measurable. On the other hand, we note that measurable functions can be highly discontinuous. For example the function which equals 1 at every rational number, and equals zero otherwise, is measurable, because it is the indicator function of a measurable set.

2 CUMULATIVE DISTRIBUTION FUNCTIONS

A simple way of describing the probabilistic properties of a random variable is in terms of the cumulative distribution function, which we now define.

Definition 4. (Cumulative distribution function) Let X be a random variable. The function $F_X : \mathbb{R} \rightarrow [0, 1]$, defined by

$$F_X(x) = \mathbb{P}(X \leq x),$$

is called the cumulative distribution function (**CDF**) of X .

Example 4. Let X be the number of heads in two independent tosses of a fair coin. In particular, $\mathbb{P}(X = 0) = \mathbb{P}(X = 2) = 1/4$, and $\mathbb{P}(X = 1) = 1/2$. Then,

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1/4, & \text{if } 0 \leq x < 1, \\ 3/4, & \text{if } 1 \leq x < 2, \\ 1, & \text{if } x \geq 2. \end{cases}$$

Example 5. (A uniform random variable and its square) Consider a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, where $\Omega = [0, 1]$, \mathcal{B} is the Borel σ -field \mathcal{B} , and \mathbb{P} is the Lebesgue measure. The random variable U defined by $U(\omega) = \omega$ is said to be uniformly distributed. Its CDF is given by

$$F_U(x) = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } 0 \leq x < 1, \\ 1, & \text{if } x \geq 1. \end{cases}$$

Consider now the random variable $X = U^2$. We have $\mathbb{P}(U^2 \leq x) = 0$, when $x < 0$, and $\mathbb{P}(U^2 \leq x) = 1$, when $x \geq 1$. For $x \in [0, 1)$, we have

$$\mathbb{P}(U^2 \leq x) = \mathbb{P}(\{\omega \in [0, 1] \mid \omega^2 \leq x\}) = \mathbb{P}(\{\omega \in [0, 1] : \omega \leq \sqrt{x}\}) = \sqrt{x}.$$

Thus,

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0, \\ \sqrt{x}, & \text{if } 0 \leq x < 1, \\ 1, & \text{if } x \geq 1. \end{cases}$$

2.1 CDF properties

The cumulative distribution function of a random variable always possesses certain properties.

Theorem 3. *Let X be a random variable, and let F be its CDF.*

- (a) **(Monotonicity)** If $x \leq y$, then $F_X(x) \leq F_X(y)$.
- (b) **(Limiting values)** We have $\lim_{x \rightarrow -\infty} F_X(x) = 0$, and $\lim_{x \rightarrow \infty} F(x) = 1$.
- (c) **(Right-continuity)** For every x , we have $\lim_{y \downarrow x} F_X(y) = F_X(x)$.

Proof:

- (a) Suppose that $x \leq y$. Then, $\{X \leq x\} \subset \{X \leq y\}$, which implies that

$$F(x) = \mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y) = F(y).$$

- (b) Since $F_X(x)$ is monotonic in x and bounded below by zero, it converges as $x \rightarrow -\infty$, and the limit is the same for every sequence $\{x_n\}$ converging to $-\infty$. So, let $x_n = -n$, and note that the sequence of events $\cap_{n=1}^{\infty} \{X \leq -n\}$ converges to the empty set. Using the continuity of probabilities, we obtain

$$\lim_{x \rightarrow -\infty} F_X(x) = \lim_{n \rightarrow \infty} F_X(-n) = \lim_{n \rightarrow \infty} \mathbb{P}(X \leq -n) = \mathbb{P}(\emptyset) = 0.$$

The proof of $\lim_{x \rightarrow \infty} F_X(x) = 1$ is similar, and is omitted.

- (c) Consider a decreasing sequence $\{x_n\}$ that converges to x . The sequence of events $\{X \leq x_n\}$ is decreasing and $\cap_{n=1}^{\infty} \{X \leq x_n\} = \{X \leq x\}$. Using the continuity of probabilities, we obtain

$$\lim_{n \rightarrow \infty} F_X(x_n) = \lim_{n \rightarrow \infty} \mathbb{P}(X \leq x_n) = \mathbb{P}(X \leq x) = F_X(x).$$

Since this is true for every such sequence $\{x_n\}$, we conclude that $\lim_{y \downarrow x} F_X(y) = F_X(x)$. \square

We note that CDFs need not be left-continuous. For instance, in Example 4, we have $\lim_{x \uparrow 0} F_X(x) = 0$, but $F_X(0) = 1/4$.

2.2 From a CDF to a probability law

Consider a function $F : \mathbb{R} \rightarrow [0, 1]$ that satisfies the three properties in Theorem 3; we call such a function a **distribution function**. Given a distribution function F , does there exist a random variable X , defined on some probability space, whose CDF, F_X , is equal to the given distribution F ? This is certainly the case for the distribution function F function that satisfies $F(x) = x$, for $x \in (0, 1)$: the uniform random variable U in Example 5 will do. More generally, the objective $F_X = F$ can be accomplished by letting $X = g(U)$, for a suitable function $g : (0, 1) \rightarrow \mathbb{R}$.

Theorem 4. *Let F be a given distribution function. Consider the probability space $([0, 1], \mathcal{B}, \mathbb{P})$, where \mathcal{B} is the Borel σ -field, and \mathbb{P} is the Lebesgue measure. There exists a measurable function $X : \Omega \rightarrow \mathbb{R}$ whose CDF F_X satisfies $F_X = F$.*

Proof: We first present the proof under an additional simplifying assumption that F is continuous and strictly increasing. Then, the range of F is the entire interval $(0, 1)$. Furthermore, F is invertible: for every $y \in (0, 1)$, there exists a unique x , denoted $F^{-1}(y)$, such that $F(x) = y$. We define $U(\omega) = \omega$ and $X(\omega) = F^{-1}(\omega)$, for every $\omega \in (0, 1)$, so that $X = F^{-1}(U)$. Note that $F(F^{-1}(\omega)) = \omega$ for every $\omega \in (0, 1)$, so that $F(X) = U$. Since F is strictly increasing, we have $X \leq x$ if and only $F(X) \leq F(x)$, or $U \leq F(x)$. (Note that this also establishes that the event $\{X \leq x\}$ is measurable, so that X is indeed a random variable.) Thus, for every $x \in \mathbb{R}$, we have

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(F(X) \leq F(x)) = \mathbb{P}(U \leq F(x)) = F(x),$$

as desired.

The general case is treated by defining a right-continuous inverse of F , called a *quantile function* of F :

$$q(y) = \inf\{x : F(x) > y\} \quad \forall 0 < y < 1.$$

It can be shown that for every y

$$\{x : y < F(x)\} \subseteq \{x : q(y) \leq x\} \subseteq \{x : y \leq F(x)\}$$

and therefore $F(U) \sim \mathbb{P}$. □

Note that the probability law of X assigns probabilities to all Borel sets, whereas the CDF only specifies the probabilities of certain intervals. Nevertheless, the CDF contains enough information to recover the law of X .

Corollary 1. *There is a one-to-one correspondence between distribution functions F and probability measures \mathbb{P} on $(\mathbb{R}, \mathcal{B})$.*

Proof: Indeed, for any CDF F Theorem 4 constructs a random variable X whose CDF $F_X = F$. On the other hand, for a random variable X , by Proposition 1, the induced set-function $\mathbb{P}_X(\cdot)$ is a probability measure on $(\mathbb{R}, \mathcal{B})$, and given a probability measure \mathbb{P}_X on $(\mathbb{R}, \mathcal{B})$ we obtain a CDF $F(c) = \mathbb{P}_X((-\infty, c])$.

It remains to check that different probability measures \mathbb{P}_X and \mathbb{P}'_X necessarily yield different CDFs. Indeed, if \mathbb{P}_X and \mathbb{P}'_X coincide on all intervals $(-\infty, c]$ then $\mathbb{P}_X = \mathbb{P}'_X$ by Proposition 1 of Lecture 2 (see Remark there) since the collection of sets $\{(-\infty, c], c \in \mathbb{R}\}$ is a generating p -system for \mathcal{B} . \square

3 DISCRETE RANDOM VARIABLES

Discrete random variables take values in a countable set. We need some notation. Given a function $f : \Omega \rightarrow \mathbb{R}$, its **range** is the set

$$f(\Omega) = \{x \in \mathbb{R} \mid \exists \omega \in \Omega \text{ such that } f(\omega) = x\}.$$

Definition 5. Discrete random variables and PMFs

- (a) *A random variable X , defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, is said to be discrete if its range $X(\Omega)$ is finite or countable.*
- (b) *If X is a discrete random variable, the function $p_X : \mathbb{R} \rightarrow [0, 1]$ defined by $p_X(x) = \mathbb{P}(X = x)$, for every x , is called the **(probability) mass function** of X , or **PMF** for short.*

Consider a discrete random variable X whose range is a finite set C . In that case, for any Borel set A , countable additivity yields

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A \cap C) = \sum_{x \in A \cap C} \mathbb{P}(X = x).$$

In particular, the CDF is given by

$$F_X(x) = \sum_{\{y \in A \cap C \mid y \leq x\}} \mathbb{P}(X = y).$$

A random variable that takes only integer values is discrete. For instance, the random variable in Example 4 (number of heads in two coin tosses) is discrete.

Also, every simple random variable is discrete, since it takes a finite number of values. However, more complicated discrete random variables are also possible.

Example 6. Let the sample space be the set \mathbb{N} of natural numbers, and consider a measure that satisfies $\mathbb{P}(n) = 1/2^n$, for every $n \in \mathbb{N}$. The random variable X defined by $X(n) = n$ is discrete.

Suppose now that the rational numbers have been arranged in a sequence, and that x_n is the n th rational number, according to this sequence. Consider the random variable Y defined by $Y(n) = x_n$. The range of this random variable is countable, so Y is a discrete random variable. Its range is the set of rational numbers, every rational number has positive probability, and the set of irrational numbers has zero probability.

We close by noting that discrete random variables can be represented in terms of indicator functions. Indeed, given a discrete random variable X , with range $\{x_1, x_2, \dots\}$, we define $A_n = \{X = x_n\}$, for every $n \in \mathbb{N}$. Observe that each set A_n is measurable (why?). Furthermore, the sets A_n , $n \in \mathbb{N}$, form a partition of the sample space. Using indicator functions, we can write

$$X(\omega) = \sum_{n=1}^{\infty} x_n I_{A_n}(\omega).$$

Conversely, suppose we are given a sequence $\{A_n\}$ of disjoint events, and a real sequence $\{x_n\}$. Define $X : \Omega \rightarrow \mathbb{R}$ by letting $X(\omega) = x_n$ if and only if $\omega \in A_n$. Then X is a discrete random variable, and $\mathbb{P}(X = x_n) = \mathbb{P}(A_n)$, for every n .

4 CONTINUOUS RANDOM VARIABLES

The definition of a continuous random variable is more subtle. It is not enough for a random variable to have a “continuous range.”

Definition 6. A random variable X , defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, is said to be **continuous** if there exists a nonnegative measurable function $f : \mathbb{R} \rightarrow [0, \infty)$ such that

$$F_X(x) = \int_{-\infty}^x f(t) dt, \quad \forall x \in \mathbb{R}. \tag{1}$$

The function f is called a **(probability) density function** (or **PDF**, for short) for X ,

There is some ambiguity in the above definition, because the meaning of the integral of a measurable function may be unclear. We will see later in this course how such an integral is defined. For now, we just note that the integral is well-defined, and agrees with the Riemann integral encountered in calculus, if the function is continuous, or more generally, if it has a finite number of discontinuities.

Since $\lim_{x \rightarrow \infty} F_X(x) = 1$, we must have $\lim_{x \rightarrow \infty} \int_{-\infty}^x f(t) dt = 1$, or

$$\int_{-\infty}^{\infty} f_X(t) dt = 1. \quad (2)$$

Any nonnegative measurable function that satisfies Eq. (2) is called a density function. Conversely, given a density function f , we can define $F(x) = \int_{-\infty}^x f(t) dt$, and verify that F is a distribution function. It follows that given a density function, there always exists a random variable whose PDF is the given density.

If a CDF F_X is differentiable at some x , the corresponding value $f_X(x)$ can be found by taking the derivative of F_X at that point. However, CDFs need not be differentiable, so this will not always work. Let us also note that a PDF of a continuous random variable is not uniquely defined. We can always change the PDF at a finite set of points, without affecting its integral, hence multiple PDFs can be associated to the same CDF. However, this nonuniqueness rarely becomes an issue. In the sequel, we will often refer to “the PDF” of X , ignoring the fact that it is nonunique.

Example 7. For a uniform random variable, we have $F_X(x) = \mathbb{P}(X \leq x) = x$, for every $x \in (0, 1)$. By differentiating, we find $f_X(x) = 1$, for $x \in (0, 1)$. For $x < 0$ we have $F_X(x) = 0$, and for $x > 1$ we have $F_X(x) = 1$; in both cases, we obtain $f_X(x) = 0$. At $x = 0$, the CDF is not differentiable. We are free to define $f_X(0)$ to be 0, or 1, or in fact any real number; the value of the integral of f_X will remain unaffected.

Example 8. Consider the random variable $X(\omega) = \omega^2$ on $[0, 1]$ which we discussed in Section 1. Let $f(t) = \frac{1}{2\sqrt{t}}$ when $t \in [0, 1]$ and $f(t) = 0$ for all other t . Then for every $x \in [0, 1]$ we have $\int_{-\infty}^x f(t) dt = \int_0^x f(t) dt = \sqrt{x} = F(x)$. We can check trivially that the equality holds for all other values of x . Thus $f(t) = \frac{1}{2\sqrt{t}}$ is the density function corresponding to this probability distribution.

Using the PDF of a continuous random variable, we can calculate the probability of various subsets of the real line. For example, we have $\mathbb{P}(X = x) = 0$, for all x , and if $a < b$,

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = \int_a^b f_X(t) dt.$$

More generally, for any Borel set B , it turns out that

$$\mathbb{P}(X \in B) = \int_B f(t) dt = \int I_B(t) f_X(t) dt,$$

and that $\mathbb{P}(X \in B) = 0$ whenever B has Lebesgue measure zero. However, more detail on this subject must wait until we develop the theory of integration of measurable functions.

We close by pointing out that not all random variables are continuous or discrete. For example, suppose that X is a discrete random variable, and that Y is a continuous random variable. Fix some $\lambda \in (0, 1)$, and define

$$F(z) = \lambda F_X(x) + (1 - \lambda) F_Y(y). \quad (3)$$

It can be verified that F is a distribution function, and therefore can be viewed as the CDF of some new random variable Z . However, the random variable Z is neither discrete, nor continuous. For an interpretation, we can visualize Z being generated as follows: we first generate the random variables X and Y ; then, with probability λ , we set $Z = X$, and with probability $1 - \lambda$, we set $Z = Y$.

Even more pathological random variables are possible. Appendix B discusses a particularly interesting one.

5 APPENDIX A – CONTINUOUS FUNCTIONS

We introduce here some standard notation and terminology regarding convergence of function values, and continuity.

- (a) Consider a function $f : \mathbb{R}^m \rightarrow \mathbb{R}$, and fix some $x \in \mathbb{R}^m$. We say that $f(y)$ **converges** to a value c , as y tends to x , if we have $\lim_{n \rightarrow \infty} f(x_n) = c$, for every sequence $\{x_n\}$ of elements of \mathbb{R}^m such that $x_n \neq x$ for all n , and $\lim_{n \rightarrow \infty} x_n = x$. In this case, we write $\lim_{y \rightarrow x} f(y) = c$.
- (b) If $f : \mathbb{R}^m \rightarrow \mathbb{R}$ and $\lim_{y \rightarrow x} f(y) = f(x)$, we say that f is **continuous** at x . If this holds for every x , we say that f is continuous.
- (c) Consider a function $f : \mathbb{R} \rightarrow \mathbb{R}$, and fix some $x \in \mathbb{R} \cup \{-\infty\}$. We say that $f(y)$ **converges** to a value c , as y decreases to x , if we have $\lim_{n \rightarrow \infty} f(x_n) = c$, for every decreasing sequence $\{x_n\}$ of elements of \mathbb{R}^m such that $x_n > x$ for all n , and $\lim_{n \rightarrow \infty} x_n = x$. In this case, we write $\lim_{y \downarrow x} f(y) = c$.
- (d) If $f : \mathbb{R} \rightarrow \mathbb{R}$ and $\lim_{y \downarrow x} f(y) = f(x)$, we say that the function f is **right-continuous** at x . If this holds for every $x \in \mathbb{R}$, we say that f is right-continuous.

- (e) Consider a function $f : \mathbb{R} \rightarrow \mathbb{R}$, and fix some $x \in \mathbb{R} \cup \{\infty\}$. We say that $f(y)$ **converges** to a value c , as y increases to x , if we have $\lim_{n \rightarrow \infty} f(x_n) = c$, for every increasing sequence $\{x_n\}$ of elements of \mathbb{R} such that $x_n < x$ for all n , and $\lim_{n \rightarrow \infty} x_n = x$. In this case, we write $\lim_{y \uparrow x} f(y) = c$.
- (f) If $f : \mathbb{R} \rightarrow \mathbb{R}$ and $\lim_{y \uparrow x} f(y) = f(x)$, we say that the function f is **left-continuous** at x . If this holds for every $x \in \mathbb{R}$, we say that f is left-continuous.

For a function $f : \mathbb{R} \rightarrow \mathbb{R}$, and some $x \in \mathbb{R}$, it is not hard to show, starting from the above definitions, that f is continuous at x if and only if it is both left-continuous and right-continuous.

6 APPENDIX B – THE CANTOR SET, AN UNUSUAL RANDOM VARIABLE, AND A SINGULAR MEASURE

Every number $x \in [0, 1]$ has a ternary expansion of the form

$$x = \sum_{i=1}^{\infty} \frac{x_i}{3^i}, \quad \text{with } x_i \in \{0, 1, 2\}. \quad (4)$$

This expansion is not unique. For example, $1/3$ admits two expansions, namely $.10000\dots$ and $.022222\dots$. Nonuniqueness occurs only for those x that admit an expansion ending with an infinite sequence of 2s. The set of such unusual x is countable, and therefore has Lebesgue measure zero.

The **Cantor set** C is defined as the set of all $x \in [0, 1]$ that have a ternary expansion that uses only 0s and 2s (no 1s allowed). The set C can be constructed as follows. Start with the interval $[0, 1]$ and remove the “middle third” $(1/3, 2/3)$. Then, from each of the remaining closed intervals, $[0, 1/3]$ and $[2/3, 1]$, remove their middle thirds, $(1/9, 2/9)$ and $(7/9, 8/9)$, resulting in four closed intervals, and continue this process indefinitely. Note that C is measurable, since it is constructed by removing a countable sequence of intervals. Also, the length (Lebesgue measure) of C is 0, since at each stage its length is multiplied by a factor of $2/3$. On the other hand, the set C has the same cardinality as the set $\{0, 2\}^{\infty}$, and is uncountable.

Consider now an infinite sequence of independent rolls of a 3-sided die, whose faces are labeled 0, 1, and 2. Assume that at each roll, each of the three possible results has the same probability, $1/3$. If we use the sequence of these rolls to form a number x , then the probability law of the resulting random variable is the Lebesgue measure (i.e., picking a ternary expansion “at random” leads to a uniform random variable).

The Cantor set can be identified with the event consisting of all roll sequences in which a 1 never occurs. (This event has zero probability, which is consistent with the fact that C has zero Lebesgue measure.)

Consider now an infinite sequence of independent tosses of a fair coin. If the i th toss results in tails, record $x_i = 0$; if it results in heads, record $x_i = 2$. Use the x_i s to form a number x , using Eq. (4). This defines a random variable X on $([0, 1], \mathcal{B})$, whose range is the set C . The probability law of this random variable is therefore concentrated on the “zero-length” set C . At the same time, $\mathbb{P}(X = x) = 0$ for every x , because any particular sequence of heads and tails has zero probability. A measure with this property is called **singular**.

The random variable X that we have constructed here is neither discrete nor continuous. Moreover, the CDF of X cannot be written as a mixture of the kind considered in Eq. (3).

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

DISCRETE RANDOM VARIABLES AND THEIR EXPECTATIONS**Contents**

1. Combinatorial probability
2. A few useful discrete random variables
3. Joint, marginal, and conditional PMFs
4. Independence of random variables
5. Expected values

1 COMBINATORIAL PROBABILITY

In this section we will briefly review some combinatorial concepts, which come in handy when performing actual computations with discrete random variables. We start with two results from analysis:

1. Exponential function as a limit, for all $x \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = e^x.$$

2. Stirling bounds on factorial

$$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}.$$

Definition 1. Let Ω be a finite sample space. The discrete uniform probability space is $(\Omega, 2^\Omega, \mathbb{P})$, where

$$\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|} \quad \forall \omega \in \Omega.$$

Moreover, for any event $A \subset \Omega$, by finite additivity,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}.$$

Such assignments form the foundation of combinatorial probability wherein one is usually interested in counting the number of elements satisfying a particular criterion. This counting is often done following an iterative procedure. We collect some specific examples.

Example 1(Permutations). *A permutation of the numbers $1, \dots, n$ is an isomorphism $\pi: \{1, \dots, n\} \rightarrow \{1, \dots, n\}$. Proceeding iteratively, there are n choices for $\pi(1)$, $n - 1$ choices for $\pi(2)$, \dots , and 1 choice for $\pi(n)$. Therefore, there are $n \cdot (n - 1) \cdots 2 \cdot 1 = n!$ possible permutations.*

Example 2(Choices). *Given a collection of n distinct objects there are $n \cdot (n - 1) \cdots (n - (k - 2)) \cdot (n - (k - 1)) = n!/(n - k)!$ ways to select k of those objects in a particular order. Moreover, there are $n!/((n - k)!k!)$ ways to select k of those objects ignoring order, as $k!$ represents all possible ways to arrange k objects. This last expression $n!/((n - k)!k!)$ is denoted $\binom{n}{k}$ and referred to as n choose k .*

Example 3(Birthday Paradox). *Given n individuals, what is the probability that no two have the same birthday? Let A = “All individuals in a group of size n have unique birthdays”. Assume that an individual’s birthday is independent of all other birthdays and occurs equally likely on any calendar day (non leap year), i.e. birthdays are independent and identically distribution uniformly on $\{1, \dots, 365\}$. A is the number of ways to uniquely select n birthdays, with $|A| = 365 \cdot 364 \cdots (365 - (n - 1))$, and the sample space is $\{1, \dots, 365\}^n$. Therefore, the resulting probability is*

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{365!}{365^n(365 - n)!}.$$

For $n = 23$, $\mathbb{P}(A) = .4927$, so it is more likely than two individuals will have the same birthday.

Example 4(Mafia Game). *Suppose that n members of the mafia are in one room and simultaneously shoot another mafia member uniformly at random (possibly themselves). Let A = “Every member is shot”. Let $\Omega = \{(\omega_1, \dots, \omega_n)\}$ be the set of assignments of mafia members to the people they shoot, i.e. $\omega_i \in$*

$\{1, \dots, n\}$ is the target of the i -th mafia member. The event A occurs for all $\omega \in \Omega$ that are permutations. Therefore, the corresponding probability is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{n!}{n^n} \leq \frac{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}}{n^n} \leq \left(\sqrt{2\pi n e}\right) e^{-n}.$$

Hence, for large n , the event A is very unlikely.

Example 5(Mafia Survival). Under the setting of the Mafia Game, let $B = \text{"The first mafia member survives"}$. This time each shooter has $(n - 1)$ admissible targets for event B to occur. Therefore, the corresponding probability is

$$\Pr(B) = \frac{(n - 1)^n}{n^n} = \left(1 - \frac{1}{n}\right)^n \xrightarrow{n} e^{-1}.$$

This result can be further generalized (see Section 2.1) to show that probability of the first mafia member dying from exactly 3 bullets is

$$\binom{n}{3} \frac{1}{n} \frac{1}{n} \frac{1}{n} \left(1 - \frac{1}{n}\right)^{n-3} \xrightarrow{n} \frac{e^{-1}}{3!}.$$

Example 6(Multinomial Coefficients). Similar to the binomial coefficient $\binom{n}{k}$, given n elements and r numbers n_i , $i = 1, \dots, r$, with $\sum_{i=1}^r n_i = n$, the multinomial coefficient expresses the number of ways those n elements can be separated into r groups of size n_i . Proceeding iteratively, there are $\binom{n}{n_1}$ choices for the first group, $\binom{n-n_1}{n_2}$ choices for the second group, \dots , and $\binom{n_r}{n_r}$ choices for the r -th group. In total this provides

$$\binom{n}{n_1} \binom{n-n_1}{n_2} \cdots \binom{n_r}{n_r} = \frac{n!}{n_1! n_2! \cdots n_r!} \triangleq \binom{n}{n_1, \dots, n_r}$$

possible choices, where this last expression is the multinomial coefficient.

2 A FEW USEFUL RANDOM VARIABLES

Recall that a random variable $X : \Omega \rightarrow \mathbb{R}$ is called discrete if its range (i.e., the set of values that it can take) is a countable set. The PMF of X is a function $p_X : \mathbb{R} \rightarrow [0, 1]$, defined by $p_X(x) = \mathbb{P}(X = x)$, and completely determines the probability law of X .

The following are some important PMFs.

- (a) **Discrete uniform** with parameters a and b , where a and b are integers with $a < b$. Here,

$$p_X(k) = 1/(b - a + 1), \quad k = a, a + 1, \dots, b,$$

and $p_X(k) = 0$, otherwise.¹

- (b) **Bernoulli** with parameter p , where $0 \leq p \leq 1$. Here, $p_X(0) = p$, $p_X(1) = 1 - p$.
- (c) **Binomial** with parameters n and p , where $n \in \mathbb{N}$ and $p \in [0, 1]$. Here,

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n.$$

A binomial random variable with parameters n and p represents the number of heads observed in n independent tosses of a coin if the probability of heads at each toss is p .

- (d) **Geometric** with parameter p , where $0 < p \leq 1$. Here,

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \dots.$$

A geometric random variable with parameter p represents the number of independent tosses of a coin until heads are observed for the first time, if the probability of heads at each toss is p .

- (e) **Poisson** with parameter λ , where $\lambda > 0$. Here,

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \dots.$$

As will be seen shortly, a Poisson random variable can be thought of as a limiting case of a binomial random variable. Note that this is a legitimate PMF (i.e., it sums to one), because of the series expansion of the exponential function, $e^\lambda = \sum_{k=0}^{\infty} \lambda^k/k!$.

- (f) **Power law** with parameter α , where $\alpha > 0$. Here,

$$p_X(k) = \frac{1}{k^\alpha} - \frac{1}{(k+1)^\alpha}, \quad k = 1, 2, \dots.$$

An equivalent but more intuitive way of specifying this PMF is in terms of the formula

$$\mathbb{P}(X \geq k) = \frac{1}{k^\alpha}, \quad k = 1, 2, \dots.$$

¹In the remaining examples, the qualification “ $p_X(k) = 0$, otherwise,” will be omitted for brevity.

Note that when α is small, the “tail” $\mathbb{P}(X \geq k)$ of the distribution decays slowly (slower than an exponential) as k increases, and in some sense such a distribution has “heavy” tails.

Notation: Let us use the abbreviations $dU(a, b)$, $Ber(p)$, $Bin(n, p)$, $Geo(p)$, $Pois(\lambda)$, and $Pow(\alpha)$ to refer the above defined PMFs. We will use notation such as $X \stackrel{d}{=} dU(a, b)$ or $X \sim dU(a, b)$ as a shorthand for the statement that X is a discrete random variable whose PMF is uniform on (a, b) , and similarly for the other PMFs we defined. We will also use the notation $X \stackrel{d}{=} Y$ to indicate that two random variables have the same PMFs.

2.1 Poisson distribution as a limit of the binomial

To get a feel for the Poisson random variable, think of a binomial random variable with very small p and very large n . For example, consider the number of typos in a book with a total of n words, when the probability p that any one word is misspelled is very small (associate a word with a coin toss that results in a head when the word is misspelled), or the number of cars involved in accidents in a city on a given day (associate a car with a coin toss that results in a head when the car has an accident). Such random variables can be well modeled with a Poisson PMF.

More precisely, the Poisson PMF with parameter λ is a good approximation for a binomial PMF with parameters n and p , i.e.,

$$e^{-\lambda} \frac{\lambda^k}{k!} \approx \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

provided $\lambda = np$, n is large, and p is small. In this case, using the Poisson PMF may result in simpler models and calculations. For example, let $n = 100$ and $p = 0.01$. Then the probability of $k = 5$ successes in $n = 100$ trials is calculated using the binomial PMF as

$$\frac{100!}{95!5!} \cdot 0.01^5 (1 - 0.01)^{95} = 0.00290.$$

Using the Poisson PMF with $\lambda = np = 100 \cdot 0.01 = 1$, this probability is approximated by

$$e^{-1} \frac{1}{5!} = 0.00306.$$

Proposition 1. (Binomial convergence to Poisson) *Let us fix some $\lambda > 0$, and suppose that $X_n \stackrel{d}{=} \text{Bin}(n, \lambda/n)$, for every n . Let $X \stackrel{d}{=} \text{Pois}(\lambda)$. Then, as $n \rightarrow \infty$, the PMF of X_n converges to the PMF of X , in the sense that $\lim_{n \rightarrow \infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$, for any $k \geq 0$.*

Proof: We have

$$\mathbb{P}(X_n = k) = \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Fix k and let $n \rightarrow \infty$. We have, for $j = 1, \dots, k$,

$$\frac{n-k+j}{n} \rightarrow 1, \quad \left(1 - \frac{\lambda}{n}\right)^{-k} \rightarrow 1, \quad \left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}.$$

Thus, for any fixed k , we obtain

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n = k) = e^{-\lambda} \frac{\lambda^k}{k!} = \mathbb{P}(X = k),$$

as claimed. \square

3 JOINT, MARGINAL, AND CONDITIONAL PMFS

In most applications, one typically deals with several random variables at once. In this section, we introduce a few concepts that are useful in such a context.

3.1 Marginal PMFs

Consider two discrete random variables X and Y associated with the same experiment. The probability law of each one of them is described by the corresponding PMF, p_X or p_Y , called a **marginal** PMF. However, the marginal PMFs do not provide any information on possible relations between these two random variables. For example, suppose that the PMF of X is symmetric around the origin. If we have either $Y = X$ or $Y = -X$, the PMF of Y remains the same, and fails to capture the specifics of the dependence between X and Y .

As another example let $X \stackrel{d}{=} \text{Bin}(n, 1/2)$. Notice that then $Y = n - X$ also has $\text{Bin}(n, 1/2)$ as a PMF. At the same time $X + Y = n$ and this is something that cannot be inferred from PMF alone.

3.2 Joint PMFs

The statistical properties of two random variables X and Y are captured by a function $p_{X,Y} : \mathbb{R}^2 \rightarrow [0, 1]$, defined by

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y),$$

called the **joint PMF** of X and Y . We think of X and Y defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Namely, $X, Y : \Omega \rightarrow \mathbb{R}$. Then the event $A = \{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\}$ is well defined. Then $\mathbb{P}(X = x, Y = y)$ is simply $\mathbb{P}(A)$. So that we can talk about the probability of this event we also need to ensure that the event is measurable.

Exercise 1. Suppose X and Y are two discrete random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Is it the case that for every $x, y \in \mathbb{R}$ the event $A = \{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\}$ is measurable with respect to \mathcal{F} ? Either prove this or construct a counterexample.

From this point on we assume that the events A described above are measurable and will omit the measurability issues. Here and in the sequel, we will use the abbreviated notation $\mathbb{P}(X = x, Y = y)$ instead of the more precise notations $\mathbb{P}(\{X = x\} \cap \{Y = y\})$ or $\mathbb{P}(X = x \text{ and } Y = y)$. More generally, the PMF of finitely many discrete random variables, X_1, \dots, X_n on the same probability space is defined by

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n).$$

Sometimes, we define a vector random variable X , by letting $X = (X_1, \dots, X_n)$, in which case the joint PMF will be denoted simply as $p_X(x)$, where now the argument x is an n -dimensional vector.

The joint PMF of X and Y determines the probability of any event that can be specified in terms of the random variables X and Y . For example if A is the set of all pairs (x, y) that have a certain property, then

$$\mathbb{P}((X, Y) \in A) = \sum_{(x,y) \in A} p_{X,Y}(x, y).$$

In fact, we can calculate the marginal PMFs of X and Y by using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x, y), \quad p_Y(y) = \sum_x p_{X,Y}(x, y).$$

The formula for $p_X(x)$ can be verified using the calculation

$$p_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y p_{X,Y}(x, y),$$

where the second equality follows by noting that the event $\{X = x\}$ is the union of the countably many disjoint events $\{X = x, Y = y\}$, as y ranges over all the different values of Y . The formula for $p_Y(y)$ is verified similarly.

3.3 Conditional PMFs

Let X and Y be two discrete random variables, defined on the same probability space, with joint PMF $p_{X,Y}$. The **conditional PMF** of X given Y is a function $p_{X|Y}$, defined by

$$p_{X|Y}(x | y) = \mathbb{P}(X = x | Y = y), \quad \text{if } \mathbb{P}(Y = y) > 0;$$

if $\mathbb{P}(Y = y) = 0$, the value of $p_{X|Y}(y | x)$ is left undefined. Using the definition of conditional probabilities, we obtain

$$p_{X|Y}(x | y) = \frac{p_{X,Y}(x, y)}{p_Y(y)},$$

whenever $p_Y(y) > 0$.

More generally, if we have random variables X_1, \dots, X_n and Y_1, \dots, Y_m , defined on the same probability space, we define a conditional PMF by letting

$$\begin{aligned} p_{X_1, \dots, X_n | Y_1, \dots, Y_m}(x_1, \dots, x_n | y_1, \dots, y_m) \\ = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n | Y_1 = y_1, \dots, Y_m = y_m) \\ = \frac{p_{X_1, \dots, X_n, Y_1, \dots, Y_m}(x_1, \dots, x_n, y_1, \dots, y_m)}{p_{Y_1, \dots, Y_m}(y_1, \dots, y_m)}, \end{aligned}$$

whenever $p_{Y_1, \dots, Y_m}(y_1, \dots, y_m) > 0$. Again, if we define $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_m)$, the shorthand notation $p_{X|Y}(x | y)$ can be used.

Note that if $p_Y(y) > 0$, then $\sum_x p_{X|Y}(x | y) = 1$, where the sum is over all x in the range of the random variable X . Thus, the conditional PMF is essentially the same as an ordinary PMF, but with redefined probabilities that take into account the conditioning event $Y = y$. Visually, if we fix y , the conditional PMF $p_{X|Y}(x | y)$, viewed as a function of x is a “slice” of the joint PMF $p_{X,Y}$, renormalized so that its entries sum to one.

4 INDEPENDENCE OF RANDOM VARIABLES

We now define the important notion of independence of random variables. We start with a general definition that applies to all types of random variables, including discrete and continuous ones. We then specialize to the case of discrete random variables.

4.1 Independence of general random variables

Intuitively, two random variables are independent if any partial information on the realized value of one random variable does not change the distribution of the other. This notion is formalized in the following definition.

Definition 2. (Independence of random variables)

(a) Let X_1, \dots, X_n be random variables defined on the same probability space.

We say that these random variables are *independent* if the events $X_1 \in B_1, \dots, X_n \in B_n$ are independent for any Borel subsets B_1, \dots, B_n of the real line. Namely,

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \cdots \mathbb{P}(X_n \in B_n),$$

for any Borel subsets B_1, \dots, B_n .

(b) Let $\{X_s \mid s \in S\}$ be a collection of random variables indexed by the elements of a (possibly infinite) index set S . We say that these random variables are *independent* if for every finite subset $\{s_1, \dots, s_n\}$ of S , the random variables X_{s_1}, \dots, X_{s_n} are independent.

Verifying the independence of random variables using the above definition (which involves arbitrary Borel sets) is rather difficult. It turns out that one only needs to examine Borel sets of the form $(-\infty, x]$.

Proposition 2. Suppose that for every n , every x_1, \dots, x_n , and every finite subset $\{s_1, \dots, s_n\}$ of S , the events $\{X_{s_i} \leq x_i\}$, $i = 1, \dots, n$, are independent. Then, the random variables X_s , $s \in S$, are independent.

The proof is a simple application of Theorem 2 from Lecture 3 applied to a generating p -system $(-\infty, x]$.

Let us define the joint CDF of the random variables X_1, \dots, X_n by

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n).$$

In view of Proposition 2, independence of X_1, \dots, X_n is equivalent to the condition

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n), \quad \forall x_1, \dots, x_n.$$

Exercise 2. Consider a collection $\{A_s \mid s \in S\}$ of events, where S is a (possibly infinite) index set. Prove that the events A_s are independent if and only if the corresponding indicator functions I_{A_s} , $s \in S$, are independent random variables.

4.2 Independence of discrete random variables

For a finite number of discrete random variables, independence is equivalent to having a joint PMF which factors into a product of marginal PMFs.

Theorem 1. Let X and Y be discrete random variables defined on the same probability space. The following are equivalent.

- (a) The random variables X and Y are independent.
- (b) For any $x, y \in \mathbb{R}$, the events $\{X = x\}$ and $\{Y = y\}$ are independent.
- (c) For any $x, y \in \mathbb{R}$, we have $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.
- (d) For any $x, y \in \mathbb{R}$ such that $p_Y(y) > 0$, we have $p_{X|Y}(x \mid y) = p_X(x)$.

Proof: The fact that (a) implies (b) is immediate from the definition of independence, since recall that the sets consisting of one point $\{x\}, \{y\}$ are Borel sets.

That (b) implies (c), and (c) implies (d) is also an immediate consequence of our definitions. Let us show that (d) implies (c). For the case when $p_Y(y) > 0$, we have $p_{X,Y}(x, y) = p_{X|Y}(x \mid y)p_Y(y) = p_X(x)p_Y(y)$. When $p_Y(y) = 0$, we have also $p_X(x)p_Y(y) = 0$. So in order to show the identity we need $p_{X,Y}(x, y) = 0$. But $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) \leq \mathbb{P}(Y = y) = 0$, and we have verified that both parts equal zero.

We complete the proof by verifying that (c) implies (a). Suppose that X and Y are independent, and let A, B , be two Borel subsets of the real line. We then

have

$$\begin{aligned}
\mathbb{P}(X \in A, Y \in B) &= \sum_{x \in A, y \in B} \mathbb{P}(X = x, Y = y) \\
&= \sum_{x \in A, y \in B} p_{X,Y}(x, y) \\
&= \sum_{x \in A, y \in B} p_X(x)p_Y(y) \\
&= \left(\sum_{x \in A} p_X(x) \right) \left(\sum_{y \in B} p_Y(y) \right) \\
&= \mathbb{P}(X \in A) \mathbb{P}(Y \in B).
\end{aligned}$$

Since this is true for any Borel sets A and B , we conclude that X and Y are independent. \square

We note that Theorem 1 generalizes to the case of multiple, but finitely many, random variables. The generalization of conditions (a)-(c) should be obvious. As for condition (d), it can be generalized to a few different forms, one of which is the following: given any subset S_0 of the random variables under consideration, the conditional joint PMF of the random variables X_s , $s \in S_0$, given the values of the remaining random variables, is the same as the unconditional joint PMF of the random variables X_s , $s \in S_0$, as long as we are conditioning on an event with positive probability.

We finally note that functions $g(X)$ and $h(Y)$ of two independent random variables X and Y must themselves be independent. This should be expected on intuitive grounds: If X is independent from Y , then the information provided by the value of $g(X)$ should not affect the distribution of Y , and consequently should not affect the distribution of $h(Y)$. Observe that when X and Y are discrete, then $g(X)$ and $h(Y)$ are random variables (the required measurability conditions are satisfied) even if the functions g and h are not measurable (why?).

Theorem 2. Let X and Y be independent discrete random variables. Let g and h be some functions from \mathbb{R} into itself. Then, the random variables $g(X)$ and $h(Y)$ are independent.

The proof is left as an exercise.

4.3 Examples

Example. Let X_1, \dots, X_n be independent Bernoulli random variables with the same

parameter p . Then, the random variable X defined by $X = X_1 + \dots + X_n$ is binomial with parameters n and p . To see this, consider n independent tosses of a coin in which every toss has probability p of resulting in a one, and let X_i be the result of the i th coin toss. Then, X is the number of ones observed in n independent tosses, and is therefore a binomial random variable.

Example. Let X and Y be independent binomial random variables with parameters (n, p) and (m, p) , respectively. Then, the random variable Z , defined by $Z = X + Y$ is binomial with parameters $(n+m, p)$. To see this, consider $n+m$ independent tosses of a coin in which every toss has probability p of resulting in a one. Let X be the number of ones in the first n tosses, and let Y be the number of ones in the last m tosses. Then, Z is the number of ones in $n+m$ independent tosses, which is binomial with parameters $(n+m, p)$.

Example. Consider n independent tosses of a coin in which every toss has probability p of resulting in a one. Let X be the number of ones obtained, and let $Y = n - X$, which is the number of zeros. The random variables X and Y are not independent. For example, $\mathbb{P}(X = 0) = (1-p)^n$ and $\mathbb{P}(Y = 0) = p^n$, but $\mathbb{P}(X = 0, Y = 0) = 0 \neq \mathbb{P}(X = 0)\mathbb{P}(Y = 0)$. Intuitively, knowing that there was a small number of heads gives us information that the number of tails must be large.

However, in sharp contrast to the intuition from the preceding example, we obtain independence when the number of coin tosses is itself random, with a Poisson distribution. More precisely, let N be a Poisson random variable with parameter λ . We assume that X has conditional PMF $p_{X|N}(\cdot | n)$ is binomial with parameters n and p (representing the number of ones observed in n coin tosses), and define $Y = N - X$, which represents the number of zeros obtained. We have the following surprising result. An intuitive justification will have to wait until we consider the Poisson process, later in this course. The proof is left as an exercise.

Theorem 3. (Splitting of a Poisson random variable) *The random variables X and Y are independent. Moreover, $X \stackrel{d}{=} \text{Pois}(\lambda p)$ and $Y \stackrel{d}{=} \text{Pois}(\lambda(1-p))$.*

5 EXPECTED VALUES

5.1 Preliminaries: infinite sums

Consider a sequence $\{a_n\}$ of nonnegative real numbers and the infinite sum $\sum_{i=1}^{\infty} a_i$, defined as the limit, $\lim_{n \rightarrow \infty} \sum_{i=1}^n a_i$, of the partial sums. The infinite sum can be finite or infinite; in either case, it is well defined, as long as we allow the limit to be an extended real number. Furthermore, it can be verified that the value of the infinite sum is the same even if we reorder the elements of the sequence $\{a_n\}$ and carry out the summation according to this different order. Because the order of the summation does not matter, we can use the notation $\sum_{n \in \mathbb{N}} a_n$ for the infinite sum. More generally, if C is a countable set and $g : C \rightarrow [0, \infty)$ is a nonnegative function, we can use the notation $\sum_{x \in C} g(x)$, which is unambiguously even without specifying a particular order in which the values $g(x)$ are to be summed.

When we consider a sequence of nonpositive real numbers, the discussion remains the same, and infinite sums can be unambiguously defined. However, when we consider sequences that involve both positive and negative numbers, the situation is more complicated. In particular, the order at which the elements of the sequence are added can make a difference.

Example. Let $a_n = (-1)^n/n$. It can be verified that the limit $\lim_{n \rightarrow \infty} \sum_{i=1}^n a_i$ exists, and is finite, but that the elements of the sequence $\{a_n\}$ can be reordered to form a new sequence $\{b_n\}$ for which the limit of $\sum_{i=1}^n b_i$ does not exist.

In order to deal with the general case, we proceed as follows. Let S be a countable set, and consider a collection of real numbers a_s , $s \in S$. Let S_+ (respectively, S_-) be the set of indices s for which $a_s \geq 0$ (respectively, $a_s < 0$). Let $S_+ = \sum_{s \in S_+} a_s$ and $S_- = \sum_{s \in S_-} |a_s|$. We distinguish four cases.

- (a) If both S_+ and S_- are finite (or equivalently, if $\sum_{s \in S} |a_s| < \infty$), we say that the sum $\sum_{s \in S} a_s$ is **absolutely convergent** and is equal to $S_+ - S_-$. In this case, for every possible arrangement of the elements of S in a sequence $\{s_n\}$, we have

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n a_{s_i} = S^+ - S_-.$$

- (b) If $S_+ = \infty$ and $S_- < \infty$, the sum $\sum_{s \in S} a_s$ is not absolutely convergent; we define it to be equal to ∞ . In this case, for every possible arrangement of the elements of S in a sequence $\{s_n\}$, we have

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n a_{s_i} = \infty.$$

- (c) If $S_+ < \infty$ and $S_- = \infty$, the sum $\sum_{s \in S} a_s$ is not absolutely convergent; we define it to be equal to $-\infty$. In this case, for every possible arrangement of the elements of S in a sequence $\{s_n\}$, we have

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n a_{s_i} = -\infty.$$

- (d) If $S_+ = \infty$ and $S_- = \infty$, the sum $\sum_{s \in S} a_s$ is left undefined. In fact, in this case, different arrangements of the elements of S in a sequence $\{s_n\}$ will result into different or even nonexistent values of the limit

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n a_{s_i}.$$

To summarize, we consider a countable sum to be well defined in cases (a)-(c), and call it absolutely convergent only in case (a).

We close by recording a related useful fact. If we have a doubly indexed family of nonnegative numbers a_{ij} , $i, j \in \mathbb{N}$, and if either (i) the numbers are nonnegative, or (ii) the sum $\sum_{(i,j)} a_{ij}$ is absolutely convergent, then

$$\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij} = \sum_{(i,j) \in \mathbb{N}^2} a_{ij}. \quad (1)$$

More important, we stress that the first equality need not hold in the absence of conditions (i) or (ii) above.

5.2 Definition of the expectation

The PMF of a random variable X provides us with several numbers, the probabilities of all the possible values of X . It is often desirable to summarize this information in a single representative number. This is accomplished by the **expectation** of X , which is a weighted (in proportion to probabilities) average of the possible values of X .

As motivation, suppose you spin a wheel of fortune many times. At each spin, one of the numbers m_1, m_2, \dots, m_n comes up with corresponding probability p_1, p_2, \dots, p_n , and this is your monetary reward from that spin. What is the amount of money that you “expect” to get “per spin”? The terms “expect” and “per spin” are a little ambiguous, but here is a reasonable interpretation.

Suppose that you spin the wheel k times, and that k_i is the number of times that the outcome is m_i . Then, the total amount received is $m_1 k_1 + m_2 k_2 + \dots + m_n k_n$. The amount received per spin is

$$M = \frac{m_1 k_1 + m_2 k_2 + \dots + m_n k_n}{k}.$$

If the number of spins k is very large, and if we are willing to interpret probabilities as relative frequencies, it is reasonable to anticipate that m_i comes up a fraction of times that is roughly equal to p_i :

$$\frac{k_i}{k} \approx p_i, \quad i = 1, \dots, n.$$

Thus, the amount of money per spin that you “expect” to receive is

$$M = \frac{m_1 k_1 + m_2 k_2 + \dots + m_n k_n}{k} \approx m_1 p_1 + m_2 p_2 + \dots + m_n p_n.$$

Motivated by this example, we introduce the following definition.

Definition 3. (Expectation) We define the **expected value** (also called the **expectation** or the **mean**) of a discrete random variable X , with PMF p_X , as

$$\mathbb{E}[X] = \sum_x x p_X(x),$$

whenever the sum is well defined, and where the sum is taken over the countable set of values in the range of X .

Observe that $\mathbb{E}[X]$ is non-negative if X only takes non-negative values with positive probability. (Namely, $p_X(x) > 0$ implies $x \geq 0$).

5.3 Properties of the expectation

We start by pointing out an alternative formula for the expectation, and leave its proof as an exercise. In particular, if X can only take nonnegative integer values, then

$$\mathbb{E}[X] = \sum_{n \geq 0} \mathbb{P}(X > n). \tag{2}$$

Example. Using this formula, it is easy to give an example of a random variable for which the expected value is infinite. Consider $X \stackrel{d}{=} \text{Pow}(\alpha)$, where $\alpha \leq 1$. Then, it can

be verified, using the fact $\sum_{n=1}^{\infty} 1/n = \infty$, that $\mathbb{E}[X] = \sum_{n \geq 0} \frac{1}{n} = \infty$. On the other hand, if $\alpha > 1$, then $\mathbb{E}[X] < \infty$.

Here is another useful fact, whose proof is again left as an exercise.

Proposition 3. *Given a discrete random variable X and a function $g : \mathbb{R} \rightarrow \mathbb{R}$, we have*

$$\mathbb{E}[g(X)] = \sum_{\{x \mid p_X(x) > 0\}} g(x)p_X(x). \quad (3)$$

More generally, this formula remains valid given a vector $X = (X_1, \dots, X_n)$ of random variables with joint PMF $p_X = p_{X_1, \dots, X_n}$, and a function $g : \mathbb{R}^n \rightarrow \mathbb{R}$.

For example, suppose that X is a discrete random variable and consider the function $g : \mathbb{R} \rightarrow \mathbb{R}$ defined by $g(x) = x^2$. Let $Y = g(X)$. In order to calculate the expectation $\mathbb{E}[Y]$ according to Definition 2, we need to first find the PMF of Y , and then use the formula $\mathbb{E}[Y] = \sum_y y p_Y(y)$. However, according to Proposition 3, we can work directly with the PMF of X , and write $\mathbb{E}[Y] = \mathbb{E}[X^2] = \sum_x x^2 p_X(x)$.

The quantity $\mathbb{E}[X^2]$ is called the **second moment** of X . More generally, if $r \in \mathbb{N}$, the quantity $\mathbb{E}[X^r]$ is called the r th moment of X . Furthermore, $\mathbb{E}[(X - \mathbb{E}[X])^r]$ is called the r th **central moment** of X . The second central moment, $\mathbb{E}[(X - \mathbb{E}[X])^2]$ is called the **variance** of X , and is denoted by $\text{var}(X)$. The square root of the variance is called the **standard deviation** of X , and is often denoted by σ_X , or just σ . Note, that for every even r , the r th moment and the r th central moment are always nonnegative; in particular, the standard deviation is always well defined.

We continue with a few more important properties of expectations. In the sequel, notations such as $X \geq 0$ or $X = c$ mean that $X(\omega) \geq 0$ or $X(\omega) = c$, respectively, for all $\omega \in \Omega$. Similarly, a statement such as “ $X \geq 0$, almost surely” or “ $X \geq 0$, a.s.” means that $\mathbb{P}(X \geq 0) = 1$.

Proposition 4. Let X and Y be discrete random variables defined on the same probability space.

- (a) If $X \geq 0$, a.s., then $\mathbb{E}[X] \geq 0$.
- (b) If $X = c$, a.s., for some constant $c \in \mathbb{R}$, then $\mathbb{E}[X] = c$.
- (c) **Linearity of expectation.** For any $a, b \in \mathbb{R}$, we have $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ (as long as the sum $a\mathbb{E}[X] + b\mathbb{E}[Y]$ is well-defined).
- (d) If $\mathbb{E}[X]$ is finite, then $\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
- (e) For every $a \in \mathbb{R}$, we have $\text{var}(aX) = a^2\text{var}(X)$.
- (f) If X and Y are independent and have finite expectations, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ and $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$.
- (g) More generally, if X_1, \dots, X_n are independent and have finite expectations, then

$$\mathbb{E}\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n \mathbb{E}[X_i],$$

and

$$\text{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{var}(X_i).$$

Remark: We emphasize that property (c) does not require independence.

Proof: We only give the proof for the case where all expectations involved are well defined and finite, and leave it to the reader to verify that the results extend to the case where all expectations involved are well defined but possibly infinite.

Parts (a) and (b) are immediate consequences of the definitions. For part (c), we use the second part of Proposition 3, and then Eq. (1), we obtain

$$\begin{aligned} \mathbb{E}[aX + bY] &= \sum_{x,y} (ax + by)p_{X,Y}(x,y) \\ &= \sum_x \left(ax \sum_y p_{X,Y}(x,y) \right) + \sum_y \left(by \sum_x p_{X,Y}(x,y) \right) \\ &= a \sum_x x p_X(x) + b \sum_y y p_Y(y) \\ &= a\mathbb{E}[X] + b\mathbb{E}[Y]. \end{aligned}$$

For part (d), we have

$$\begin{aligned}\text{var}(X) &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2.\end{aligned}$$

where the second equality made use of property (c).

Part (e) follows easily from (d) and (c). For part (f), we apply Proposition 3 and then use independence to obtain

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x,y} xy p_{X,Y}(x,y) \\ &= \sum_{x,y} xy p_X(x)p_Y(y) \\ &= \left(\sum_x x p_X(x) \right) \left(\sum_y y p_Y(y) \right) \\ &= \mathbb{E}[X] \mathbb{E}[Y].\end{aligned}$$

Furthermore, using property (d), we have

$$\begin{aligned}\text{var}(X+Y) &= \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \\ &= \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY] - (\mathbb{E}[X])^2 - (\mathbb{E}[Y])^2 - 2\mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

Using the equality $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, the above expression becomes $\text{var}(X) + \text{var}(Y)$. The proof of part (g) is similar and is omitted. \square

Remark: The equalities in part (f) need not hold in the absence of independence. For example, consider a random variable X that takes either value 1 or -1 , with probability $1/2$. Then, $\mathbb{E}[X] = 0$, but $\mathbb{E}[X^2] = 1$. If we let $Y = X$, we see that $\mathbb{E}[XY] = \mathbb{E}[X^2] = 1 \neq 0 = (\mathbb{E}[X])^2$. Furthermore, $\text{var}(X+Y) = \text{var}(2X) = 4\text{var}(X)$, while $\text{var}(X) + \text{var}(Y) = 2$.

Exercise 3. Show that $\text{var}(X) = 0$ if and only if there exists a constant c such that $\mathbb{P}(X = c) = 1$.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MORE ON DISCRETE RANDOM VARIABLES AND THEIR EXPECTATIONS

Contents

1. Comments on expected values
2. Expected values of some common random variables
3. Covariance and correlation
4. Indicator variables and the inclusion-exclusion formula
5. Conditional expectations

1 COMMENTS ON EXPECTED VALUES

- (a) Recall that $\mathbb{E}[X]$ is well defined unless both sums $\sum_{x:x<0} xp_X(x)$ and $\sum_{x:x>0} xp_X(x)$ are infinite. Furthermore, $\mathbb{E}[X]$ is well-defined and finite if and only if both sums are finite. This is the same as requiring that

$$\mathbb{E}[|X|] = \sum_x |x| p_X(x) < \infty.$$

Random variables that satisfy this condition are called **integrable**.

- (b) Note that for any random variable X , $\mathbb{E}[X^2]$ is always well-defined (whether finite or infinite), because all the terms in the sum $\sum_x x^2 p_X(x)$ are nonnegative. If we have $\mathbb{E}[X^2] < \infty$, we say that X is **square integrable**.
- (c) Using the inequality $|x| \leq 1 + x^2$, we have $\mathbb{E}[|X|] \leq 1 + \mathbb{E}[X^2]$, which shows that a square integrable random variable is always integrable. Similarly, for every positive integer r , if $\mathbb{E}[|X|^r]$ is finite then it is also finite for every $l < r$ (fill details).

Exercise 1. Recall that the r -the central moment of a random variable X is $\mathbb{E}[(X - \mathbb{E}[X])^r]$. Show that if the r -th central moment of an almost surely non-negative random variable X is finite, then its l -th central moment is also finite for every $l < r$.

- (d) Because of the formula $\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, we see that: (i) if X is square integrable, the variance is finite; (ii) if X is integrable, but not square integrable, the variance is infinite; (iii) if X is not integrable, the variance is undefined.

2 EXPECTED VALUES OF SOME COMMON RANDOM VARIABLES

In this section, we use either the definition or the properties of expectations to calculate the mean and variance of a few common discrete random variables.

- (a) **Bernoulli(p)**. Let X be a Bernoulli random variable with parameter p . Then,

$$\begin{aligned}\mathbb{E}[X] &= 1 \cdot p + 0 \cdot (1 - p) = p, \\ \text{var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 1^2 \cdot p + 0^2 \cdot (1 - p) - p^2 = p(1 - p).\end{aligned}$$

- (b) **Binomial(n, p)**. Let X be a binomial random variable with parameters n and p . We note that X can be expressed in the form $X = \sum_{i=1}^n X_i$, where X_1, \dots, X_n are independent Bernoulli random variables with a common parameter p . It follows that

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = np.$$

Furthermore, using the independence of the random variables X_i , we have

$$\text{var}(X) = \sum_{i=1}^n \text{var}(X_i) = np(1 - p).$$

- (c) **Geometric(p)**. Let X be a geometric random variable with parameter p . We will use the formula $\mathbb{E}[X] = \sum_{n=0}^{\infty} \mathbb{P}(X > n)$. We observe that

$$\mathbb{P}(X > n) = \sum_{j=n+1}^{\infty} (1 - p)^{j-1} p = (1 - p)^n,$$

which implies that

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} (1-p)^n = \frac{1}{p}.$$

The variance of X is given by

$$\text{var}(X) = \frac{1-p}{p^2},$$

but we defer the derivation to a later section.

- (d) **Poisson(λ).** Let X be a Poisson random variable with parameter λ . A direct calculation yields

$$\begin{aligned}\mathbb{E}[X] &= e^{-\lambda} \sum_{n=0}^{\infty} n \frac{\lambda^n}{n!} \\ &= e^{-\lambda} \sum_{n=1}^{\infty} n \frac{\lambda^n}{n!} \\ &= e^{-\lambda} \sum_{n=1}^{\infty} \frac{\lambda^n}{(n-1)!} \\ &= \lambda e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \\ &= \lambda.\end{aligned}$$

The variance of X turns out to satisfy $\text{var}(X) = \lambda$, but we defer the derivation to a later section. We note, however, that the mean and the variance of a Poisson random variable are exactly what one would expect, on the basis of the formulae for the mean and variance of a binomial random variable, and taking the limit as $n \rightarrow \infty, p \rightarrow 0$, while keeping np fixed at λ .

- (e) **Power(α).** Let X be a random variable with a power law distribution with parameter α . We have

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} \mathbb{P}(X > k) = \sum_{k=0}^{\infty} \frac{1}{(k+1)^{\alpha}}.$$

If $\alpha \leq 1$, the expected value is seen to be infinite. For $\alpha > 1$, the sum is finite, but a closed form expression is not available; it is known as the Riemann zeta function, and is denoted by $\zeta(\alpha)$.

3 COVARIANCE AND CORRELATION

3.1 Covariance

The **covariance** of two square integrable random variables X and Y is denoted by $\text{cov}(X, Y)$, and is defined by

$$\text{cov}(X, Y) = \mathbb{E} \left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) \right].$$

When $\text{cov}(X, Y) = 0$, we say that X and Y are **uncorrelated**.

Note that, under the square integrability assumption, the covariance is always well-defined and finite. This is a consequence of the fact that $|XY| \leq (X^2 + Y^2)/2$, which implies that XY , as well as $(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])$, are integrable.

Roughly speaking, a positive or negative covariance indicates that the values of $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ obtained in a single experiment “tend” to have the same or the opposite sign, respectively. Thus, the sign of the covariance provides an important qualitative indicator of the relation between X and Y .

We record a few properties of the covariance, which are immediate consequences of its definition:

- (a) $\text{cov}(X, X) = \text{var}(X)$;
- (b) $\text{cov}(X, Y + a) = \text{cov}(X, Y)$;
- (c) $\text{cov}(X, Y) = \text{cov}(Y, X)$;
- (d) $\text{cov}(X, aY + bZ) = a \cdot \text{cov}(X, Y) + b \cdot \text{cov}(X, Z)$.

An alternative formula for the covariance is

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],$$

as can be verified by a simple calculation. Recall from last lecture that if X and Y are independent, we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, which implies that $\text{cov}(X, Y) = 0$. Thus, if X and Y are independent, they are also uncorrelated. However, the reverse is not true, as illustrated by the following example.

Example. Suppose that the pair of random variables (X, Y) takes the values $(1, 0)$, $(0, 1)$, $(-1, 0)$, and $(0, -1)$, each with probability $1/4$. Thus, the marginal PMFs of X and Y are symmetric around 0, and $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. Furthermore, for all possible value pairs (x, y) , either x or y is equal to 0, which implies that $XY = 0$ and $\mathbb{E}[XY] = 0$. Therefore,

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0,$$

and X and Y are uncorrelated. However, X and Y are not independent since, for example, a nonzero value of X fixes the value of Y to zero.

3.2 Variance of the sum of random variables

The covariance can be used to obtain a formula for the variance of the sum of several (not necessarily independent) random variables. In particular, if X_1, X_2, \dots, X_n are random variables with finite variance, we have

$$\text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2) + 2\text{cov}(X_1, X_2),$$

and, more generally,

$$\text{var} \sum_{i=1}^n X_i = \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{cov}(X_i, X_j).$$

This can be seen from the following calculation, where for brevity, we denote $\tilde{X}_i = X_i - \mathbb{E}[X_i]$:

$$\begin{aligned} \text{var} \sum_{i=1}^n X_i &= \mathbb{E} \left[\sum_{i=1}^n \tilde{X}_i \right]^2 \\ &= \mathbb{E} \left[\sum_{i=1}^n \sum_{j=1}^n \tilde{X}_i \tilde{X}_j \right] \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1}^n \mathbb{E}[\tilde{X}_i^2] + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbb{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{cov}(X_i, X_j). \end{aligned}$$

3.3 Correlation coefficient

The **correlation coefficient** $\rho(X, Y)$ of two random variables X and Y that have nonzero and finite variances is defined as

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.$$

(The simpler notation ρ will also be used when X and Y are clear from the context.) It may be viewed as a normalized version of the covariance $\text{cov}(X, Y)$.

Theorem 1. Let X and Y be discrete random variables with positive variance, and correlation coefficient equal to ρ .

(a) We have $-1 \leq \rho \leq 1$.

(b) We have $\rho = 1$ (respectively, $\rho = -1$) if and only if there exists a positive (respectively, negative) constant a such that $Y - \mathbb{E}[Y] = a(X - \mathbb{E}[X])$, with probability 1.

The proof of Theorem 1 relies on the Schwarz (or Cauchy-Schwarz) inequality, given below.

Proposition 1. (Cauchy-Schwarz inequality) For any two random variables, X and Y , with finite variance, we have

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Proof: Let us assume that $\mathbb{E}[Y^2] \neq 0$; otherwise, we have $Y = 0$ with probability 1, and hence $\mathbb{E}[XY] = 0$, so the inequality holds. We have

$$\begin{aligned} 0 &\leq \mathbb{E}\left[\left(X - \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}Y\right)^2\right] \\ &= \mathbb{E}\left[X^2 - 2\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}XY + \frac{(\mathbb{E}[XY])^2}{(\mathbb{E}[Y^2])^2}Y^2\right] \\ &= \mathbb{E}[X^2] - 2\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}\mathbb{E}[XY] + \frac{(\mathbb{E}[XY])^2}{(\mathbb{E}[Y^2])^2}\mathbb{E}[Y^2] \\ &= \mathbb{E}[X^2] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[Y^2]}, \end{aligned}$$

i.e., $\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$. □

Proof of Theorem 1:

(a) Let $\tilde{X} = X - \mathbb{E}[X]$ and $\tilde{Y} = Y - \mathbb{E}[Y]$. Using the Schwarz inequality, we get

$$(\rho(X, Y))^2 = \frac{(\mathbb{E}[\tilde{X}\tilde{Y}])^2}{\mathbb{E}[\tilde{X}^2]\mathbb{E}[\tilde{Y}^2]} \leq 1,$$

and hence $|\rho(X, Y)| \leq 1$.

(b) One direction is straightforward. If $\tilde{Y} = a\tilde{X}$, then

$$\rho(X, Y) = \frac{\mathbb{E}[\tilde{X}a\tilde{X}]}{\sqrt{\mathbb{E}[\tilde{X}^2]\mathbb{E}[(a\tilde{X})^2]}} = \frac{a}{|a|},$$

which equals 1 or -1 depending on whether a is positive or negative.

To establish the reverse direction, let us assume that $(\rho(X, Y))^2 = 1$, which implies that $\mathbb{E}[\tilde{X}^2]\mathbb{E}[\tilde{Y}^2] = (\mathbb{E}[\tilde{X}\tilde{Y}])^2$. Using the inequality established in the proof of Proposition 1, we conclude that the random variable

$$\tilde{X} - \frac{\mathbb{E}[\tilde{X}\tilde{Y}]}{\mathbb{E}[\tilde{Y}^2]}\tilde{Y}$$

is equal to zero, with probability 1. It follows that, with probability 1,

$$\tilde{X} = \frac{\mathbb{E}[\tilde{X}\tilde{Y}]}{\mathbb{E}[\tilde{Y}^2]}\tilde{Y} = \sqrt{\frac{\mathbb{E}[\tilde{X}^2]}{\mathbb{E}[\tilde{Y}^2]}}\rho(X, Y)\tilde{Y}.$$

Note that the sign of the constant ratio of \tilde{X} and \tilde{Y} is determined by the sign of $\rho(X, Y)$, as claimed. \square

Example. Consider n independent tosses of a coin with probability of a head equal to p . Let X and Y be the numbers of heads and of tails, respectively, and let us look at the correlation coefficient of X and Y . Here, we have $X + Y = n$, and also $\mathbb{E}[X] + \mathbb{E}[Y] = n$. Thus,

$$X - \mathbb{E}[X] = -Y - \mathbb{E}[Y].$$

We will calculate the correlation coefficient of X and Y , and verify that it is indeed equal to -1 .

We have

$$\begin{aligned} \text{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= -\mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= -\text{var}(X). \end{aligned}$$

Hence, the correlation coefficient is

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{-\text{var}(X)}{\sqrt{\text{var}(X)\text{var}(X)}} = -1.$$

4 INDICATOR VARIABLES AND THE INCLUSION-EXCLUSION FORMULA

Indicator functions are special discrete random variables that can be useful in simplifying certain derivations or proofs. In this section, we develop the inclusion-exclusion formula and apply it to a matching problem.

Recall that with every event A , we can associate its **indicator function**, which is a discrete random variable $I_A : \Omega \rightarrow \{0, 1\}$, defined by $I_A(\omega) = 1$ if $\omega \in A$, and $I_A(\omega) = 0$ otherwise. Note that $I_{A^c} = 1 - I_A$ and that $\mathbb{E}[I_A] = \mathbb{P}(A)$. These simple observations, together with the linearity of expectations turn out to be quite useful.

4.1 The inclusion-exclusion formula

Note that $I_{A \cap B} = I_A I_B$, for every $A, B \in \mathcal{F}$. Therefore,

$$\begin{aligned} I_{A \cup B} &= 1 - I_{(A \cup B)^c} = 1 - I_{A^c \cap B^c} = 1 - I_{A^c} I_{B^c} \\ &= 1 - (1 - I_A)(1 - I_B) = I_A + I_B - I_A I_B. \end{aligned}$$

Taking expectations of both sides, we obtain

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

an already familiar formula.

We now derive a generalization, known as the inclusion-exclusion formula. Suppose we have a collection of events A_j , $j = 1, \dots, n$, and that we are interested in the probability of the event $B = \bigcup_{j=1}^n A_j$. Note that

$$I_B = 1 - \prod_{j=1}^n (1 - I_{A_j}).$$

We begin with the easily verifiable fact that for any real numbers a_1, \dots, a_n , we have

$$\begin{aligned} \prod_{j=1}^n (1 - a_j) &= 1 - \sum_{1 \leq j \leq n} a_j + \sum_{1 \leq i < j \leq n} a_i a_j - \sum_{1 \leq i < j < k \leq n} a_i a_j a_k \\ &\quad + \cdots + (-1)^n a_1 \cdots a_n. \end{aligned}$$

We replace a_j by I_{A_j} , and then take expectations of both sides, to obtain

$$\begin{aligned}\mathbb{P}(B) &= \sum_{1 \leq j \leq n} \mathbb{P}(A_j) - \sum_{1 \leq i < j \leq n} \mathbb{P}(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} \mathbb{P}(A_i \cap A_j \cap A_k) \\ &\quad - \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap \cdots \cap A_n).\end{aligned}$$

4.2 The matching problem

Suppose that n people throw their hats in a box, where $n \geq 2$, and then each person picks one hat at random. (Each hat will be picked by exactly one person.) We interpret “at random” to mean that every permutation of the n hats is equally likely, and therefore has probability $1/n!$.

In an alternative model, we can visualize the experiment sequentially: the first person picks one of the n hats, with all hats being equally likely; then, the second person picks one of the remaining $n - 1$ remaining hats, with every remaining hat being equally likely, etc. It can be verified that under this second model, every permutation has probability $1/n!$, so the two models are equivalent.

We are interested in the mean, variance, and PMF of a random variable X , defined as the number of people that get back their own hat.¹ This problem is best approached using indicator variables.

For the i th person, we introduce a random variable X_i that takes the value 1 if the person selects his/her own hat, and takes the value 0 otherwise. Note that

$$X = X_1 + X_2 + \cdots + X_n.$$

Since $\mathbb{P}(X_i = 1) = 1/n$ and $\mathbb{P}(X_i = 0) = 1 - 1/n$, the mean of X_i is

$$\mathbb{E}[X_i] = 1 \cdot \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = \frac{1}{n},$$

which implies that

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] = n \cdot \frac{1}{n} = 1.$$

In order to find the variance of X , we first find the variances and covariances of the random variables X_i . We have

$$\text{var}(X_i) = \frac{1}{n} \left(1 - \frac{1}{n}\right).$$

¹For more results on various extensions of the matching problem, see L.A. Zager and G.C. Vergheze, “Caps and robbers: what can you expect?,” *College Mathematics Journal*, v. 38, n. 3, 2007, pp. 185-191.

For $i \neq j$, we have

$$\begin{aligned}
\text{cov}(X_i, X_j) &= \mathbb{E} \left[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j]) \right] \\
&= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \mathbb{E}[X_j] \\
&= \mathbb{P}(X_i = 1 \text{ and } X_j = 1) - \mathbb{P}(X_i = 1) \mathbb{P}(X_j = 1) \\
&= \mathbb{P}(X_i = 1) \mathbb{P}(X_j = 1 | X_i = 1) - \mathbb{P}(X_i = 1) \mathbb{P}(X_j = 1) \\
&= \frac{1}{n} \cdot \frac{1}{n-1} - \frac{1}{n^2} \\
&= \frac{1}{n^2(n-1)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{var}(X) &= \text{var} \left(\sum_{i=1}^n X_i \right) \\
&= \sum_{i=1}^n \text{var}(X_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{cov}(X_i, X_j) \\
&= n \cdot \frac{1}{n} \left(1 - \frac{1}{n} \right) + 2 \cdot \frac{n(n-1)}{2} \cdot \frac{1}{n^2(n-1)} \\
&= 1.
\end{aligned}$$

Finding the PMF of X is a little harder. Let us first dispense with some easy cases. We have $\mathbb{P}(X = n) = 1/n!$, because there is only one (out of the $n!$ possible) permutations under which every person receives their own hat. Furthermore, the event $X = n-1$ is impossible: if $n-1$ persons have received their own hat, the remaining person must also have received their own hat.

Let us continue by finding the probability that $X = 0$. Let A_i be the event that the i th person gets their own hat, i.e., $X_i = 1$. Note that the event $X = 0$ is the same as the event $\cap_i A_i^c$. Thus, $\mathbb{P}(X = 0) = 1 - \mathbb{P}(\cup_{i=1}^n A_i)$. Using the inclusion-exclusion formula, we have

$$\mathbb{P}(\cup_{i=1}^n A_i) = \sum_i \mathbb{P}(A_i) - \sum_{i < j} \mathbb{P}(A_i \cap A_j) + \sum_{i < j < k} \mathbb{P}(A_i \cap A_j \cap A_k) + \dots.$$

Observe that for every fixed distinct indices i_1, i_2, \dots, i_k , we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = \frac{1}{n} \cdot \frac{1}{n-1} \cdots \frac{1}{n-k+1} = \frac{(n-k)!}{n!}. \quad (1)$$

Thus,

$$\begin{aligned}\mathbb{P}(\cup_{i=1}^n A_i) &= n \cdot \frac{1}{n} - \frac{n}{2} \frac{(n-2)!}{n!} + \frac{n}{3} \frac{(n-3)!}{n!} + \cdots + (-1)^{n+1} \frac{n}{n} \frac{(n-n)!}{n!} \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1} \frac{1}{n!}.\end{aligned}$$

We conclude that

$$\mathbb{P}(X = 0) = \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!}. \quad (2)$$

Note that $\mathbb{P}(X = 0) \rightarrow e^{-1}$, as $n \rightarrow \infty$.

To conclude, let us now fix some integer r , with $0 < r \leq n-2$, and calculate $\mathbb{P}(X = r)$. The event $\{X = r\}$ can only occur as follows: for some subset S of $\{1, \dots, n\}$, of cardinality r , the following two events, B_S and C_S , occur:

- B_S : for every $i \in S$, person i receives their own hat;
- C_S : for every $i \notin S$, person i does not receive their own hat.

We then have

$$\{X = r\} = \bigcup_{S: |S|=r} B_S \cap C_S.$$

The events $B_S \cap C_S$ for different subsets S are disjoint. Furthermore, by symmetry, $\mathbb{P}(B_S \cap C_S)$ is the same for every S of cardinality r . Thus,

$$\begin{aligned}\mathbb{P}(X = r) &= \sum_{S: |S|=r} \mathbb{P}(B_S \cap C_S) \\ &= \binom{n}{r} \mathbb{P}(B_S) \mathbb{P}(C_S | B_S).\end{aligned}$$

Note that

$$\mathbb{P}(B_S) = \frac{(n-r)!}{n!},$$

by the same argument as in Eq. (1). Conditioned on the event that the r persons in the set S have received their own hats, the event C_S will materialize if and only if none of the remaining $n-r$ persons receive their own hat. But this is the same situation as the one analyzed when we calculated the probability that $X = 0$, except that n needs to be replaced by $n-r$. We conclude that

$$P(C_S | B_S) = \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-r} \frac{1}{(n-r)!}.$$

Putting everything together, we conclude that

$$\begin{aligned}\mathbb{P}(X = r) &= \binom{n}{r} \frac{(n-r)!}{n!} \left(\frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-r} \frac{1}{(n-r)!} \right) \\ &= \frac{1}{r!} \left(\frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-r} \frac{1}{(n-r)!} \right).\end{aligned}$$

Note that for each fixed r , the probability $\mathbb{P}(X = r)$ converges to $e^{-1}/r!$, as $n \rightarrow \infty$, which corresponds to a Poisson distribution with parameter 1. An intuitive justification is as follows. The random variables X_i are not independent (in particular, their covariance is nonzero). On the other hand, as $n \rightarrow \infty$, they are “approximately independent”. Furthermore, the success probability for each person is $1/n$, and the situation is similar to the one in our earlier proof that the binomial PMF approaches the Poisson PMF.

5 CONDITIONAL EXPECTATIONS

We have already defined the notion of a conditional PMF, $p_{X|Y}(\cdot | y)$, given the value of a random variable Y . Similarly, given an event A , we can define a conditional PMF $p_{X|A}$, by letting $p_{X|A}(x) = \mathbb{P}(X = x | A)$. In either case, the conditional PMF, as a function of x , is a bona fide PMF (a nonnegative function that sums to one). As such, it is natural to associate a (conditional) expectation to the (conditional) PMF.

Definition 1. Given an event A , such that $\mathbb{P}(A) > 0$, and a discrete random variable X , the **conditional expectation** of X given A is defined as

$$\mathbb{E}[X | A] = \sum_x x p_{X|A}(x),$$

provided that the sum is well-defined.

Note that the preceding also provides a definition for a conditional expectation of the form $\mathbb{E}[X | Y = y]$, for any y such that $p_Y(y) > 0$: just let A be the event $\{Y = y\}$, which yields

$$\mathbb{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y).$$

We note that the conditional expectation is always well defined when either the random variable X is nonnegative, or when the random variable X is integrable. In particular, whenever $\mathbb{E}[|X|] < \infty$, we also have $\mathbb{E}[|X| | Y = y] < \infty$,

for every y such that $p_Y(y) > 0$. To verify the latter assertion, note that for every y such that $p_Y(y) > 0$, we have

$$\sum_x |x| p_{X|Y}(x | y) = \sum_x |x| \frac{p_{X,Y}(x, y)}{p_Y(y)} \leq \frac{1}{p_Y(y)} \sum_x |x| p_X(x) = \frac{\mathbb{E}[|X|]}{p_Y(y)}.$$

The converse, however, is not true: it is possible that $\mathbb{E}[|X| | Y = y]$ is finite for every y that has positive probability, while $\mathbb{E}[|X|] = \infty$. This is left as an exercise.

The conditional expectation is essentially the same as an ordinary expectation, except that the original PMF is replaced by the conditional PMF. As such, the conditional expectation inherits all the properties of ordinary expectations (cf. Proposition 4 in the notes for Lecture 6).

5.1 The total expectation theorem

A simple calculation yields

$$\begin{aligned} \sum_y \mathbb{E}[X | Y = y] p_Y(y) &= \sum_y \sum_x x p_{X|Y}(x | y) p_Y(y) \\ &= \sum_y \sum_x x p_{X,Y}(x, y) \\ &= \mathbb{E}[X]. \end{aligned}$$

Note that this calculation is rigorous if X is nonnegative or integrable.

Suppose now that $\{A_i\}$ is a countable family of disjoint events that forms a partition of the probability space Ω . Define a random variable Y by letting $Y = i$ if and only if A_i occurs. Then, $p_Y(i) = \mathbb{P}(A_i)$, and $\mathbb{E}[X | Y = i] = \mathbb{E}[X | A_i]$, which yields

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X | A_i] \mathbb{P}(A_i).$$

Example. (The mean of the geometric.) Let X be a random variable with parameter p , so that $p_X(k) = (1-p)^{k-1}p$, for $p \in \mathbb{N}$. We first observe that the geometric distribution is memoryless: for $k \in \mathbb{N}$, we have

$$\begin{aligned} \mathbb{P}(X - 1 = k | X > 1) &= \frac{\mathbb{P}(X = k + 1, X > 1)}{\mathbb{P}(X > 1)} \\ &= \frac{\mathbb{P}(X = k + 1)}{\mathbb{P}(X > 1)} \\ &= \frac{(1-p)^k p}{1-p} = (1-p)^{k-1} p \\ &= \mathbb{P}(X = k). \end{aligned}$$

In words, in a sequence of repeated i.i.d., trials, given that the first trial was a failure, the distribution of the remaining trials, $X - 1$, until the first success is the same as the unconditional distribution of the number of trials, X , until the first success. In particular, $\mathbb{E}[X - 1 \mid X > 1] = \mathbb{E}[X]$.

Using the total expectation theorem, we can write

$$\mathbb{E}[X] = \mathbb{E}[X \mid X > 1]\mathbb{P}(X > 1) + \mathbb{E}[X \mid X = 1]\mathbb{P}(X = 1) = (1 + \mathbb{E}[X])(1 - p) + 1 \cdot p.$$

We solve for $\mathbb{E}[X]$, and find that $\mathbb{E}[X] = 1/p$.

Similarly,

$$\mathbb{E}[X^2] = \mathbb{E}[X^2 \mid X > 1]\mathbb{P}(X > 1) + \mathbb{E}[X^2 \mid X = 1]\mathbb{P}(X = 1).$$

Note that

$$\mathbb{E}[X^2 \mid X > 1] = \mathbb{E}[(X - 1)^2 \mid X > 1] + \mathbb{E}[2(X - 1) + 1 \mid X > 1] = \mathbb{E}[X^2] + (2/p) + 1.$$

Thus,

$$\mathbb{E}[X^2] = (1 - p)(\mathbb{E}[X^2] + (2/p) + 1) + p,$$

which yields

$$\mathbb{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}.$$

We conclude that

$$\text{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1 - p}{p^2}.$$

Example. Suppose we flip a biased coin N times, independently, where N is a Poisson random variable with parameter λ . The probability of heads at each flip is p . Let X be the number of heads, and let Y be the number of tails. Then,

$$\mathbb{E}[X \mid N = n] = \sum_{m=0}^{\infty} m\mathbb{P}(X = m \mid N = n) = \sum_{m=0}^n m \binom{n}{m} p^m (1-p)^{n-m}.$$

But X is just the expected number of heads in n trials, so that $\mathbb{E}[X \mid N = n] = np$.

Let us now calculate $\mathbb{E}[N \mid X = m]$. We have

$$\begin{aligned} \mathbb{E}[N \mid X = m] &= \sum_{n=0}^{\infty} n\mathbb{P}(N = n \mid X = m) \\ &= \sum_{n=m}^{\infty} n \frac{\mathbb{P}(N = n, X = m)}{\mathbb{P}(X = m)} \\ &= \sum_{n=m}^{\infty} n \frac{\mathbb{P}(X = m \mid N = n)\mathbb{P}(N = n)}{\mathbb{P}(X = m)} \\ &= \sum_{n=m}^{\infty} n \frac{\frac{n}{m} p^m (1-p)^{n-m} (\lambda^n / n!) e^{-\lambda}}{\mathbb{P}(X = m)}. \end{aligned}$$

Recall that $X \stackrel{d}{=} \text{Pois}(\lambda p)$, so that $\mathbb{P}(X = m) = e^{-\lambda p}(\lambda p)^m/m!$. Thus, after some cancellations, we obtain

$$\begin{aligned}\mathbb{E}[N | X = m] &= \sum_{n=m}^{\infty} n \frac{(1-p)^{n-m} \lambda^{n-m} e^{-\lambda(1-p)}}{(n-m)!} \\ &= \sum_{n=m}^{\infty} (n-m) \frac{(1-p)^{n-m} \lambda^{n-m} e^{-\lambda(1-p)}}{(n-m)!} \\ &\quad + m \sum_{n=m}^{\infty} \frac{(1-p)^{n-m} \lambda^{n-m} e^{-\lambda(1-p)}}{(n-m)!} \\ &= \lambda(1-p) + m.\end{aligned}$$

A faster way of obtaining this result is as follows. From Theorem 3 in the notes for Lecture 6, we have that X and Y are independent, and that Y is Poisson with parameter $\lambda(1-p)$. Therefore,

$$\mathbb{E}[N | X = m] = \mathbb{E}[X | X = m] + \mathbb{E}[Y | X = m] = m + \mathbb{E}[Y] = m + \lambda(1-p).$$

Exercise. (Simpson's "paradox") Let S be an event and X, Y discrete random variables, all defined on a common probability space. Show that

$$\mathbb{P}[S|X = 0, Y = y] > \mathbb{P}[S|X = 1, Y = y] \quad \forall y$$

does not imply

$$\mathbb{P}[S|X = 0] \geq \mathbb{P}[S|X = 1].$$

Thus in a clinical trial comparing two treatments (indexed by X) a drug can be more successful on each group of patients (indexed by Y) yet be less successful overall.

5.2 The conditional expectation as a random variable

Let X and Y be two discrete random variables. For any fixed value of y , the expression $\mathbb{E}[X | Y = y]$ is a real number, which however depends on y , and can be used to define a function $\phi : \mathbb{R} \rightarrow \mathbb{R}$, by letting $\phi(y) = \mathbb{E}[X | Y = y]$. Consider now the random variable $\phi(Y)$; this random variable takes the value $\mathbb{E}[X | Y = y]$ whenever Y takes the value y , which happens with probability $\mathbb{P}(Y = y)$. This random variable will be denoted as $\mathbb{E}[X | Y]$. (Strictly speaking, one needs to verify that this is a measurable function, which is left as an exercise.)

Example. Let us return to the last example and find $\mathbb{E}[X | N]$ and $\mathbb{E}[N | X]$. We found that $\mathbb{E}[X | N = n] = np$. Thus $\mathbb{E}[X | N] = Np$, i.e., it is a random variable that takes the value np with probability $\mathbb{P}(N = n) = (\lambda^n/n!)e^{-\lambda}$. We found that $\mathbb{E}[N | X = m] = \lambda(1-p) + m$. Thus $\mathbb{E}[N | X] = \lambda(1-p) + X$.

Note further that

$$\mathbb{E}[\mathbb{E}[X | N]] = \mathbb{E}[Np] = \lambda p = \mathbb{E}[X],$$

and

$$\mathbb{E}[\mathbb{E}[N | X]] = \lambda(1 - p) + \mathbb{E}[X] = \lambda(1 - p) + \lambda p = \lambda = \mathbb{E}[N].$$

This is not a coincidence; the equality $\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]$ is always true, as we shall now see. In fact, this is just the total expectation theorem, written in more abstract notation.

Theorem 2. *Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a measurable function such that $Xg(Y)$ is either nonnegative or integrable. Then,*

$$\mathbb{E} \mathbb{E}[X | Y]g(Y) = \mathbb{E}[Xg(Y)].$$

In particular, by letting $g(y) = 1$ for all y , we obtain $\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]$.

Proof: We have

$$\begin{aligned} \mathbb{E} \mathbb{E}[X | Y]g(Y) &= \sum_y \mathbb{E}[X | Y = y]g(y)p_Y(y) \\ &= \sum_y \sum_x x p_{X|Y}(x | y)g(y)p_Y(y) \\ &= \sum_{x,y} x g(y) p_{X,Y}(x, y) = \mathbb{E}[Xg(Y)]. \end{aligned}$$

□

The formula in Theorem 2 can be rewritten in the form

$$\mathbb{E}[(\mathbb{E}[X | Y] - X)g(Y)] = 0. \quad (3)$$

Here is an interpretation. We can think of $\mathbb{E}[X | Y]$ as an estimate of X , on the basis of Y , and $\mathbb{E}[X | Y] - X$ as an estimation error. The above formula says that the estimation error is uncorrelated with every function of the original data.

Equation (3) can be used as the basis for an abstract definition of conditional expectations. Namely, we define the conditional expectation as a random variable of the form $\phi(Y)$, where ϕ is a measurable function, that has the property

$$\mathbb{E}[(\phi(Y) - X)g(Y)] = 0,$$

for every measurable function g . The merits of this definition is that it can be used for all kinds of random variables (discrete, continuous, mixed, etc.). However, for this definition to be sound, there are two facts that need to be verified:

- (a) Existence: It turns out that as long as X is integrable, a function ϕ with the above properties is guaranteed to exist. We already know that this is the case for discrete random variables: the conditional expectation as defined in the beginning of this section does have the desired properties. For general random variables, this is a nontrivial and deep result. It will be revisited later in this course.
- (b) Uniqueness: It turns out that there is essentially only one function ϕ with the above properties. More precisely, any two functions with the above properties are equal with probability 1.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

ABSTRACT INTEGRATION — I**Contents**

1. Preliminaries
2. The main result
3. The limitations of the Riemann integral
4. The integral of a nonnegative simple function
5. The integral of a nonnegative function
6. The general case

The material in these notes can be found in practically every textbook that includes basic measure theory, although the order with which various properties are proved can be somewhat different in different sources.

1 PRELIMINARIES

The objective of these notes is to define the integral $\int g d\mu$ [sometimes also denoted $\int g(\omega) d\mu(\omega)$] of a measurable function $g : \Omega \rightarrow \bar{\mathbb{R}}$, defined on a measure space $(\Omega, \mathcal{F}, \mu)$. We remind that $\bar{\mathbb{R}}$ is the set of real values extended by ∞ and $-\infty$.

Special cases:

- (a) If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $X : \Omega \rightarrow \bar{\mathbb{R}}$ is measurable (i.e., an extended-valued random variable), the integral $\int X d\mathbb{P}$ is also denoted $\mathbb{E}[X]$, and is called the expectation of X .
- (b) If we are dealing with the measure space $(\mathbb{R}, \mathcal{B}, \lambda)$, where \mathcal{B} is the Borel σ -field and λ is the Lebesgue measure, the integral $\int g d\lambda$ is often denoted as $\int g(x) dx$, and is meant to be a generalization of the usual integral encountered in calculus.

The program: We will define the integral $\int g d\mu$ for progressively general classes of measurable functions:

- (a) Finite nonnegative functions g that take finitely many values (“simple functions”). In this case, the integral is just a suitably weighted sum of the values of g . When g is in fact a random variable, we recognize it as a discrete random variable since its range is finite.
- (b) Nonnegative functions g . Here, the integral will be defined by approximating g from below by a sequence of simple functions.
- (c) General functions g . This is done by decomposing g in the form $g = g_+ - g_-$, where g_+ and g_- are nonnegative functions, and letting $\int g d\mu = \int g_+ d\mu - \int g_- d\mu$.

We will be focusing on the integral over the entire set Ω . The integral over a (measurable) subset B of Ω is defined by letting

$$\int_B g d\mu = \int (1_B g) d\mu.$$

Here 1_B is the indicator function of the set B that

$$(1_B g)(\omega) = \begin{cases} g(\omega), & \text{if } \omega \in B, \\ 0, & \text{if } \omega \notin B. \end{cases}$$

Throughout, we will use the term “almost everywhere” to mean “for all ω outside a zero-measure subset of Ω .” For the special case of probability measures, we will often use the alternative terminology “almost surely,” or “a.s.” for short. Thus, if X and Y are random variables, we have $X = Y$, a.s., if and only if $\mathbb{P}(X = Y) = \mathbb{P}(\{\omega : X(\omega) \neq Y(\omega)\}) = 0$.

In the sequel, an inequality $g \leq h$ between two functions will be interpreted as “ $g(\omega) \leq h(\omega)$, for all $\omega \in \Omega$.” Similarly, “ $g \leq h$, a.e.” means that $g(\omega) \leq h(\omega)$, for all ω outside a zero-measure set.” The notation “ $g_n \uparrow g$ ” will mean that for every ω , the sequence $g_n(\omega)$ is monotone nondecreasing and converges to $g(\omega)$. Finally, “ $g_n \uparrow g$, a.e.” will mean that the monotonic convergence of $g_n(\omega)$ to $g(\omega)$ holds for all ω outside a zero-measure set.

2 THE MAIN RESULT

Once the construction is carried out, integrals of nonnegative functions will always be well-defined. For general functions, integrals will be left undefined only when an expression of the form $-\infty$ is encountered.

The following properties will turn out to be true, whenever the integrals or expectations involved are well-defined. On the left, we show the general version; on the right, we show the same property, specialized to the case of probability measures. In property 8, the convention $\infty = 0$ will be in effect.

- | | |
|---|--|
| 1. $\int 1_B d\mu = \mu(B)$ | $\mathbb{E}[1_B] = \mathbb{P}(B)$ |
| 2. $g \geq 0 \Rightarrow \int g d\mu \geq 0$ | $X \geq 0 \Rightarrow \mathbb{E}[X] \geq 0$ |
| 3. $g = 0, \text{a.e.} \Rightarrow \int g d\mu = 0$ | $X = 0, \text{ a.s.} \Rightarrow \mathbb{E}[X] = 0$ |
| 4. $g \leq h \Rightarrow \int g d\mu \leq \int h d\mu$
(assuming both $\int g$ and $\int h$ exist) | $X \leq Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$ |
| 4'. $g \leq h, \text{a.e.} \Rightarrow \int g d\mu \leq \int h d\mu$ | $X \leq Y, \text{a.s.} \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$ |
| 5. $g = h, \text{a.e.} \Rightarrow \int g d\mu = \int h d\mu$
and both \int exist or do not exist simultaneously | $X = Y, \text{a.s.} \Rightarrow \mathbb{E}[X] = \mathbb{E}[Y]$ |
| 6. $[g \geq 0, \text{a.e., and} \int g d\mu = 0] \Rightarrow g = 0, \text{a.e.}$ | $[X \geq 0, \text{ a.s., and} \mathbb{E}[X] = 0] \Rightarrow X = 0, \text{a.s.}$ |
| 7. $\int (g + h) d\mu = \int g d\mu + \int h d\mu$
(assuming RHS is well-defined) | $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ |
| 8. $\int (ag) d\mu = a \int g d\mu$ | $\mathbb{E}[aX] = a\mathbb{E}[X]$ |
| 9. $0 \leq g_n \uparrow g \Rightarrow \int g_n d\mu \uparrow \int g d\mu$ | $0 \leq X_n \uparrow X, \Rightarrow \mathbb{E}[X_n] \uparrow \mathbb{E}[X]$ |
| 9'. $0 \leq g_n \uparrow g, \text{a.e.} \Rightarrow \int g_n d\mu \uparrow \int g d\mu$ | $0 \leq X_n \uparrow X, \text{a.s.} \Rightarrow \mathbb{E}[X_n] \uparrow \mathbb{E}[X]$ |
| 10. $g \geq 0 \Rightarrow \nu(B) = \int_B g d\mu$ is a measure | $[f \geq 0 \text{ and } \int f d\mathbb{P} = 1]$
$\Rightarrow \nu(B) = \int_B f d\mathbb{P}$ is a probability measure |

Property 7, the linearity of expectations is central. Also, Property 9, and its generalization, property 9' is known as the Monotone Convergence Theorem (MCT), and is a cornerstone of integration theory.

3 THE LIMITATIONS OF THE RIEMANN INTEGRAL

Before proceeding, it is worth understanding why the traditional integral encountered in calculus is not adequate for our purposes. Let us recall the definition of the (Riemann) integral $\int_a^b g(x) dx$. We subdivide the interval $[a, b]$ using a finite sequence $\sigma = (x_1, x_2, \dots, x_n)$ of points that satisfy $a = x_1 < x_2 <$

$\dots < x_n = b$, and define

$$U(\sigma) = \sum_{i=1}^{n-1} \left(\max_{x_i \leq x < x_{i+1}} g(x) \right) \cdot (x_{i+1} - x_i),$$

$$L(\sigma) = \sum_{i=1}^{n-1} \left(\min_{x_i \leq x < x_{i+1}} g(x) \right) \cdot (x_{i+1} - x_i).$$

Thus, $U(\sigma)$ and $L(\sigma)$ are approximations of the “area under the curve from above and from below, respectively. We say that the integral $\int_a^b g(x) dx$ is well-defined, and equal to a number if

$$\limsup_{\sigma} L(\sigma) = \liminf_{\sigma} U(\sigma) = c,$$

where \limsup and \liminf are taken with respect to sequences of partitions (x_1, x_2, \dots, x_n) whose resolution $\max_{0 \leq i \leq n-1} (x_{i+1} - x_i)$ converges to zero. In this case, we also say that g is Riemann-integrable over $[a, b]$. Intuitively, we want the upper and lower approximants $U(\sigma)$ and $L(\sigma)$ to agree, in the limit of very fine subdivisions of the interval $[a, b]$.

It is known that if $f : \mathbb{R} \rightarrow \mathbb{R}$ is Riemann-integrable over every interval $[a, b]$, then f is continuous almost everywhere (i.e., there exists a set of Lebesgue measure zero, such that f is continuous at every $x \notin S$). This is a severe limitation on the class of Riemann-integrable functions.

Example. Let Q be the set of rational numbers $[0, 1]$. Let $g = 1_Q$. For any $\sigma = (x_1, x_2, \dots, x_n)$, and every i , the interval $[x_i, x_{i+1})$ contains a rational number, and also an irrational number. Thus $\max_{x_i \leq x < x_{i+1}} g(x) = 1$ and $\min_{x_i \leq x < x_{i+1}} g(x) = 0$. It follows that $U(\sigma) = 1$ and $L(\sigma) = 0$, for all σ , and $\sup_{\sigma} L(\sigma) \neq \inf_{\sigma} U(\sigma)$. Therefore 1_Q is not Riemann integrable. On the other hand if we consider a uniform distribution over $[0, 1]$, and the binary random variable Q , we have $\mathbb{P}(1_Q = 1) = 0$, and we would like to be able to say $\mathbb{E}[1_Q] = \int_{[0,1]} 1_Q(x) dx = 0$. This indicates that a different definition is in order.

4 THE INTEGRAL OF A NONNEGATIVE SIMPLE FUNCTION

A function $g : \Omega \rightarrow \mathbb{R}$ is called **simple** if it is measurable, finite- and takes only finitely many different values. In particular, a simple function can be written as

$$g(\omega) = \sum_{i=1}^k a_i 1_{A_i}(\omega), \quad \forall \omega \in \Omega, \tag{1}$$

where k is a (finite) nonnegative integer, the coefficients $\in \mathbb{R}$ are real values and the A_i are measurable sets.

Note that a simple function can have several representations of the form (1). For example, $1_{[0,2]}$ and $1_{[0,1]} + 1_{(1,2]}$ are two representations of the same function. For another example, note that $1_{[0,2]} + 1_{[1,2]} = 1_{[0,1]} + 2 \cdot 1_{[1,2]}$. On the other hand, if we require the A_i to be distinct and the sets to form a partition of Ω , it is not hard to see that there is only one possible representation, which we will call the **canonical** representation. More concretely, in the canonical representation, we let $\{a_1, \dots, a_k\}$ be the range of g , where the a_i are distinct, and $A_i = \{\omega \mid g(\omega) = a_i\}$.

Definition 1. If g is a simple function, of the form (1), its integral is defined by

$$\int g d\mu = \sum_{i=1}^k a_i \mu(A_i).$$

(If $a_i = 0$ and $\mu(A_i) = \infty$, we assume $a_i \mu(A_i) = 0$.)

Before continuing, we need to make sure that Definition 1 is sound, in the following sense. If we consider two alternative representations of the same simple function, we need to ensure that the resulting value of μ is the same. Technically, we need to show the following:

$$\text{if } \sum_{i=1}^k a_i 1_{A_i} = \sum_{i=1}^m b_i 1_{B_i}, \text{ then } \sum_{i=1}^k a_i \mu(A_i) = \sum_{i=1}^m b_i \mu(B_i).$$

This is left as an exercise for the reader.

Example. We have $1_{[0,2]} + 1_{[1,2]} = 1_{[0,1]} + 2 \cdot 1_{[1,2]}$. The first representation leads to $\mu([0,2]) + \mu([1,2])$, the second to $\mu([0,1]) + 2\mu([1,2])$. Using the fact $\mu([0,2]) = \mu([0,1]) + \mu([1,2])$ (finite additivity), we see that the two values are indeed equal.

For the case where the underlying measure is a probability measure, a simple function $X : \Omega \rightarrow \mathbb{R}$ is called a simple random variable, and its integral $\int X d\mathbb{P}$ is also denoted as $\mathbb{E}[X]$. We then have

$$\mathbb{E}[X] = \sum_{i=1}^k a_i \mathbb{P}(A_i).$$

If the coefficients a_i are distinct and exhaust all the possible values of X , then

by taking $A_i = \{\omega \mid X(\omega) = a_i\}$, we obtain

$$\mathbb{E}[X] = \sum_{i=1}^k a_i \mathbb{P}(\{\omega \mid X(\omega) = a_i\}) = \sum_{i=1}^k a_i \mathbb{P}(X = a_i),$$

which agrees with the elementary definition $\mathbb{E}[X]$ for discrete random variables.

Note, for future reference, that the sum or difference of two simple functions is also simple.

4.1 Verification of various properties for the case of simple functions

For the various properties listed in Section 2, we will use the shorthand “property S-A” and “property N-A”, to refer to “property A for the special case of nonnegative simple functions” and “property A for nonnegative measurable functions,” respectively.

We note a few immediate consequences of the definition. For \mathcal{F} , the function 1_B is simple and $\int 1_B d\mu = \mu(B)$, which verifies **property 1**. In particular, when Q is the set of rational numbers and Lebesgue measure, we have $\int 1_Q d\mu = \mu(Q) = 0$, as desired. Note that a nonnegative simple function has a representation of the form (1) with all nonnegative. It follows that $\int g d\mu \geq 0$, which verifies **property S-2**.

Suppose now that a simple function satisfies $\int g d\mu = 0$, a.e. Then, it has a canonical representation of the form $g = \sum_{i=1}^k a_i 1_{A_i}$, where $\mu(A_i) = 0$, for every i for which $a_i \neq 0$. Definition 1 implies that $\int g d\mu = 0$, which verifies **property S-3**.

Let us now verify the linear **property S-7**. Let g and h be nonnegative simple functions. Using canonical representations, we can write

$$g = \sum_{i=1}^k a_i 1_{A_i}, \quad h = \sum_{j=1}^m b_j 1_{B_j},$$

where the disjoint sets A_i or the sets B_j form a partition of Ω . Then, the sets $A_i \cap B_j$ are also disjoint, and form a partition of Ω . We have

$$g + h = \sum_{i=1}^k \sum_{j=1}^m (a_i + b_j) 1_{A_i \cap B_j}.$$

Therefore,

$$\begin{aligned}
\int (g + h) d\mu &= \sum_{i=1}^k \sum_{j=1}^m (a_i + b_j) \mu(A_i \cap B_j) \\
&= \sum_{i=1}^k a_i \sum_{j=1}^m \mu(A_i \cap B_j) + \sum_{j=1}^m b_j \sum_{i=1}^k \mu(A_i \cap B_j) \\
&= \sum_{i=1}^k a_i \mu(A_i) + \sum_{j=1}^m b_j \mu(B_j) \\
&= \int g d\mu + \int h d\mu.
\end{aligned}$$

(The first and fourth equalities follow from Definition 1. The third equality made use of finite additivity for the sets A_i or the sets B_j form a partition of Ω .)

Property S-8 is an immediate consequence of Definition 1. We only need to be careful for the case where $d\mu = \infty$ and $a = 0$. We have $g = 0$ and so $\int(ga)d\mu = 0$. On the other hand $\int gd\mu = 0 \cdot \infty$. We have agreed to assume that $0 \cdot \infty = 0$ and under this convention the identity is verified.

By combining properties S-7 and S-8, with -1 we see that, for simple functions g and h we have $\int(g - h)d\mu = \int g d\mu - \int h d\mu$.

We now verify **property S-4'**, which also implies **property S-4** as a special case. Suppose that $g \leq h$, a.e. We then have $g = g + q$, for a simple function q such that $q \geq 0$, a.e. In particular $q = q_+ - q_-$, where $q_+ \geq 0$, $q_- \geq 0$, and $q_- = 0$, a.e. Thus $h - g = g + q_+ - q_-$. Note that q , q_+ , and q_- are all simple functions. Using the linearity property S-7, and then properties S-3, S-2, we obtain

$$\int h d\mu = \int g d\mu + \int q_+ d\mu - \int q_- d\mu = \int g d\mu + \int q_+ d\mu \geq \int g d\mu.$$

We next verify **property S-5**. If $g = h$, a.e., then we have both $h \leq g$, a.e., and $g \leq h$, a.e. Thus, $\int g d\mu \leq \int h d\mu$, and $\int g d\mu \geq \int h d\mu$, which implies that $\int g d\mu = \int h d\mu$.

We finally verify **property S-6**. Suppose that $g \geq 0$, a.e., and $\int g d\mu = 0$. We write $g = g_+ - g_-$, where $g_+ \geq 0$ and $g_- \geq 0$. Then, $g_- = 0$, a.e., and $\int g_- d\mu = 0$. Thus, using property S-7, $\int g_+ d\mu = \int g d\mu + \int g_- d\mu = 0$. Note that g_+ is simple. Hence, its canonical representation is of the form $g_+ = \sum_{i=1}^k a_i 1_{A_i}$, with $a_i \geq 0$. Since $\sum_{i=1}^k a_i \mu(A_i) = 0$, it follows that $\mu(A_i) = 0$, for every i such that $a_i > 0$. From finite additivity, we conclude

that $\mu(\cup_i A_i) = 0$, where the union $\cup_i A_i$ is over i such that $a_i > 0$. Therefore, $g_+ = 0$, a.e., and also $g = 0$, a.e.

5 THE INTEGRAL OF A NONNEGATIVE FUNCTION

The integral of a nonnegative function will be defined by approximating from below, using simple functions.

Definition 2. For a nonnegative (extended-valued) measurable function: $\Omega \rightarrow [0, \infty]$, we let $S(g)$ be the set of all nonnegative simple (hence automatically measurable) functions q that satisfy $0 \leq q \leq g$, and define

$$\int g d\mu = \sup_{q \in S(g)} \int q d\mu.$$

We will now verify that with this definition, properties N-2 to N-10 are all satisfied. This is easy for some (e.g., property N-2). Most of our effort will be devoted to establishing properties N-7 (linearity) and N-9 (monotone convergence theorem).

The arguments that follow will make occasional use of the following continuity property for monotonic sequences of measurable sets. If $B_i \uparrow B$, then $\mu(B_i) \uparrow \mu(B)$. This property was established in the notes for Lecture 1, for the special case where μ is a probability measure, but the same proof applies to the general case.

5.1 Verification of some easy properties

Throughout this subsection, we assume that g is measurable and nonnegative.

Property N-2: For every $q \in S(g)$, we have $\int q d\mu \geq 0$ (property S-2). Thus, $\int g d\mu = \sup_{q \in S(g)} \int q d\mu \geq 0$.

Property N-3: If $g = 0$, a.e., and $0 \leq q \leq g$, then $q = 0$, a.e. Therefore, $\int q d\mu = 0$ for every $q \in S(g)$ (by property S-3), which implies that $\int g d\mu = 0$.

Property N-4: Suppose that $0 \leq g \leq h$. Then, $S(g) \subset S(h)$, which implies that

$$\int g d\mu = \sup_{q \in S(g)} \int q d\mu \leq \sup_{q \in S(h)} \int q d\mu = \int h d\mu.$$

Property N-5: Suppose that $g = h$, a.e. Let $A = \{\omega \mid g(\omega) = h(\omega)\}$, and note that the complement of A has zero measure, so that $1_A q = q$, a.e., for any

function q . Then,

$$\begin{aligned}\int g d\mu &= \sup_{q \in S(g)} \int q d\mu = \sup_{q \in S(g)} \int 1_A q d\mu \leq \sup_{q \in S(1_A g)} \int q d\mu \\ &\leq \sup_{q \in S(h)} \int q d\mu = \int h d\mu.\end{aligned}$$

A symmetrical argument yields $\int h d\mu \leq \int g d\mu$.

Exercise: Justify the above sequence of equalities and inequalities.

Property N-4': Suppose that $g \leq h$, a.e. Then, there exists a function such that $g' \leq h$ and $g = g'$, a.e. Property N-5 yields $\int g d\mu = \int g' d\mu$. Property N-4 yields $\int g' d\mu \leq \int h d\mu$. These imply that $\int g d\mu \leq \int h d\mu$.

Property N-6: Suppose that $g \geq 0$ but the relation $g = 0$, a.e., is not true. We will show that $\int g d\mu > 0$. Let $B = \{\omega \mid g(\omega) > 0\}$. Then, $\mu(B) > 0$. Let $B_n = \{\omega \mid g(\omega) > 1/n\}$. Then, $B_n \uparrow B$ and, therefore, $\mu(B_n) \uparrow \mu(B) > 0$. This shows that for some n we have $\mu(B_n) > 0$. Note that $g \geq (1/n)1_{B_n}$. Then, properties S-4, S-8, and 1 yield

$$\int g d\mu \geq \int \frac{1}{n} \cdot 1_{B_n} d\mu = \frac{1}{n} \int 1_{B_n} d\mu = \frac{1}{n} \mu(B_n) > 0.$$

Property N-8, when $a \geq 0$: If $a = 0$, the result is immediate. Assume that $a > 0$. It is not hard to see that $S(g)$ if and only if $q \in S(ag)$. Thus,

$$\int (ag) d\mu = \sup_{q \in S(ag)} \int q d\mu = \sup_{aq \in S(ag)} \int (aq) d\mu = \sup_{q \in S(g)} \int (aq) d\mu = a \int q d\mu.$$

5.2 Proof of the Monotone Convergence Theorem

We first provide the proof **Property N-9**, for the special case where g is equal to a simple function, and then generalize.

Let q be a nonnegative simple function, represented in the form $\sum_{i=1}^k a_i 1_{A_i}$, where the a_i are finite nonnegative numbers, and where the measurable sets form a partition of Ω . Let g_n be a sequence of nonnegative measurable (not necessarily simple) functions such that $g_n \uparrow q$. The limit $\lim_{n \rightarrow \infty} \int g_n d\mu$ exists by monotonicity. We need to show that this limit is $\int q d\mu$. We distinguish between two different cases, depending on whether $\int q d\mu$ is finite or infinite.

- (i) Suppose that $\int q d\mu = \infty$. This implies that there exists some i for which $a_i > 0$ and $\mu(A_i) = \infty$. Fix such an i and let

$$B_n = \{\omega \in A_i \mid g_n(\omega) > a_i/2\}.$$

For every $\omega \in A_i$, there exists some n such that $g_n(\omega) > a_i/2$. Therefore, $B_n \uparrow A_i$. From the continuity of measures, we obtain $\mu(B_n) \uparrow \infty$. Now, note that $g_n \geq (a_i/2)1_{B_n}$. Then, using property N-4, we have

$$\int g_n d\mu \geq \frac{a_i}{2} \mu(B_n) \uparrow \infty = \int q d\mu.$$

- (ii) Suppose now that $\int q d\mu < \infty$. Then, $\mu(A_i) < \infty$, for all $i \in S$ for which $a_i > 0$. Let

$$A = \bigcup_{\{i: a_i > 0\}} A_i.$$

By finite additivity, we have $\mu(A) < \infty$. Let us fix a positive integer r such that $1/r < a_i$ for every i such that $a_i > 0$. Let

$$B_n = \{\omega \in A \mid g_n(\omega) \geq q(\omega) - (1/r)\}.$$

We observe that $B_n \uparrow A$ and, by continuity $\mu(B_n) \uparrow \mu(A)$. Since $\mu(A) = \mu(B_n) + \mu(A \setminus B_n)$, and $\mu(A) < \infty$, this also yields $\mu(A \setminus B_n) \downarrow 0$.

Note that $1_A q = q$, a.e. Using properties, S-5 and S-7, we have

$$\int q d\mu = \int 1_A q d\mu = \int 1_{B_n} q d\mu + \int 1_{A \setminus B_n} q d\mu. \quad (2)$$

For $\omega \in B_n$, we have $g_n(\omega) + (1/r) \geq q(\omega)$. Thus, $g_n + (1/r)1_{B_n} \geq 1_{B_n}q$. Using properties N-4 and S-7, together with Eq. (2), we have

$$\begin{aligned} \int g_n d\mu + \int \frac{1}{r} 1_{B_n} d\mu &\geq \int 1_{B_n} q d\mu = \int q d\mu - \int 1_{A \setminus B_n} q d\mu \\ &\geq \int q d\mu - a\mu(A \setminus B_n), \end{aligned}$$

where $a = \max_i a_i$. By taking the limit as $n \rightarrow \infty$, we obtain

$$\lim_{n \rightarrow \infty} \int g_n d\mu + \frac{1}{r} \mu(A) \geq \int q d\mu.$$

Since this is true for every $r > 1/(\min_i a_i)$, we must have

$$\lim_{n \rightarrow \infty} \int g_n d\mu \geq \int q d\mu.$$

On the other hand, we have $g_n \leq q$, so that $\int g_n d\mu \leq \int q d\mu$, and $\lim_{n \rightarrow \infty} \int g_n d\mu \leq \int q d\mu$.

We now turn to the general case. We assume that $g_n \uparrow g$. Suppose that $q \in S(g)$, so that $0 \leq q \leq g$. We have

$$0 \leq \min\{g_n, q\} \uparrow \min\{g, q\} = q.$$

Therefore,

$$\lim_{n \rightarrow \infty} \int g_n d\mu \geq \lim_{n \rightarrow \infty} \int \min\{g_n, q\} d\mu = \int q d\mu.$$

(The inequality above uses property N-4; the equality relies on the fact that we already proved the MCT for the case where the limit function is simple.) By taking the supremum over $q \in S(g)$, we obtain

$$\lim_{n \rightarrow \infty} \int g_n d\mu \geq \sup_{q \in S(g)} \int q d\mu = \int g d\mu.$$

On the other hand, we have $g_n \leq g$, so that $\int g_n d\mu \leq \int g d\mu$. Therefore, $\lim_{n \rightarrow \infty} \int g_n d\mu \leq \int g d\mu$, which concludes the proof of property N-9.

To prove property '9 suppose that $g_n \uparrow g$, a.e. Then, there exist functions g'_n and g' , such that $g_n = g'_n$, a.e., $g = g'$, a.e., and $g'_n \uparrow g'$. By combining properties N-5 and N-9, we obtain

$$\lim_{n \rightarrow \infty} \int g_n d\mu = \lim_{n \rightarrow \infty} \int g'_n d\mu = \int g' d\mu = \int g d\mu.$$

5.3 Approximating g from below using “special” simple functions

Let g be a nonnegative measurable function. From the definition of $\int g d\mu$, it follows that there exists a sequence $q_n \in S(g)$ such that $\int q_n d\mu \rightarrow \int g d\mu$. This does not provide us with much information on the sequence. In contrast, the construction that follows provides us with a concrete way of approximating $\int g d\mu$.

For any positive integer r , we define a function $g_r : \Omega \rightarrow \mathbb{R}$ by letting

$$g_r(\omega) = \begin{cases} r, & \text{if } g(\omega) \geq r \\ \frac{i}{2^r}, & \text{if } \frac{i}{2^r} \leq g(\omega) < \frac{i+1}{2^r}, \quad i = 0, 1, \dots, r2^r - 1 \end{cases}$$

In words, the function g_r is a quantized version of g . For every ω , the value of $g(\omega)$ is first capped at r , and then rounded down to the nearest multiple of $\frac{1}{2^r}$.

We note a few properties of g_r that are direct consequences of its definition.

- (a) For every r , the function g_r is simple (and, in particular, measurable).

- (b) We have $0 \leq g_r \uparrow g$; that is, for every, we have $g_r(\omega) \uparrow g(\omega)$.
- (c) If g is bounded above by and $r \geq c$, then $|g_r(\omega) - g(\omega)| \leq 1/2^r$, for every ω .

Statement (b) above gives us a transparent characterization of the set of measurable functions. Namely, a nonnegative function is measurable if and only if it is the monotonic and pointwise limit of simple functions. Indeed, we have established in earlier lectures that pointwise limits of measurable functions are measurable. On the other hand, we just showed that every measurable function is a pointwise monotone limit of simple functions. (While the discussion above was about non-negative functions, see how you can extend it to the general case). Furthermore, the MCT indicates that $g_r d\mu \uparrow \int g d\mu$, for this particular choice of simple functions g_r . (In an alternative line of development of the subject, some texts start by defining $g d\mu$ as the limit of $g_r d\mu$.)

5.4 Linearity

We now prove linearity (**property N-7**). Let g_r and h_r be the approximants of g and h , respectively, defined in Section 5.3. Since $\uparrow g$ and $h_r \uparrow h$, we have $(g_r + h_r) \uparrow (g + h)$. Therefore, using the MCT and property S-7 (linearity for simple functions),

$$\begin{aligned} \int (g + h) d\mu &= \lim_{r \rightarrow \infty} \int (g_r + h_r) d\mu \\ &= \lim_{r \rightarrow \infty} \left(\int g_r d\mu + \int h_r d\mu \right) \\ &= \lim_{r \rightarrow \infty} \int g_r d\mu + \lim_{r \rightarrow \infty} \int h_r d\mu \\ &= \int g d\mu + \int h d\mu. \end{aligned}$$

5.5 Using an integral to define a measure

In order to prove the last property (**N-10**) one uses countable additivity for the measure μ , a limiting argument based on the approximations by simple functions, and the MCT. The detailed proof is left as an exercise for the reader.

6 THE GENERAL CASE

Consider now a measurable function: $\Omega \rightarrow \overline{\mathbb{R}}$. Let $A_+ = \{\omega \mid g(\omega) > 0\}$ and $A_- = \{\omega \mid g(\omega) < 0\}$; note that these are measurable sets. Let $g \cdot 1_{A_+}$ and $g_- = -1_{A_-} g$; note that these are nonnegative (possibly extended-valued) measurable functions. We then have $= g_+ - g_-$, and we define

$$\int g d\mu = \int g_+ d\mu - \int g_- d\mu.$$

The integral $\int g d\mu$ is well-defined, as long as we do not have both $d\mu$ and $\int g_- d\mu$ equal to infinity.

With this definition, verifying properties 3-6 and 8 is not too difficult, and there are no surprises. We decompose the function g and h into negative and positive parts, and apply the properties already proved for the nonnegative case. The details are left as an exercise. The only extra work is needed for property 7. It is clear, however, that establishing **property 7** is equivalent to **property N-7** and the following statement:

$$g \geq 0, h \geq 0 \Rightarrow \int (g - h) d\mu = \int gd\mu - \int hd\mu \quad (3)$$

In the special case $g \geq h$ property (3) is just a consequence of **property N-7**:

$$g \geq h \geq 0 \Rightarrow \int gd\mu = \int (g - h + h) d\mu = \int (g - h) d\mu + \int hd\mu. \quad (4)$$

The general case of (3) can be shown via the following argument:

$$\int (g - h) d\mu \triangleq \int (g - h) 1\{g > h\} d\mu - \int (h - g) 1\{g \leq h\} d\mu \quad (5)$$

$$\begin{aligned} &= \int g 1\{g > h\} d\mu - \int h 1\{g > h\} d\mu \\ &\quad + \int g 1\{g \leq h\} d\mu - \int h 1\{g \leq h\} d\mu \end{aligned} \quad (6)$$

$$= \int gd\mu - \int hd\mu, \quad (7)$$

where (6) is from (4) and (7) is from

$$\int g 1\{g > h\} d\mu + \int g 1\{g \leq h\} d\mu = \int gd\mu$$

by **property N-7**.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

ABSTRACT INTEGRATION — II**Contents**

1. Borel-Cantelli revisited
2. Connections between abstract integration and elementary definitions of integrals and expectations
3. Fatou's lemma
4. Dominated convergence theorem

In the previous lecture:

- (a) We defined the notion of an integral of a measurable function with respect to a measure ($\int g d\mu$), which subsumes the special case of expectations ($\mathbb{E}[X] = \int X d\mathbb{P}$), where X is a random variable, and \mathbb{P} is a probability measure.
- (b) We saw that integrals are always well-defined, though possibly infinite, if the function being integrated is nonnegative.
- (c) For a general function g , we decompose it as the sum $g = g_+ - g_-$ of a positive and a negative function, and integrate each piece separately. The integral is well defined unless both $\int g_+ d\mu$ and $\int g_- d\mu$ happen to be infinite.
- (d) We saw that integrals obey a long list natural properties, including linearity:
$$\int(g + h) d\mu = \int g d\mu + \int h d\mu.$$
- (e) We stated the Monotone Convergence Theorem (MCT), according to which, if $\{g_n\}$ is a nondecreasing sequence of nonnegative measurable functions that converge pointwise to a function g , then $\lim_{n \rightarrow \infty} \int g_n d\mu = \int g d\mu$.
- (f) Finally, we saw that for every nonnegative measurable function g , we can find an nondecreasing sequence of nonnegative simple functions that converges (pointwise) to g .

1 BOREL-CANTELLI REVISITED

Recall that one of the Borel-Cantelli lemmas states that if $\sum_{i=1}^{\infty} \mathbb{P}(A_i) < \infty$, then $\mathbb{P}(A_i \text{ i.o.}) = 0$. In this section, we rederive this result using the new machinery that we have available.

Let X_i be the indicator function of the event A_i , so that $\mathbb{E}[X_i] = \mathbb{P}(A_i)$. Thus, by assumption $\sum_{i=1}^{\infty} \mathbb{E}[X_i] < \infty$. The random variables $\sum_{i=1}^n X_i$ are nonnegative and form an increasing sequence, as n increases. Furthermore,

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n X_i = \sum_{i=1}^{\infty} X_i,$$

pointwise; that is, for every ω , we have $\lim_{n \rightarrow \infty} \sum_{i=1}^n X_i(\omega) = \sum_{i=1}^{\infty} X_i(\omega)$.

We can now apply the MCT, and then the linearity property of expectations (for finite sums), to obtain

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{\infty} X_i\right] &= \lim_{n \rightarrow \infty} \mathbb{E}\left[\sum_{i=1}^n X_i\right] \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{E}[X_i] \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{P}(A_i) \\ &= \sum_{i=1}^{\infty} \mathbb{P}(A_i) \\ &< \infty. \end{aligned}$$

This implies that $\sum_{i=1}^{\infty} X_i < \infty$, a.s. (This is intuitively obvious, but a short formal proof is actually needed.) It follows that, with probability 1, only finitely many of the events A_i can occur. Equivalently, the probability that infinitely many of the events A_i occur is zero, i.e., $\mathbb{P}(A_i \text{ i.o.}) = 0$.

2 CONNECTIONS BETWEEN ABSTRACT INTEGRATION AND ELEMENTARY DEFINITIONS OF INTEGRALS AND EXPECTATIONS

Abstract integration would not be useful theory if it were inconsistent with the more elementary notions of integration. For discrete random variables taking values in a finite range, this consistency is automatic because of the definition of an integral of a simple function. We will now verify some additional aspects of this consistency.

2.1 Connection with Riemann integration.

We state here the following reassuring result. Let $\Omega = \mathbb{R}$ and λ be the Lebesgue measure (considered on either Borel σ -algebra, or its completion the Lebesgue σ -algebra). Suppose that f is a Riemann integrable function on some interval $[a, b]$. Then, f is *Lebesgue-measurable* and its Lebesgue integral equals the Riemann integral:

$$\int_a^b f(x) dx = \int_{[a,b]} f d\lambda = \int_{\mathbb{R}} 1_{[a,b]} f d\lambda$$

In particular, every *Borel* function's Lebesgue integral coincides with its Riemann integral whenever the latter exists.

Proof (optional). Consider an arbitrary finite partition σ of $[a, b]$. Corresponding to each σ there is a piece-wise constant (hence simple) function $f_\sigma(x) \leq f(x)$ and $f'_\sigma(x) \geq f(x)$ such that

$$\int_{[a,b]} f_\sigma d\lambda = L(\sigma) \tag{1}$$

$$\int_{[a,b]} f'_\sigma d\lambda = U(\sigma) \tag{2}$$

where $L(\sigma)$ and $U(\sigma)$ are lower and upper Darboux sums (defined in Lecture 7). There exists a sequence of partitions σ_n , each refining the previous one, such that

$$L(\sigma_n) \nearrow \sup_{\sigma} L(\sigma) \tag{3}$$

$$U(\sigma_n) \searrow \inf_{\sigma} U(\sigma). \tag{4}$$

On the other hand the corresponding sequences of functions f_{σ_n} and f'_{σ_n} are monotone, hence converging:

$$f_{\sigma_n}(x) \nearrow \underline{f}(x) \leq f(x) \quad \forall x \in [a, b] \tag{5}$$

$$f'_{\sigma_n}(x) \searrow \overline{f}(x) \geq f(x) \quad \forall x \in [a, b] \tag{6}$$

From (1)-(2) and Riemann integrability we conclude that

$$\int_{[a,b]} \underline{f} d\lambda = \int_{[a,b]} \overline{f} d\lambda = \int_a^b g(x) dx. \tag{7}$$

Consequently, $\int |\overline{f} - \underline{f}| d\lambda = 0$ and thus

$$\underline{f}(x) = \overline{f}(x) = f(x) \quad \text{for a.e. } x$$

implying f is Lebesgue measurable (and coincides with some Borel measurable function except on a set of measure zero). The equality of Lebesgue and Riemann integrals of f in turn follows from (7). \square

2.2 Evaluating expectations by integrating on different spaces

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $X : \Omega \rightarrow \mathbb{R}$, be a random variable. We then obtain a second probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$, where \mathcal{B} is the Borel σ -field, and \mathbb{P}_X is the probability law of X , defined by

$$\mathbb{P}_X(A) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\}), \quad A \in \mathcal{B}.$$

Consider now a measurable function $g : \mathbb{R} \rightarrow \mathbb{R}$, and use it to define a new random variable $Y = g(X)$, and a corresponding probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P}_Y)$. The expectation of Y can be evaluated in three different ways, that is, by integrating over either of the three spaces we have introduced.

Theorem 1. We have

$$\int Y d\mathbb{P} = \int g d\mathbb{P}_X = \int y d\mathbb{P}_Y,$$

and all three \int 's exist or do not exist simultaneously.

Proof: We follow the “standard program”: first establish the result for simple functions, then take the limit to deal with nonnegative functions, and finally generalize.

Let g be a simple function, which takes values in a finite set y_1, \dots, y_k . Using the definition of the integral of a simple function we have

$$\begin{aligned} \int Y d\mathbb{P} &= \sum_{y_i} y_i \mathbb{P}_Y(Y = y_i) \\ &= \sum_{y_i} y_i \mathbb{P}(\{\omega \mid Y(\omega) = y_i\}) \\ &= \sum_{y_i} y_i \mathbb{P}(\{\omega \mid g(X(\omega)) = y_i\}). \end{aligned}$$

Similarly,

$$\int g d\mathbb{P}_X = \sum_{y_i} y_i \mathbb{P}_X(\{x \mid g(x) = y_i\}).$$

However, from the definition of \mathbb{P}_X , we obtain

$$\begin{aligned}\mathbb{P}_X(\{x \mid g(x) = y_i\}) &= \mathbb{P}_X(g^{-1}(y_i)) \\ &= \mathbb{P}(\{\omega \mid X(\omega) \in g^{-1}(y_i)\}) \\ &= \mathbb{P}(\{\omega \mid g(X(\omega)) = y_i\}),\end{aligned}$$

and the equalities in the theorem follow, for simple functions.

Let now g be nonnegative function, and let $\{g_n\}$ be an increasing sequence of nonnegative simple functions that converges to g . Note that $g_n(X)$ converges monotonically to $g(X)$. We then have

$$\int Y d\mathbb{P} = \int g(X) d\mathbb{P} = \lim_{n \rightarrow \infty} \int g_n(X) d\mathbb{P} = \lim_{n \rightarrow \infty} \int g_n d\mathbb{P}_X = \int g d\mathbb{P}_X.$$

(The second equality is the MCT; the third is the result that we already proved for simple functions; the last equality is once more the MCT.)

The case of general (not just nonnegative) functions follows easily from the above – the details are omitted. This proves the theorem. \square

2.3 The case of continuous random variables, described by PDFs

We can now revisit the development of continuous random variables (Lecture 4), in a more rigorous manner. We say that a random variable $X : \Omega \rightarrow \mathbb{R}$ is continuous if its CDF can be written in the form

$$F_X(x) = \mathbb{P}(X \leq x) = \int 1_{(-\infty, x]} f d\lambda, \quad \forall x \in \mathbb{R},$$

where λ is Lebesgue measure, and f is a nonnegative measurable function with $\int_{\mathbb{R}} f d\lambda = 1$. Recall that by Theorem 4 of Lecture 4 to each CDF there corresponds a unique probability measure \mathbb{P}_X on $(\mathbb{R}, \mathcal{B})$. In this case \mathbb{P}_X has a particularly simple expression:

$$\mathbb{P}_X(A) = \int_A f d\lambda \tag{8}$$

for any Borel set A . (Obviously, the CDF of \mathbb{P}_X is F_X . The fact that (8) defines a valid measure is **property 10** from Lecture 7.)

When f is Riemann integrable and the set $A = [a, b]$ is an interval, we can also write $\mathbb{P}_X(A) = \int_a^b f(x) dx$, where the latter integral is an ordinary Riemann integral.

Theorem 2. For any measurable function g we have

$$\mathbb{E}[g(X)] = \int g d\mathbb{P}_X = \int (gf) d\lambda$$

where all \mathbb{E} and \int 's exist or do not exist simultaneously.

Note: Since integrals of non-negative functions always exist, this also gives a convenient criterion: $\mathbb{E}[g(X)]$ is finite iff $\int |g| f d\lambda < \infty$.

Proof: The first equality was shown in Theorem 1. So, let us concentrate on the second. Following the usual program, let us first consider the case where g is a simple function, of the form $g = \sum_{i=1}^k a_i 1_{A_i}$, for some measurable disjoint subsets A_i of the real line. We have

$$\begin{aligned} \int g d\mathbb{P}_X &= \sum_{i=1}^k a_i \mathbb{P}_X(A_i) \\ &= \sum_{i=1}^k a_i \int_{A_i} f d\lambda \\ &= \sum_{i=1}^k \int a_i 1_{A_i} f d\lambda \\ &= \int \sum_{i=1}^k a_i 1_{A_i} f d\lambda \\ &= \int (gf) d\lambda. \end{aligned}$$

The first equality is the definition of the integral for simple functions. The second uses Eq. (8). The fourth uses linearity of integrals. The fifth uses the definition of g .

Suppose now that g is a nonnegative function, and let $\{g_n\}$ be an increasing sequence of nonnegative functions that converges to g , pointwise. Since f is nonnegative, note that $g_n f$ also increases monotonically and converges to gf . Then,

$$\int g d\mathbb{P}_X = \lim_{n \rightarrow \infty} \int g_n d\mathbb{P}_X = \lim_{n \rightarrow \infty} \int (g_n f) d\lambda = \int (gf) d\lambda.$$

The first and the third equality above is the MCT. The middle equality is the result we already proved, for the case of a simple function g_n .

Finally, if g is not nonnegative, the result is proved by considering separately the positive and negative parts of g . \square

When g and f are “nice” functions, e.g., piecewise continuous, Theorem 2 yields the familiar formula

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx,$$

where the integral is now an ordinary (improper) Riemann integral.

3 FATOU'S LEMMA

Note that for any two random variables, we have $\min\{X, Y\} \leq X$ and $\min\{X, Y\} \leq Y$. Taking expectations, we obtain $\mathbb{E}[\min\{X, Y\}] \leq \min\{\mathbb{E}[X], \mathbb{E}[Y]\}$. Fatou's lemma is in the same spirit, except that infinitely many random variables are involved, as well as a limiting operation, so some additional technical conditions are needed.

Theorem 3. *Let $f_n \geq 0$ be measurable, then*

$$\int \liminf_{n \rightarrow \infty} f_n d\mu \leq \liminf_{n \rightarrow \infty} \int f_n d\mu$$

Proof: Fix some n . We have

$$\inf_{k \geq n} f_k \leq f_m, \quad \forall m \geq n.$$

Integrating both sides, we obtain

$$\int \inf_{k \geq n} f_k d\mu \leq \int f_m d\mu, \quad \forall m \geq n.$$

Taking the infimum of both sides with respect to m , we obtain

$$\int \inf_{k \geq n} f_k d\mu \leq \inf_{m \geq n} \int f_m d\mu \tag{9}$$

The statement of the Theorem follows from (9) after taking the limit $\lim_{n \rightarrow \infty}$. Indeed, the sequence $\inf_{k \geq n} f_k$ is nonnegative and nondecreasing with n , and converges to $\liminf_{n \rightarrow \infty} f_n$. Therefore, from MCT we obtain

$$\lim_{n \rightarrow \infty} \int \inf_{k \geq n} f_k d\mu = \int \lim_{n \rightarrow \infty} \inf_{k \geq n} f_k \triangleq \int \liminf_{n \rightarrow \infty} f_n d\mu$$

Similarly, the limit as $n \rightarrow \infty$ of the right-hand side of (9) converges to $\liminf \int f_n d\mu$. \square

Corollary 1. Let Y be a random variable that satisfies $\mathbb{E}[|Y|] < \infty$.

- (a) If $Y \leq X_n$, for all n , then $\mathbb{E}[\liminf_{n \rightarrow \infty} X_n] \leq \liminf_{n \rightarrow \infty} \mathbb{E}[X_n]$.
- (b) If $X_n \leq Y$, for all n , then $\mathbb{E}[\limsup_{n \rightarrow \infty} X_n] \geq \limsup_{n \rightarrow \infty} \mathbb{E}[X_n]$.

Proof: Apply Theorem 3 to $X_n - Y$ or $Y - X_n$. \square

4 DOMINATED CONVERGENCE THEOREM

The dominated convergence theorem complements the MCT by providing an alternative set of conditions under which a limit and an expectation can be interchanged.

Theorem 4. (DCT) Consider a sequence of random variables $\{X_n\}$ that converges to X a.e. Suppose that $|X_n| \leq Y$, for all n , where Y is a non-negative random variable that satisfies $\mathbb{E}[Y] < \infty$. Then, $\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = \mathbb{E}[X]$.

Proof: Let $A \subset \Omega$ be the set of outcomes ω along which $X_n(\omega) \rightarrow X(\omega)$ as $n \rightarrow \infty$. Then $\mathbb{P}(A^c) = 0$. Let $\tilde{X}_n(\omega) = X_n(\omega)$ for $\omega \in A$ and $= 0$ otherwise. Similarly, let $\tilde{X}(\omega) = X(\omega)$, $\omega \in A$ and $= 0$ otherwise. Then $\mathbb{E}[\tilde{X}_n] = \mathbb{E}[X_n]$, $\mathbb{E}[\tilde{X}] = \mathbb{E}[X]$ and $\tilde{X}_n \rightarrow \tilde{X}$ for all ω . Thus we may assume, without the loss of generality that $X_n(\omega) \rightarrow X(\omega)$ for all ω .

Since $-Y \leq X_n \leq Y$, we can apply both parts of Fatou's lemma , to obtain

$$\mathbb{E}[X] = \mathbb{E}[\liminf_{n \rightarrow \infty} X_n] \leq \liminf_{n \rightarrow \infty} \mathbb{E}[X_n] \leq \limsup_{n \rightarrow \infty} \mathbb{E}[X_n] \leq \mathbb{E}[\limsup_{n \rightarrow \infty} X_n] = \mathbb{E}[X].$$

This proves that

$$\mathbb{E}[X] = \liminf_{n \rightarrow \infty} \mathbb{E}[X_n] = \limsup_{n \rightarrow \infty} \mathbb{E}[X_n].$$

In particular, the limit $\lim_{n \rightarrow \infty} \mathbb{E}[X_n]$ exists and equals $\mathbb{E}[X]$. \square

Remark: We note that the DCT remains valid for general measures, not just for probability measures (the proof is the same). However, the following statement (Bounded Convergence Theorem), is specific to probability measures: *If there exists a constant $c \in \mathbb{R}$ such that $|X_n| \leq c$, a.s., for all n , then $\lim_n \mathbb{E}[X_n] = \mathbb{E}[\lim_n X_n]$.*

Corollary 2. Suppose that $\sum_{n=1}^{\infty} \mathbb{E}[|Z_n|] < \infty$. Then,

$$\sum_{n=1}^{\infty} \mathbb{E}[Z_n] = \mathbb{E}\left[\sum_{n=1}^{\infty} Z_n\right].$$

Proof: By the monotone convergence theorem, applied to $Y_n = \sum_{k=1}^n |Z_k|$, we have

$$\mathbb{E}\left[\sum_{n=1}^{\infty} |Z_n|\right] = \sum_{n=1}^{\infty} \mathbb{E}[|Z_n|] < \infty.$$

Let $X_n = \sum_{i=1}^n Z_i$ and note that $\lim_{n \rightarrow \infty} X_n = \sum_{i=1}^{\infty} Z_i$. We observe that $|X_n| \leq \sum_{i=1}^{\infty} |Z_i|$, which has finite expectation, as shown earlier. The result follows from the dominated convergence theorem. \square

Exercise: Can you prove Corollary 1 directly from the monotone convergence theorem, without appealing to the DCT or Fatou's lemma?

Theorems such as MCT and DCT impose assumptions additional to the assumption that $X_n \rightarrow X$ a.e. that insure that $\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$. It should not be surprising that, in general just having $X_n \rightarrow X$ a.e. is not enough. Here is a counter-example. Let $\Omega = [0, 1]$, let \mathcal{F} be the Borel sigma-field \mathcal{B} on $[0, 1]$, and let \mathbb{P} be the uniform (Lebesgue) probability measure. Let $X(\omega) = 0$ for all $\omega \in [0, 1]$. Let

$$X_n(\omega) = \begin{cases} n, & \text{when } \omega \in (0, \frac{1}{n}); \\ 0, & \text{when } \omega = 0 \text{ or } \omega \in (\frac{1}{n}, 1]. \end{cases}$$

Verify that $X_n(\omega) \rightarrow 0$ for all ω , but $\mathbb{E}[X_n] = n(1/n) = 1$ and thus $\mathbb{E}[X_n] \rightarrow 0$ does not hold.

The example above shows that DCT does not hold unless we make an additional assumption, such as $|X_n| \leq Y$ for some random variable Y with $\mathbb{E}[Y] < \infty$. However, the sequence X_n is not increasing.

Exercise:

- (a) Establish the following generalization of the MCT. Suppose X_n is a.e. increasing sequence of random variables, but suppose X_n are not necessarily non-negative. Let $\lim_n X_n = X$ a.e. Suppose $X_n \geq Y$ a.e. for some random variable Y . Finally, suppose the expectations of X_n, X and Y are all finite. Establish that $\lim_n \mathbb{E}[X_n] = \mathbb{E}[X]$.
- (b) Construct a sequence of random variables X_n which is increasing a.e., but $\mathbb{E}[X_n]$ does not converge to $\mathbb{E}[X]$, where $X = \lim_n X_n$ a.e.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Lecture 9

Fall 2018

PRODUCT MEASURE AND FUBINI'S THEOREM

Contents

1. Product measure
2. Fubini's theorem

In elementary math and calculus, we often interchange the order of summation and integration. The discussion here is concerned with conditions under which this is legitimate.

1 PRODUCT MEASURE

Consider two probabilistic experiments with probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$, respectively. We are interested in forming a probabilistic model of a “joint experiment” in which the original two experiments are carried out independently.

1.1 The sample space of the joint experiment

If the first experiment has an outcome ω_1 , and the second has an outcome ω_2 , then the outcome of the joint experiment is the pair (ω_1, ω_2) . This leads us to define a new sample space $\Omega = \Omega_1 \times \Omega_2$.

1.2 The σ -algebra of the joint experiment

Next, we need a σ -algebra on Ω . If $A_1 \in \mathcal{F}_1$, we certainly want to be able to talk about the event $\{\omega_1 \in A_1\}$ and its probability. In terms of the joint experiment, this would be the same as the event

$$A_1 \times \Omega_1 = \{(\omega_1, \omega_2) \mid \omega_1 \in A_1, \omega_2 \in \Omega_2\}.$$

Thus, we would like our σ -algebra on Ω to include all sets of the form $A_1 \times \Omega_2$, (with $A_1 \in \mathcal{F}_1$) and by symmetry, all sets of the form $\Omega_1 \times A_2$ (with $(A_2 \in \mathcal{F}_2)$. This leads us to the following definition.

Definition 1. We define $\mathcal{F}_1 \times \mathcal{F}_2$ as the smallest σ -algebra of subsets of $\Omega_1 \times \Omega_2$ that contains all sets of the form $A_1 \times \Omega_2$ and $\Omega_1 \times A_2$, where $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$.

Note that the notation $\mathcal{F}_1 \times \mathcal{F}_2$ is misleading: this is not the Cartesian product of \mathcal{F}_1 and \mathcal{F}_2 !

Since σ -fields are closed under intersection, we observe that if $A_i \in \mathcal{F}_i$, then $A_1 \times A_2 = (A_1 \times \Omega_2) \cap (\Omega_1 \cap A_2) \in \mathcal{F}_1 \times \mathcal{F}_2$. It turns out (and is not hard to show) that $\mathcal{F}_1 \times \mathcal{F}_2$ can also be defined as the smallest σ -algebra containing all sets of the form $A_1 \times A_2$, where $A_i \in \mathcal{F}_i$. Alternatively, suppose \mathcal{F}_1 and \mathcal{F}_2 are generated by algebras $\mathcal{F}_{0,1}, \mathcal{F}_{0,2}$. That is $\mathcal{F}_i = \sigma(\mathcal{F}_{0,i}), i = 1, 2$. Then $\mathcal{F}_1 \times \mathcal{F}_2$ is also the smallest σ -algebra containing all sets of the form $A_1 \times A_2$, where $A_i \in \mathcal{F}_{0,i}$.

In the sequel, we will talk about $g : \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$ – measurable functions with respect to $\mathcal{F}_1 \times \mathcal{F}_2$. Recall, this means that for any Borel set $B \subset \mathbb{R}$, the set $\{(\omega_1, \omega_2) \mid g(\omega_1, \omega_2) \in B\}$ belongs to the σ -algebra $\mathcal{F}_1 \times \mathcal{F}_2$. As a practical matter, it is enough to verify that for any scalar c , the set $\{(\omega_1, \omega_2) \mid g(\omega_1, \omega_2) \leq c\}$ is measurable. Other than using this definition directly, how else can we verify that such a function g is measurable? The basic tools at hand are the following:

- (a) continuous functions from \mathbb{R}^2 to \mathbb{R} are measurable;
- (b) indicator functions of measurable sets are measurable;
- (c) combining measurable functions in the usual ways (e.g., adding them, multiplying them, taking limits, etc.) results in measurable functions.

The following proposition gives further information about $\mathcal{F}_1 \times \mathcal{F}_2$ and functions measurable with respect to it.

Proposition 1. Let $E \in \mathcal{F}_1 \times \mathcal{F}_2$ then for every $\omega_1 \in \Omega_1$ the set

$$E_{\omega_1} \triangleq \{\omega_2 \mid (\omega_1, \omega_2) \in E\}$$

belongs to \mathcal{F}_2 . Consequently, for every $\mathcal{F}_1 \times \mathcal{F}_2$ -measurable function f and every ω_1 the function

$$f_{\omega_1}(\omega_2) \triangleq f(\omega_1, \omega_2)$$

is \mathcal{F}_2 -measurable.

Remark: E_{ω_1} and f_{ω_1} are called slices of E and f at ω_1 , respectively.

Proof. Fix some ω_1 and define a collection of sets

$$\mathcal{L} = \{E \in \mathcal{F}_1 \times \mathcal{F}_2 \mid E_{\omega_1} \in \mathcal{F}_2\}.$$

When $E = A_1 \times A_2$ the set E_{ω_1} is either empty or equal to A_2 . Thus \mathcal{L} contains all the rectangles. On the other hand, for any sequence E_j we have

$$(\cup_j E_j)_{\omega_1} = \bigcup_j E_j)_{\omega_1}$$

and

$$(E^c)_{\omega_1} = (E_{\omega_1})^c.$$

Thus \mathcal{L} is closed under countable unions and complements. Hence \mathcal{L} is a σ -algebra, which by minimality of $\mathcal{F}_1 \times \mathcal{F}_2$ must be equal to the latter. This shows these statement for sets.

Next, a slice of a simple function

$$f = \sum_{i=1}^N a_i 1_{E_i}$$

at ω_1 is itself a simple (hence measurable) function on $(\Omega_2, \mathcal{F}_2)$. This follows from what was just shown for slices of sets. For the general f we have $f = \lim_{r \rightarrow \infty} f_r$, where f_r are simple functions. Since the slice of each f_r is \mathcal{F}_2 measurable and the class of \mathcal{F}_2 -measurable functions is closed under taking limits the result follows. \square

1.3 The product measure

We now define a measure, to be denoted by $\mathbb{P}_1 \times \mathbb{P}_2$ (or just \mathbb{P} , for short) on the measurable space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$. To capture the notion of independence, we require that

$$\mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2), \quad \forall A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2. \quad (1)$$

Theorem 1. There exists a unique measure \mathbb{P} on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$ that has property (1). Furthermore, for every $E \in \mathcal{F}_1 \times \mathcal{F}_2$ measure $\mathbb{P}(E)$ satisfies

$$\mathbb{P}(E) = \int \mathbb{P}_2(E_{\omega_1}) \mathbb{P}_1(d\omega_1) \quad (2)$$

$$= \int \mathbb{P}_1(E_{\omega_2}) \mathbb{P}_2(d\omega_2). \quad (3)$$

Proof. Uniqueness follows from the fact that $A_1 \times A_2$ is a generating p -system for $\mathcal{F}_1 \times \mathcal{F}_2$ (see Proposition 1 in Lecture 2). We only need to show existence. We start by showing that for every $E \in \mathcal{F}_1 \times \mathcal{F}_2$ the function

$$f_E(\omega_1) \triangleq \mathbb{P}_2(E_{\omega_1})$$

is \mathcal{F}_1 -measurable. Note that $\mathbb{P}_2(E_{\omega_1})$ is well-defined by Proposition 1. Define a collection

$$\mathcal{L} = \{E : f_E \text{ is } \mathcal{F}_1\text{-measurable}\}.$$

When $E = A_1 \times A_2$ the function $f_E(\omega_1) = \mathbb{P}_2(A_2)1_{A_1}(\omega_1)$, which is clearly measurable. Thus \mathcal{L} contains all rectangles. Next, if E and F are disjoint then so are E_{ω_1} and F_{ω_1} . Consequently,

$$f_{E \cup F}(\omega_1) = f_E(\omega_1) + f_F(\omega_2) \quad \text{if } E \cap F = \emptyset. \quad (4)$$

This implies that \mathcal{L} contains all finite unions of disjoint rectangles. The latter is an algebra of sets (since $(A_1 \times A_2)^c$ can be written as disjoint union of 3 rectangles). Finally, if $E_j \nearrow E$ and $E_j \in \mathcal{L}$ then

$$f_{E_j} \nearrow f_E \quad (5)$$

and therefore f_E is \mathcal{F}_1 -measurable. Same argument applies to $E_j \searrow E$. All in all \mathcal{L} is a monotone class, containing an algebra that generates $\mathcal{F}_1 \times \mathcal{F}_2$. So $\mathcal{L} = \mathcal{F}_1 \times \mathcal{F}_2$.

We now *define* for any $E \in \mathcal{F}_1 \times \mathcal{F}_2$

$$\mathbb{P}(E) \triangleq \int f_E(\omega_1) \mathbb{P}_1(d\omega_1). \quad (6)$$

It is evident that this assignment satisfies (1). Finite additivity of \mathbb{P} follows from (4). It remains to show σ -additivity, which in turn is equivalent to continuity. The latter follows from (5) and the MCT.

Thus, existence of \mathbb{P} is established. Furthermore, definition (6) is just a restatement of (2). Regarding (3), construct another measure \mathbb{P}' by exchanging roles of $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ in (6). So constructed \mathbb{P}' automatically satisfies (3). Moreover, \mathbb{P}' also verifies (1) and hence coincides with \mathbb{P} on a p -system of rectangles $A \times B$. By Proposition 1 of Lecture 2 we have: $\mathbb{P}' = \mathbb{P}$. \square

The above discussion extends to the case of any finite number of probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, 2, \dots, k$. In particular there exists a unique measure \mathbb{P} on $\Omega = \Omega_1 \times \dots \times \Omega_k$ such that for every collection of sets $A_i \in \mathcal{F}_i$,

$$\mathbb{P}(A_1 \times \dots \times A_k) = \mathbb{P}(A_1) \times \dots \times \mathbb{P}(A_k).$$

The corresponding σ -algebra on Ω is the smallest σ -algebra containing all sets of the form $A_1 \times \dots \times A_k$ where $A_i \in \mathcal{F}_i$. Moreover, this extends to a countable collections of probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, 2, \dots$, but now the measure is only defined when a finite collection of the $\{A_i\}$ are not Ω_k , i.e. $i = 1, 2, \dots, k$

$$\mathbb{P}(A_1 \times \dots \times A_k \times \dots \times A_{k+1} \times \dots \times A_{k+2} \times \dots) = \mathbb{P}(A_1) \times \dots \times \mathbb{P}(A_k).$$

1.4 Beyond probability measures

Everything in these notes extends to the case where instead of probability measures \mathbb{P}_i , we are dealing with general measures μ_i , under the assumptions that the measures μ_i are **σ -finite**. (A measure μ is called σ -finite if the set Ω can be partitioned into a countable union of sets, each of which has finite measure.)

The most relevant example of a σ -finite measure is the Lebesgue measure on the real line. Indeed, the real line can be broken into a countable sequence of intervals $(n, n + 1]$ each of which has finite Lebesgue measure.

1.5 The product measure on \mathbb{R}^2

The two-dimensional plane \mathbb{R}^2 is the Cartesian product of \mathbb{R} with itself. We endow each copy of \mathbb{R} with the Borel σ -field \mathcal{B} and one-dimensional Lebesgue measure. The resulting σ -field $\mathcal{B} \times \mathcal{B}$ is called the Borel σ -field on \mathbb{R}^2 . The resulting product measure on \mathbb{R}^2 is called two-dimensional Lebesgue measure, to be denoted here by λ_2 . The measure λ_2 corresponds to the natural notion of area. For example,

$$\lambda_2([a, b] \times [c, d]) = \lambda([a, b]) \cdot \lambda([c, d]) = (b - a) \cdot (d - c).$$

More generally, for any “nice” set of the form encountered in calculus, e.g., sets of the form $A = \{(x, y) \mid f(x, y) \leq c\}$, where f is a continuous function, $\lambda_2(A)$ coincides with the usual notion of the area of A .

Remark for those of you who know a little bit of topology – otherwise ignore it. We could define the Borel σ -field on \mathbb{R}^2 as the σ -field generated by the collection of open subsets of \mathbb{R}^2 . (This is the standard way of defining Borel sets in topological spaces.) It turns out that this definition results in the same σ -field as the method of Section 1.2.

2 FUBINI'S THEOREM

Fubini's theorem is a powerful tool that provides conditions for interchanging the order of integration in a double integral. Given that sums are essentially special cases of integrals (with respect to discrete measures), it also gives conditions for interchanging the order of summations, or the order of a summation and an integration. In this respect, it subsumes results such as Corollary 1 at the end of the notes for Lecture 12.

Fubini's theorem holds under two different sets of conditions: (a) nonnegative functions g (compare with the MCT); (b) functions g whose absolute value has a finite integral (compare with the DCT). We state the two versions separately, because of some subtle differences.

The two statements below are taken verbatim from the text by Adams & Guillemin, with minor changes to conform to our notation.

Theorem 2. Let $g : \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$ be a nonnegative measurable function.

Let $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2$ be a product measure. Then,

(a) $\int_{\Omega_2} g(\omega_1, \omega_2) d\mathbb{P}_2$ is a measurable function of ω_1 .

(b) $\int_{\Omega_1} g(\omega_1, \omega_2) d\mathbb{P}_1$ is a measurable function of ω_2 .

(c) We have

$$\begin{aligned} \int_{\Omega_1} \left[\int_{\Omega_2} g(\omega_1, \omega_2) d\mathbb{P}_2 \right] d\mathbb{P}_1 &= \int_{\Omega_2} \left[\int_{\Omega_1} g(\omega_1, \omega_2) d\mathbb{P}_1 \right] d\mathbb{P}_2 \\ &= \int_{\Omega_1 \times \Omega_2} g(\omega_1, \omega_2) d\mathbb{P}. \end{aligned}$$

Note that some of the integrals above may be infinite, but this is not a problem; since everything is nonnegative, expressions of the form $\infty - \infty$ do not arise.

Proof. For simple functions $g = \sum_{i=1}^n a_i 1_{E_i}, E_i \in \mathcal{F}_1 \times \mathcal{F}_2$ statement

(a) follows from measurability of $\omega_1 \mapsto \mathbb{P}_2(E_{\omega_1})$ established in the proof of Theorem 1. For a general g consider a sequence of simple functions

$$g_r(\omega_1, \omega_2) \nearrow g(\omega_1, \omega_2) \quad \forall \omega_1, \omega_2$$

as $r \rightarrow \infty$. Then we have shown that

$$f_r(\omega_1) = \int_{\mathcal{E}_2} g_r(\omega_1, \omega_2) d\mathbb{P}_2$$

are \mathcal{F}_1 measurable and monotonically increasing $f_r \nearrow f$. By the MCT

$$f(\omega_1) \triangleq \lim_{r \rightarrow \infty} \int_{\mathcal{E}_2} g_r(\omega_1, \omega_2) d\mathbb{P}_2 \quad (7)$$

$$= \int_{\mathcal{E}_2} \lim_{r \rightarrow \infty} g_r(\omega_1, \omega_2) d\mathbb{P}_2 \quad (8)$$

$$= \int_{\mathcal{E}_2} g(\omega_1, \omega_2) d\mathbb{P}_2. \quad (9)$$

Since f is a limit of measurable f_r 's – f must be measurable. By (9) the integral over \mathcal{E}_2 is also \mathcal{F}_1 measurable. This establishes (a) and (b) by symmetry. Finally (c), for a simple function g is just (2)-(3), while for a general function g we just need to integrate (7) interchanging \int and \lim by the MCT at will. \square

Recall now that a function is said to be **integrable** if it is measurable and the integral of its absolute value is finite.

Theorem 3. Let $g : \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$ be a measurable function such that

$$\int_{\Omega_1 \times \Omega_2} |g(\omega_1, \omega_2)| d\mathbb{P} < \infty, \quad (10)$$

where $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2$.

- (a) For almost all $\omega_1 \in \Omega_1$, $g(\omega_1, \omega_2)$ is an integrable function of ω_2 .
- (b) For almost all $\omega_2 \in \Omega_2$, $g(\omega_1, \omega_2)$ is an integrable function of ω_1 .
- (c) There exists an integrable function $h : \Omega_1 \rightarrow \mathbb{R}$ such that $\int_{\Omega_2} g(\omega_1, \omega_2) d\mathbb{P}_2 = h(\omega_1)$, a.s. (i.e., except for a set of ω_1 of zero \mathbb{P}_1 -measure for which $\int_{\Omega_2} g(\omega_1, \omega_2) d\mathbb{P}_2$ is undefined or infinite).
- (d) There exists an integrable function $h : \Omega_2 \rightarrow \mathbb{R}$ such that $\int_{\Omega_1} g(\omega_1, \omega_2) d\mathbb{P}_1 = h(\omega_2)$, a.s. (i.e., except for a set of ω_2 of zero \mathbb{P}_2 -measure for which $\int_{\Omega_1} g(\omega_1, \omega_2) d\mathbb{P}_1$ is undefined or infinite).
- (e) We have

$$\begin{aligned} \int_{\Omega_1} \left[\int_{\Omega_2} g(\omega_1, \omega_2) d\mathbb{P}_2 \right] d\mathbb{P}_1 &= \int_{\Omega_2} \left[\int_{\Omega_1} g(\omega_1, \omega_2) d\mathbb{P}_1 \right] d\mathbb{P}_2 \\ &= \int_{\Omega_1 \times \Omega_2} g(\omega_1, \omega_2) d\mathbb{P}. \end{aligned}$$

Remarks:

1. Both Theorems remain valid when dealing with σ -finite measures, such as the Lebesgue measure on \mathbb{R}^2 . This provides us with conditions for the familiar calculus formula

$$\int \int g(x, y) dx dy = \int \int g(x, y) dy dx.$$

2. In order to apply Theorem 3, we need a practical method for checking the integrability condition (10). Here, Theorem 2 comes to the rescue. Indeed, by Theorem 2, we have

$$\int_{\Omega_1 \times \Omega_2} |g(\omega_1, \omega_2)| d\mathbb{P} = \int_{\Omega_1} \int_{\Omega_2} |g(\omega_1, \omega_2)| d\mathbb{P}_2 d\mathbb{P}_1,$$

so all we need is to work with the right hand side, and integrate one variable at a time, possibly also using some bounds on the way.

Proof. By now converting from a non-negative case to integrable case should be familiar. Theorem 3 is no exception: Given a function g , decompose it into its positive and negative parts, apply Theorem 2 to each part, and in the process make sure that you do not encounter expressions of the form $\infty - \infty$. We omit the details. \square

3 SOME CAUTIONARY EXAMPLES

We give a few examples where Fubini's theorem does not apply.

3.1 Nonnegativity and integrability

Suppose that both of our sample spaces are the nonnegative integers: $\Omega_1 = \Omega_2 = \{1, 2, \dots\}$. The σ -fields \mathcal{F}_1 and \mathcal{F}_2 consist of all subsets of Ω_1 and Ω_2 , respectively. Then, $\sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ is composed of all subsets of $\{1, 2, \dots\}^2$. Let both \mathbb{P}_1 and \mathbb{P}_2 be the counting measure, i.e. $\mathbb{P}(A) = |A|$. This means that

$$\int_A g d\mathbb{P}_1 = \sum_{a \in A} f(a), \quad \int_B h d\mathbb{P}_2 = \sum_{b \in B} h(b),$$

and

$$\int_C f d(\mathbb{P}_1 \times \mathbb{P}_2) = \sum_{c \in C} f(c).$$

Consider the function f defined by $f(m, m) = 1$, $f(m, m + 1) = -1$, and $f = 0$ elsewhere. It is easier to visualize f with a picture:

$$\begin{array}{ccccc} 1 & -1 & 0 & 0 & \dots \\ 0 & 1 & -1 & 0 & \dots \\ 0 & 0 & 1 & -1 & \dots \\ 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

So,

$$\begin{aligned} \int_{\Omega_1} \int_{\Omega_2} f d\mathbb{P}_2 d\mathbb{P}_1 &= \sum_n \sum_m f(n, m) = 0 \\ &\neq 1 = \sum_m \sum_n f(n, m) = \int_{\Omega_2} \int_{\Omega_1} f d\mathbb{P}_1 d\mathbb{P}_2. \end{aligned}$$

In this example, the conditions of Fubini's theorem fail to hold: the function f is neither nonnegative nor integrable.

3.2 σ -finiteness

Let $\Omega_1 = (0, 1)$, let \mathcal{F}_1 be the Borel sets, and let \mathbb{P}_1 be the Lebesgue measure. Let $\Omega_2 = (0, 1)$ let \mathcal{F}_2 be the set of all subsets of $(0, 1)$, and let \mathbb{P}_2 be the counting measure. In particular, for every infinite (countable or uncountable) subset of $(0, 1)$, $\mathbb{P}_2(A) = \infty$.

Let $f(x, y) = 1$ if $x = y$, and $f(x, y) = 0$ otherwise. Then,

$$\int_{\Omega_1} \int_{\Omega_2} f(x, y) d\mathbb{P}_2(y) d\mathbb{P}_1(x) = \int_{\Omega_1} \mathbf{1} d\mathbb{P}_1(y) = 1,$$

but

$$\int_{\Omega_2} \int_{\Omega_1} f(x, y) d\mathbb{P}_1(x) d\mathbb{P}_2(y) = \int_{\Omega_2} 0 d\mathbb{P}_2(y) = 0.$$

In this example, the conditions of Fubini's theorem fail to hold: the measure on $(0, 1)$ is not σ -finite.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Lecture 10

Fall 2018

CONTINUOUS RANDOM VARIABLES

Contents

1. Continuous random variables
2. Examples
3. Expected values
4. Joint distributions
5. Independence
6. Radon-Nikodym derivative

Readings: For a less technical version of this material, but with more discussion and examples, see Sections 3.1-3.5 of [BT] and Sections 4.1-4.5 of [GS].

1 CONTINUOUS RANDOM VARIABLES

Recall¹ that a random variable $X : \Omega \rightarrow \mathbb{R}$ is said to be continuous if its CDF can be written in the form

$$\mathbb{P}(X \leq x) = F_X(x) = \int_{(-\infty, x)} f_X(t) dt,$$

for some nonnegative measurable function $f : \mathbb{R} \rightarrow [0, \infty)$, which is called the Probability Density Function (PDF) of X . We then have, for any Borel set

$$\mathbb{P}(X \in B) = \int_B f_X(x) dx = \int_{\mathbb{R}} I_B(x) f_X(x) dx \quad (1)$$

Technical remark: All integrals from now on are understood as Lebesgue integrals, unless stated otherwise. In particular, $\int_a^b f(t) dt$ is a shorthand notation for $\int_{\mathbb{R}} 1_{(a,b)}(t) f(t) d\lambda(t)$, where λ is Lebesgue measure on $(\mathbb{R}, \mathcal{B})$.

¹The reader should revisit Section 4 of the notes for Lecture 5.

We note that f_X should be more appropriately called “a” (as opposed to “the”) PDF of X , because it is not unique. For example, if we modify a finite number of points, its integral is unaffected, so multiple densities can correspond to the same CDF. It turns out, however, that any two densities associated with the same CDF are equal except on a set of Lebesgue measure zero.

A PDF is in some ways similar to a PMF, except that the value cannot be interpreted as a probability. In particular, the value $f_X(x)$ can be greater than one for some x . Recall Example 8 from lecture 5. There the density was $1/(2\sqrt{t})$ over $t \in (0, 1]$ which is larger than one for small values. Instead, the proper intuitive interpretation is the fact that f_X is continuous over a small interval $[x, x + \delta]$, then

$$\mathbb{P}(x \leq X \leq x + \delta) \approx f_X(x)\delta.$$

Also it is instructive to recall *fundamental theorem of calculus*: If $F_X(x)$ is continuous and differentiable everywhere except countably many points, then

$$F_X(x) = \int_{-\infty}^x F'_X(t)dt$$

This provides a simple rule to find PDF from CDF in most cases of practical interest.

Remark: The fact that a random variable is continuous has no bearing on the continuity of X as a function from Ω into \mathbb{R} . In fact, we have not even defined what it means for a function on Ω to be continuous. But even in the special case where $\Omega = \mathbb{R}$, we can have a discontinuous function $X: \mathbb{R} \rightarrow \mathbb{R}$ which is a continuous random variable. Here is an example. Let the underlying probability measure on Ω be the Lebesgue measure on the unit interval. Let

$$X(\omega) = \begin{cases} \omega, & 0 \leq \omega \leq 1/2, \\ 1 + \omega, & 1/2 < \omega \leq 1. \end{cases}$$

The function X is discontinuous. The random variable takes values in the set $[0, 1/2] \cup (3/2, 2]$. Furthermore, it is not hard to check that X is a continuous random variable with PDF given by

$$f_X(x) = \begin{cases} 1, & x \in [0, 1/2] \cup (3/2, 2] \\ 0 & \text{otherwise.} \end{cases}$$

2 EXAMPLES

We present here a number of important examples of continuous random variables.

2.1 Uniform

This is perhaps the simplest continuous random variable. Consider an interval $[a, b]$, and let

$$F_X(x) = \begin{cases} 0, & x \leq a, \\ (x - a)/(b - a), & a < x \leq b, \\ 1, & x > b. \end{cases}$$

It is easy to check that F_X satisfies the required properties of CDFs. We denote this distribution by $U(a, b)$. We find that a corresponding PDF is given by $f_X(x) = (dF_X/dx)(x) = \frac{1}{b-a}$ for $x \in [a, b]$, and $f_X(x) = 0$, otherwise. When $[a, b] = [0, 1]$, the probability law of a uniform random variable is just the Lebesgue measure on $[0, 1]$.

2.2 Exponential

Fix some $\lambda > 0$. Let $F_X(x) = 1 - e^{-\lambda x}$, for $x \geq 0$, and $F_X(x) = 0$, for $x < 0$. It is easy to check that F_X satisfies the required properties of CDFs. A corresponding PDF is $f_X(x) = \lambda e^{-\lambda x}$, for $x \geq 0$, and $f_X(x) = 0$, for $x < 0$. We denote this distribution by $\text{Exp}(\lambda)$ and write

$$X \sim \text{Exp}(\lambda).$$

(Recall notation $\stackrel{d}{=}$ and \sim which stand for "distributed as ...")

The exponential distribution can be viewed as a "limit" of a geometric distribution. Indeed, if we fix some $\delta > 0$ and consider the values $F_X(k\delta) = 1 - e^{-\lambda\delta k}$, for $k = 1, 2, \dots$. Check that this is $\mathbb{P}(Y \leq k)$, where Y is geometrically distributed with parameter $p = 1 - e^{-\lambda\delta}$. Intuitively, the exponential distribution corresponds to a limit of a situation where every time units, we toss a coin whose success probability is δ , and let X be the time elapsed until the first success. We will revisit this intuition later on in the course.

The distribution $\text{Exp}(\lambda)$ has the following very important **memorylessness** property.

Theorem 1. *Let X be an exponentially distributed random variable. Then, for every $x, t \geq 0$, we have $\mathbb{P}(X > x + t \mid X > x) = \mathbb{P}(X > t)$.*

Proof: Let X be exponential with parameter λ . We have

$$\begin{aligned}\mathbb{P}(X > x + t \mid X > x) &= \frac{\mathbb{P}(X > x + t, X > x)}{\mathbb{P}(X > x)} = \frac{\mathbb{P}(X > x + t)}{\mathbb{P}(X > x)} \\ &= \frac{e^{-\lambda(x+t)}}{e^{-\lambda x}} = e^{-\lambda t} = \mathbb{P}(X > t).\end{aligned}$$

□

Exponential random variables are often used to model memoryless arrival processes, in which the elapsed waiting time does not affect our probabilistic model of the remaining time until an arrival. For example, suppose that the time until the next bus arrival is an exponential random variable with parameter $\lambda = 1/5$ (in minutes). Thus, there is probability¹ that you will have to wait for at least 5 minutes. Suppose that you have already waited for 10 minutes. The probability that you will have to wait for at least another five minutes is still the same, e^{-1} .

Semigroup property of exponential: let $X_1 \sim \text{Exp}(\lambda_1)$, $X_2 \sim \text{Exp}(\lambda_2)$, and $X_1 \perp\!\!\!\perp X_2$. then

$$\min(X_1, X_2) \sim \text{Exp}(\lambda_1 + \lambda_2).$$

note that \min defines a commutative operation $\mathbb{R}_+ \cup \{+\infty\}$ with $+\infty$ serving the role of identity. We can see that exponential distribution ~~related~~ min operations on $\mathbb{R}_+ \cup \{+\infty\}$: the addition of λ 's is equivalent to \min of X 's. For this statement to hold in full, one naturally understands $\text{Exp}(+\infty)$ as $X = 0$ a.s., and $X \sim \text{Exp}(0)$ as $X = +\infty$ a.s..

2.3 Normal distribution

Perhaps the most widely used distribution is the normal distribution which is also called Gaussian distribution. It involves parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, and the density

$$X \sim N(\mu, \sigma^2) : f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Note also that this PDF is symmetric around μ . Namely $f_X(\mu + x) = f_X(\mu - x)$ for every $x \in \mathbb{R}$. We need to show that this is a legitimate PDF, i.e., that it integrates to one. The special case $\mu = 0, \sigma = 1$ corresponds to the claim

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = 1 \tag{2}$$

and will be established later, when we deal with transformation of random variables. For now let us assume this and show that the same applies to the case of general μ, σ . We introduce a change of variables $t = (x - \mu)/\sigma$ implying $dx = dt/\sigma$. The range $x \in (-\infty, +\infty)$ implies the range $t \in (-\infty, +\infty)$. Then

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} = 1.$$

We use the notation $N(\mu, \sigma^2)$ to denote the normal distribution with parameters μ, σ . The distribution $N(0, 1)$ is referred to as the **standard normal** distribution; a corresponding random variable is also said to be standard normal.

There is no closed form formula for the corresponding CDF, but numerical tables are available. These tables can also be used to find probabilities associated with general normal variables. This is because of the fact that if $X \sim N(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim N(0, 1)$. Thus,

$$\mathbb{P}(X \leq c) = \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \frac{c - \mu}{\sigma}\right) = \Phi((c - \mu)/\sigma),$$

where Φ is the CDF of the standard normal, available from the normal tables.

Semigroup property of normal: Let $X_1 \sim N(0, \sigma_1^2)$, $X_2 \sim N(0, \sigma_2^2)$, and $X_1 \perp\!\!\!\perp X_2$. Then

$$X_1 + X_2 \sim N(0, \sigma_1^2 + \sigma_2^2).$$

2.4 Cauchy distribution

Here, there is only one parameter and

$$X \sim \text{Ca}(t) : f_X(x) = \frac{1}{\pi} \frac{t}{t^2 + x^2}, x \in \mathbb{R}$$

It is an exercise in calculus to show $\int_{-\infty}^{\infty} f(t)dt = 1$, so that f_X is indeed a PDF. The corresponding distribution is called a Cauchy distribution.

Semigroup property of Cauchy: Let $X_1 \sim \text{Ca}(t_1)$, $X_2 \sim \text{Ca}(t_2)$ and $X_1 \perp\!\!\!\perp X_2$. Then

$$X_1 + X_2 \sim \text{Ca}(t_1 + t_2)$$

2.5 Gamma distribution

Gamma distribution is parameterized by two positive reals: shape parameter $a > 0$ and (inverse) scale parameter $c > 0$.

$$X \sim \Gamma(a, c) : f_X(x) = \frac{c^a x^{a-1} e^{-cx}}{\Gamma(a)}, x > 0$$

Semigroup property of Gamma: Let $X_1 \sim \Gamma(a_1, c)$, $X_2 \sim \Gamma(a_2, c)$ and $X_1 \perp\!\!\!\perp X_2$. Then

$$X_1 + X_2 \sim \Gamma(a_1 + a_2, c)$$

2.6 Power law

We have already defined discrete power law distributions. We present here a continuous analog. Our starting point is to introduce tail probabilities that decay according to power law $\mathbb{P}(X > x) = \beta/x^\alpha$, for $x \geq c > 0$, for some parameters $\alpha, c > 0$. In this case, the CDF is given $F_X(x) = 1 - \beta/x^\alpha$, $x \geq c$, and $F_X(x) = 0$, otherwise. In order for X to be a continuous random variable, F_X cannot have a jump at $x = c$, and we therefore need $c = c^\alpha$ and $F_X(x) = 1 - c^\alpha/x^\alpha$. The corresponding density is of the form

$$f_X(t) = \frac{dF_X}{dx}(t) = \frac{\alpha c^\alpha}{t^{\alpha+1}}.$$

3 EXPECTED VALUES

Similar to the discrete case, given a continuous random variable with PDF f_X , we have a simple rule to compute the expectation:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

(This was shown in Lecture 8.) This integral is well defined and finite if $\int_{-\infty}^{\infty} f_X(x) dx < \infty$, in which case we say that the random variable is **integrable**. The integral is also well defined, but infinite, if one, but not both, of the integrals $\int_{-\infty}^0 x f_X(x) dx$ and $\int_0^{\infty} x f_X(x) dx$ is infinite. If both of these integrals are infinite, the expected value is not defined.

Practically all of the results developed for discrete random variables carry over to the continuous case. Many of them, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, have the exact same form. We list below two results in which sums need to be replaced by integrals.

Proposition 1. Let X be a nonnegative random variable, i.e., $\mathbb{P}(X < 0) = 0$. Then

$$\mathbb{E}[X] = \int_0^{\infty} (1 - F_X(t)) dt.$$

Proof: We have

$$\begin{aligned}\int_0^\infty (1 - F_X(t)) dt &= \int_0^\infty \mathbb{P}(X > t) dt = \int_{\mathbb{R}_+} dt \int_\Omega 1\{X(\omega) > t\} d\mathbb{P}(\omega) \\ &= \int_\Omega d\mathbb{P}(\omega) \left[\int_{\mathbb{R}_+} 1\{t < X(\omega)\} dt \right] = \int_\Omega X(\omega) d\mathbb{P} \triangleq \mathbb{E}[X]\end{aligned}$$

(The interchange of the order of integration is by Fubini's theorem for non-negative functions.) \square

Proposition 2. Let X be a continuous random variable with density, and suppose that $g : \mathbb{R} \rightarrow \mathbb{R}$ is a (Borel) measurable function. Then

$$\mathbb{E}[g(X)] = \int_{-\infty}^\infty g(t) f_X(t) dt$$

(i.e. the integral and the expectation exist or do not exist simultaneously, and are equal in the latter case).

Proof: This was shown in Lecture 8. \square

Note that for this result to hold, the random variables need not be continuous.

4 JOINT DISTRIBUTIONS

Definition 1. Given a pair of random variables X and Y , defined on the same probability space, their **joint distribution** $\mathbb{P}_{X,Y}$ is a probability measure on $(\mathbb{R} \times \mathbb{R}, \mathcal{B} \times \mathcal{B})$ defined as

$$\mathbb{P}_{X,Y}[B] \triangleq \mathbb{P}[(X, Y) \in B]$$

for every $B \in \mathcal{B} \times \mathcal{B}$.

Exercise: Show that set $\{\omega : (X(\omega), Y(\omega)) \in B\}$ is measurable for any $B \in \mathcal{B} \times \mathcal{B}$. (This provides another justification for the definition of product algebra.)

The **joint CDF** of X, Y is defined as

$$F_{X,Y}(x, y) = \mathbb{P}[X \leq x, Y \leq y]$$

and we say that X, Y are **jointly continuous** if there exists a measurable $f_{X,Y} :$

$\mathbb{R}^2 \rightarrow [0, \infty)$ such that their joint CDF satisfies

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) du dv.$$

The function $f_{X,Y}$ is called a joint PDF of X and Y .

At those points where a joint PDF is continuous, we have

$$\frac{\partial^2 F}{\partial x \partial y}(x, y) = \frac{\partial^2 F}{\partial y \partial x}(x, y) = f_{X,Y}(x, y).$$

Similar to what was mentioned for the case of a single random variable, for any Borel subset B of \mathbb{R}^2 , we have

$$\mathbb{P}_{X,Y}[B] = \int_B f_{X,Y}(x, y) dx dy = \int_{\mathbb{R}^2} 1_B(x, y) f_{X,Y}(x, y) dx dy. \quad (3)$$

Furthermore, if B has Lebesgue measure zero, then $\mathbb{P}_{X,Y}(B) = 0$.

We observe that by (3) and Fubini's theorem

$$\mathbb{P}(X \leq x) = \int_{-\infty}^x \int_{-\infty}^{\infty} f_{X,Y}(u, v) du dv.$$

Thus, X itself is a continuous random variable, with **marginal PDF**

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy.$$

We have just argued that X and Y are jointly continuous, then X (and, similarly, Y) is a continuous random variables. The converse is not true. For a trivial counterexample, let X be a continuous random variable, and let $Y = X$. Then the set $\{(x, y) \in \mathbb{R}^2 \mid x = y\}$ has zero area (zero Lebesgue measure), but unit probability, which is impossible for jointly continuous random variables. In particular, the corresponding probability law \mathbb{P}^2 is neither discrete nor continuous, hence qualifies as “singular.”

Proposition 2 has a natural extension to the case of multiple random variables.

Proposition 3. Let X and Y be jointly continuous with PDF $f_{X,Y}$, and suppose that $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a (Borel) measurable function such that $g(X)$ is integrable. Then,

$$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u, v) f_{X,Y}(u, v) du dv.$$

4.1 Multivariate measurable functions

A non-trivial assumption is for the joint PDF_{X,Y} to be a measurable function on \mathbb{R}^2 . How can we ensure that? Of course simple functions are measurable, as are their limits, limsup's and liminf's. However, a criterion frequently used in practice is the following. (It is also an excellent exercise for getting some practice with product-algebras!)

Proposition 4. *Let $\phi(x, y)$ be a function on \mathbb{R}^2 such that*

1. *$y \mapsto \phi(x, y)$ is measurable for every fixed $x \in \mathbb{R}$*
2. *$x \mapsto \phi(x, y)$ is right-continuous for every fixed $y \in \mathbb{R}$*

Then ϕ is jointly measurable in (x, y) .

Proof. First, it is instructive to understand the proof of Borel measurability of any right-continuous function $f \mapsto f(x)$. Let $R_a = \{x \in \mathbb{Q} : f(x) < a\}$. Then for any $a \in \mathbb{R}$ it follows

$$\{f(x) < a\} = \bigcup_{\epsilon_1 > 0} \bigcap_{\epsilon_2 > 0} \bigcup_{r \in R_{a-\epsilon_1}} [r - \epsilon_2, r]. \quad (4)$$

Here we write (abusing notation) _{$\epsilon > 0$} to mean union over arbitrary sequence of $\epsilon_n \searrow 0$, so that resulting operations are countable. If (4) holds, $\{f(x) < a\}$ belongs to \mathcal{B} and thus f is Borel.

To understand (4) note that the set

$$L_b \triangleq \bigcap_{\epsilon_2 > 0} \bigcup_{r \in R_b} [r - \epsilon_2, r]$$

corresponds to all points on the real-line that are decreasing limits of elements of R_b . For a right continuous function

$$f(x) < b \Rightarrow x \in L_b \Rightarrow f(x) \leq b$$

And thus (4) follows.

Now to prove Proposition, we only need to notice that $\{\phi(r, y) < a\}$ are measurable subsets of \mathbb{R} by assumption. Hence, by setting

$$L_b \triangleq \bigcap_{\epsilon_2 > 0} \bigcup_{r \in \mathbb{Q}} [r - \epsilon_2, r] \times \{y : \phi(r, y) < b\}$$

²It may also be helpful to remember that right-continuity $f : \mathbb{R} \rightarrow \mathbb{R}$ is equivalent to usual continuity when topology on the domain is refined declaring sets open.

we infer that

$$\phi(x, y) < b \Rightarrow (x, y) \in L_b \Rightarrow \phi(x, y) \leq b.$$

Thus, we have

$$\{(x, y) : \phi(x, y) < a\} = \bigcup_{\epsilon > 0} L_{a-\epsilon}$$

which is a countable combination of measurable rectangles. \square

5 INDEPENDENCE

Recall that two random variables X and Y , are said to be independent if for any two Borel subsets B_1 and B_2 , of the real line, we have $\mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}(X \in B_1)\mathbb{P}(Y \in B_2)$. This is equivalent to saying $\mathbb{P}_{X,Y} = \mathbb{P}_X \times \mathbb{P}_Y$, which explains why product of measures corresponds to independence.

Similar to the discrete case (cf. Proposition 1 and Theorem 1 in Section 3 of Lecture 5), simpler criteria for independence are available.

Theorem 2. Let X and Y be jointly continuous random variables defined on the same probability space. The following are equivalent.

- (a) The random variables X and Y are independent.
- (b) For any $x, y \in \mathbb{R}$, the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent.
- (c) For any $x, y \in \mathbb{R}$, we have $F_{X,Y}(x, y) = F_X(x)F_Y(y)$.
- (d) If f_X , f_Y , and $f_{X,Y}$ are corresponding marginal and joint densities, then $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, for all (x, y) except possibly on a set that has Lebesgue measure zero.

The proof parallels the proofs in Lecture 6, except for the last condition, for which the argument is simple when the densities are continuous functions (simply differentiate the CDF), but requires more care otherwise.

6 RADON-NIKODYM DERIVATIVE

In this section, we address a natural question: Given a random variable X (or X, Y) how do we know if it is (jointly) continuous?

Notice that Lebesgue measure plays a distinguished role in the definition of continuity. Thus a more general approach requires the following definition:

Definition 2. Let $(\Omega, \mathcal{F}, \lambda)$ be a measure space. Let μ be another measure on (Ω, \mathcal{F}) . Then function $f : \Omega \rightarrow \mathbb{R}_+$ is called a Radon-Nikodym derivative $\frac{d\mu}{d\lambda}$ if

$$\mu[E] = \int_E f d\lambda$$

for any $E \in \mathcal{F}$.

According to this definition X is a continuous random variable if and only if there exists a Radon-Nikodym derivative $\frac{d\mathbb{P}_X}{d\text{Leb}}$ on \mathbb{R} . Similarly, X and Y are jointly continuous if $\frac{d\mathbb{P}_{X,Y}}{d\text{Leb}}$ exists on \mathbb{R}^2 , etc. One simple consequence of (1) is that X cannot be a continuous random variable if it has atoms, namely if $\mathbb{P}[X = a] \neq 0$ for some $a \in \mathbb{R}$. However, as “singular” example in Section 4 shows the absence of atoms is not sufficient for continuity. The following definition and Theorem describe the necessary and sufficient condition:

Definition 3. Measure μ is absolutely continuous with respect to λ (notation: $\mu \ll \lambda$), if for every E

$$\lambda(E) = 0 \Rightarrow \mu(E) = 0.$$

Note that from (1) we see: X is continuous then $\mathbb{P}_X \ll \text{Leb}$ and similarly for joint continuity. Remarkably, the converse holds as well:

Theorem 3(Radon-Nikodym). Let μ and λ be σ -finite measures on (Ω, \mathcal{F}) . There exists a Radon-Nikodym derivative $\frac{d\mu}{d\lambda}$ if and only if $\mu \ll \lambda$.

Proof of this theorem is outside of the scope of this class (mainly it relies on some basic properties of Hilbert spaces).

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Fall 2018

Lecture 11

CONTINUOUS RANDOM VARIABLES - II

Contents

1. Review of joint distributions
2. From conditional distribution to joint (Markov kernels)
3. From joint to conditional (disintegration)
4. Example: The bivariate normal distribution
5. Conditional expectation
6. Mixed versions of Bayes' rule

1 REVIEW OF JOINT DISTRIBUTIONS

Recall that two random variables X and Y are said to be jointly continuous if there exists a nonnegative measurable function $f_{X,Y}$ such that

$$\mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) dv du.$$

Once we have in our hands a general definition of integrals, this can be used to establish that for every Borel subset of \mathbb{R}^2 , we have

$$\mathbb{P}((X, Y) \in B) = \int_B f_{X,Y}(u, v) du dv.$$

Furthermore, X is itself a continuous random variable, with density f_X given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy.$$

Finally, recall that $\mathbb{E}[g(X)] = \int g(x)f_X(x)dx$. Similar to the discrete case, the expectation of $g(X) = X^m$ and $g(X) = (X - \mathbb{E}[X])^m$ is called the m th moment and the m th central moment, respectively, of X . In particular, $\text{var}(X) \triangleq \mathbb{E}[(X - \mathbb{E}[X])^2]$ is the variance of X .

We note that all of the definitions and formulas have obvious extensions to the case of more than two random variables.

2 MARKOV KERNELS

Random variables X and Y endowed with a product measures $\mathbb{P}_X \times \mathbb{P}_Y$ are necessarily independent $X \perp\!\!\!\perp Y$. How do we construct $\mathbb{P}_{X,Y}$ for dependent variables? One method is to define X and Y on the same probability space and compute $\mathbb{P}_{X,Y}$ using the Definition given in Lecture 10. Another method involves the following concept:

Definition 1. $K : \Omega_1 \times \mathcal{F}_2 \rightarrow [0, 1]$ is called a transition probability kernel (or a Markov kernel) acting from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$ if:

1. $K(\omega_1, \cdot)$ is a probability measure on $(\Omega_2, \mathcal{F}_2)$ for each $\omega_1 \in \Omega_1$
2. $\omega_1 \mapsto K(\omega_1, B)$ is an \mathcal{F}_1 -measurable function for each $B \in \mathcal{F}_2$.

In some disciplines, it is common to abuse notation and say “Let $K : (\Omega_1, \mathcal{F}_1) \rightarrow (\Omega_2, \mathcal{F}_2)$ be a Markov kernel” or even “Let $K : \Omega_1 \rightarrow \Omega_2$ be a Markov kernel”, even though K is not a map between spaces.

Example: When Ω_1 and Ω_2 are finite, any Markov kernel K acting from $(\Omega_1, 2^{\Omega_1})$ to $(\Omega_2, 2^{\Omega_2})$ is simply an $|\Omega_1| \times |\Omega_2|$ matrix of non-negative values with row-sums all equal to 1. Such matrices are called stochastic (or right-stochastic, or row-stochastic).

Example: The following transition probability kernel acts between $(\mathbb{R}, \mathcal{B})$ and $(\mathbb{R}, \mathcal{B})$. It is called the *additive Gaussian noise channel*:

$$K(x, dy) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}} dy \quad x, y \in \mathbb{R}.$$

This kernel “blurs” every point into a Gaussian cloud of width σ .

Theorem 1. For any probability measure \mathbb{P}_1 and transition probability kernel K there exists a unique probability measure π (denoted $\mathbb{P}_1 \times K$) on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$ such that

$$\pi(A \times B) = \int_A K(\omega_1, B) \mathbb{P}_1(d\omega_1).$$

Furthermore, whenever $f \geq 0$ or f is π -integrable we have

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) d\pi = \int_{\Omega_1} \mathbb{P}_1(d\omega_1) \int_{\Omega_2} f(\omega_1, \omega_2) K(\omega_1, d\omega_2). \quad (1)$$

Proof. Repeat the steps in the proofs of Theorems 2 and 3 in Lecture 9 with trivial modifications. \square

The measure π on $\Omega_1 \times \Omega_2$ corresponds to the following stochastic experiment:

- Draw ω_1 in accordance with distribution $\mathbb{P}_1(\cdot)$.
- Then draw ω_2 in accordance with distribution $K(\omega_1, \cdot)$.
- Output pair (ω_1, ω_2) .

Caution: Many different kernels can lead to the same product measure, i.e.

$$\mathbb{P}_1 \times K = \mathbb{P}_1 \times K' \neq K = K'.$$

Indeed if $\mathbb{P}_1(A) = 0$, then $K(x, \cdot)$ can be defined arbitrarily for all $x \in A$ without affecting the product measure.

2.1 Measure-kernel-function

Markov kernels can act on functions and on measures and these actions are associative.

Proposition 1. Let K be a Markov kernel from $(\Omega_1, \mathcal{F}_1)$ to $(\Omega_2, \mathcal{F}_2)$. Then

1. The kernel K pulls back non-negative functions f on Ω_2 to non-negative functions on Ω_1 :

$$(Kf)(\omega_1) \triangleq \int_{\Omega_2} f(\omega_2) K(\omega_1, d\omega_2),$$

and this map $\omega_1 \mapsto (Kf)(\omega_1)$ is \mathcal{F}_1 measurable.

2. The kernel K pushes forward probability measures from Ω_1 to Ω_2 . Namely for each μ on $(\Omega_1, \mathcal{F}_1)$ there exists a unique probability measure $\nu = \mu K$ on $(\Omega_2, \mathcal{F}_2)$ satisfying

$$\nu(B) = \int_{\Omega_1} K(\omega_1, B) d\mu. \quad (2)$$

3. These actions are compatible: for any μ on Ω_1 and $f \geq 0$ on Ω_2

$$\int_{\Omega_1} (Kf)(\omega_1) d\mu = \int_{\Omega_2} f(\omega_2) d\nu. \quad (3)$$

Proof. We only sketch the details. For 1 notice that measurability of Kf for simple functions follows from the definition of Markov kernel. This extends to general functions by taking limits. For 2 notice that by the MCT assignment (2) indeed defines a σ -additive probability measure. Finally, 3 is obvious for simple functions and otherwise take limits. \square

It is common to denote the integral $\int f d\mu$ as μf or $\mu(f)$, i.e. the action of μ on f . In such notation, result (3) can be stated as

$$(\mu K)f = \mu(Kf), \quad (4)$$

and this justifies the so-called *measure-kernel-function* notation: $\mu K f$ (without parentheses). When Ω_1 and Ω_2 are finite it is customary to represent a measures μ as a row-vector, a kernel K as a stochastic matrix and a function f as a column vector. In that case, (4) is equivalent to associativity of matrix multiplication.

2.2 Conditional CDFs and PDFs

Here we give a general method for constructing Markov kernels (and via Theorem 1 – joint distributions $\mathbb{P}_{X,Y}$).

Proposition 2. *The following define Markov kernels acting from $(\mathbb{R}, \mathcal{B})$ to itself:*

- (a) *Let $f_{X|Y}(x|y)$ be a non-negative function jointly measurable in (x, y) and satisfying¹*

$$\int_{\mathbb{R}} f_{X|Y}(x|y) dx = 1 \quad y \in \mathbb{R}, \quad (5)$$

then

$$K(y, dx) = f_{X|Y}(x|y) dx \quad (6)$$

defines a Markov kernel.

¹Such functions are known as conditional PDFs.

- (b) Let $F_{X|Y}(x|y)$ be a function jointly measurable in (x, y) , such that $F_{X|Y}(\cdot|y)$ is a CDF² for every $y \in \mathbb{R}$. Then there exists a unique Markov kernel s.t.

$$K(y, (a, b]) = F_{X|Y}(b|y) - F_{X|Y}(a|y). \quad (7)$$

Proof. Part (a) is easy: (6) is a measure for every fixed y by (5). The function

$$y \mapsto \int_B f_{X|Y}(x|y) dx$$

is measurable for every $B \in \mathcal{B}$ by Fubini's theorem. For Part (b) again (7) extends to a unique probability measure. We need to verify that the map

$$y \mapsto K(y, B)$$

is measurable for every B . For $B = \bigcup_{i=1}^n (a_i, b_i]$ – a finite disjoint union of intervals – this follows from (7) and measurability of finite sums. Then define the collection:

$$\mathcal{L} = \{B \in \mathcal{B} : y \mapsto K(y, B) \text{-measurable function}\}.$$

We have shown that \mathcal{L} contains algebra of finite unions of intervals $(a, b]$. It is easy to show that \mathcal{L} is a monotone class. Thus, $\mathcal{L} = \mathcal{B}$ and we are done. \square

Example. Take PDF f_Y and conditional PDF $f_{X|Y}$. Let

$$\mathbb{P}_Y(dy) = f_Y(y)dy \quad (8)$$

$$K(y, dx) = f_{X|Y}(x|y)dx \quad (9)$$

Then the product measure $\pi = \mathbb{P}_Y \times K$ constructed in Theorem 1 satisfies

$$\pi(dx dy) = f_{X|Y}(x|y)f_Y(y) dx dy$$

In particular, π is a jointly continuous distribution with density $f_{X,Y} = f_{X|Y}f_Y$.

3 DISINTEGRATION OF JOINT DISTRIBUTIONS

Main question we will address here: given $\mathbb{P}_{X,Y}$ does there exist \mathbb{P}_Y and K such that $\mathbb{P}_{X,Y} = \mathbb{P}_Y \times K$?

²Such functions $F_{X|Y}$ are known as conditional CDFs.

Definition 2. Let $\mathbb{P}_{X,Y}$ be a probability measure on $(\mathbb{R}^2, \mathcal{B}^2)$ with marginal \mathbb{P}_Y . A Markov kernel $K(y, \cdot)$ is called a regular branch of **conditional probability** for X given Y (denoted $\mathbb{P}_{X|Y}(\cdot|y)$) if

$$\mathbb{P}_{X,Y} = \mathbb{P}_Y \times K \quad (10)$$

in the sense of Theorem 1. Equivalently, if

$$\mathbb{P}_{X,Y}[A \times B] = \int_B K(y, A) d\mathbb{P}_Y \quad (11)$$

for all $A, B \in \mathcal{B}$. Furthermore, if K is defined via (6) then $f_{X|Y}$ is called a **conditional PDF**, and if K is defined via (7) then $F_{X|Y}$ is called a **conditional CDF**.

Note: One should not confuse “a regular branch of conditional probability” (which is a Markov kernel) with conditional probability $\mathbb{P}[X \in A|Y]$ (which is a random variable; see below). It should also be clear that neither $\mathbb{P}_{X|Y}$ (a kernel), nor $f_{X|Y}$ (a conditional PDF, when it exists), nor $F_{X|Y}$ (a conditional CDF) are unique. Finally, equivalence of (10) and (11) follows from the fact that $\{A \times B\}$ is a generating p -system for $\mathcal{B} \times \mathcal{B}$.

3.1 Simple case: jointly-continuous $\mathbb{P}_{X,Y}$

For the case of discrete random variables, the conditional CDF is defined by $F_{X|Y}(x|y) = \mathbb{P}(X \leq x | Y = y)$, for any y such that $\mathbb{P}(Y = y) > 0$. However, this definition cannot be extended to the continuous case because $\mathbb{P}(Y = y) = 0$, for every y . Instead, we should think of $F_{X|Y}(x|y)$ as a limit of $\mathbb{P}(X \leq x | y \leq Y \leq y + \delta)$, as δ decreases to zero. Note that

$$\begin{aligned} F_{X|Y}(x|y) &\approx \mathbb{P}(X \leq x | y \leq Y \leq y + \delta) \\ &= \frac{\mathbb{P}(X \leq x, y \leq Y \leq y + \delta)}{\mathbb{P}(y \leq Y \leq y + \delta)} \\ &\approx \frac{\int_{-\infty}^x \int_y^{y+\delta} f_{X,Y}(u, v) dv du}{\delta f_Y(y)} \\ &\approx \frac{\delta \int_{-\infty}^x f_{X,Y}(u, y) du}{\delta f_Y(y)} \\ &= \frac{\int_{-\infty}^x f_{X,Y}(u, y) du}{f_Y(y)}. \end{aligned}$$

This heuristic motivates the next result.

Proposition 3. *Let $f_{X,Y}$ be a joint PDF. Then*

- (a) *Let f_Y be (any) marginal PDF of Y . Then the following is a conditional CDF of X given Y*

$$F_{X|Y}(x|y) = \int_{-\infty}^x \frac{f_{X,Y}(u,y)}{f_Y(y)} du,$$

for every y satisfying the following: a) $f_Y(y) > 0$; b) $\int f_{X,Y}(u,y) du < \infty$; and c) $\int f_{X,Y}(u,y) du = f_Y(y)$. For other y we set $F_{X|Y}(x|y) = 1\{x \geq 0\}$.

- (b) *The following is a conditional PDF of X given Y*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

for every y such that $f_Y(y) > 0$. For any other y we set $f_{X|Y}(x|y) = 1\{0 \leq x \leq 1\}$.

Proof. Joint measurability in (x,y) follows from Fubini (in both cases). Next, it is clear that \mathbb{P}_Y -measure of all y 's satisfying conditions a)-c) is 1. Thus definition of $F_{X|Y}$ for other y 's is immaterial for (11). For “good” y 's from the DCT we have

$$\lim_{x \searrow x_0} \int_{-\infty}^x f_{X,Y}(u,y) du = \int_{-\infty}^{x_0} f_{X,Y}(u,y) du,$$

which shows that $F_{X|Y}(\cdot|y)$ is right-continuous. $F_{X|Y}$ is clearly monotone. The property $\lim_{x \rightarrow -\infty} F_{X|Y}(x|y) = 0$ follows from the DCT again. Also

$$\lim_{x \rightarrow \infty} F_{X|Y}(x|y) = \int_{-\infty}^{\infty} \frac{f_{X,Y}(u,y)}{f_Y(y)} du = 1,$$

since the integral of the numerator is exactly $f_Y(y)$, by condition c).

The proof concludes by a verification of (11) which is left as an exercise. \square

3.2 General case: arbitrary $\mathbb{P}_{X,Y}$

Theorem 2 (Disintegration). *Let $\mathbb{P}_{X,Y}$ be a probability measure on $(\mathbb{R}^2, \mathcal{B}^2)$. Then there exists a regular branch of conditional probability $\mathbb{P}_{X|Y}(\cdot|y)$ of X given Y , i.e.*

$$\mathbb{P}_{X,Y} = \mathbb{P}_Y \times \mathbb{P}_{X|Y}.$$

We will prove this result in Section 5.1. We note that similar disintegration works for product spaces other than $\mathbb{R} \times \mathbb{R}$. E.g. X can take values in any complete metric space (not just \mathbb{R}), while Y can be arbitrary. For the proof see [Cinlar, Section II.4.2].

4 EXAMPLE: THE BIVARIATE NORMAL DISTRIBUTION

Let us fix some $\rho \in (-1, 1)$ and consider the function, called the **standard bivariate normal PDF**,

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right).$$

Let X and Y be two jointly continuous random variables, defined on the same probability space, whose joint PDF is f . Therefore, their law satisfies

$$\mathbb{P}_{X,Y}(dx, dy) = f(x, y) dx dy.$$

Proposition 4. (a) The function f is indeed a PDF (integrates to 1).

(b) The marginal density of X and Y is $N(0, 1)$, the standard normal PDF.

(c) We have $\rho(X, Y) = \rho$. Also, X and Y are independent iff $\rho = 0$.

(d) The conditional density of X , given $Y = y$, is $N(\rho y, 1 - \rho^2)$. In other words,

$$K(y, dx) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(x-\rho y)^2}{2(1-\rho^2)}} dx$$

is a regular branch of conditional probability for X given Y (i.e. $\mathbb{P}_{X,Y} = N(0, 1) \times K$).

Interpretation of (d): $X = \rho Y + \sqrt{1-\rho^2} Z$, where $Y \perp\!\!\!\perp Z$ are standard normals.

Proof: We will use repeatedly the fact that $1/(\sqrt{2\pi}\sigma) \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ is a PDF (namely, the PDF of the $N(\mu, \sigma^2)$ distribution), and thus integrates to one.

(a)-(b) We note that $x^2 - 2\rho xy + y^2 = x^2 - 2\rho xy + \rho^2 y^2 + (1 - \rho^2)y^2$, and

obtain

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x, y) dx = \frac{\exp\left(-\frac{(1-\rho^2)y^2}{2(1-\rho^2)}\right)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \\ &= \frac{\exp(-y^2/2)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx \end{aligned}$$

But we recognize

$$\frac{1}{\sqrt{2\pi(1-\rho^2)}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx$$

as the PDF of the $N(\rho y, 1 - \rho^2)$ distribution. Thus, the integral of this density equals one, and we obtain

$$f_Y(y) = \frac{\exp(-y^2/2)}{\sqrt{2\pi}},$$

which is the standard normal PDF. Since $\int_{-\infty}^{\infty} f_Y(y) dy = 1$, we conclude that $f(x, y)$ integrates to one, and is a legitimate joint PDF. Furthermore, we have verified that the marginal PDF of Y (and by symmetry, also the marginal PDF of X) is the standard normal PDF, $N(0, 1)$.

- (c) We have $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY]$, since X and Y are standard normal, and therefore have zero mean. We now have

$$\mathbb{E}[XY] = \iint xyf(x, y) dy dx.$$

Applying the same trick as above, we obtain for every y ,

$$\int xf(x, y) dx = \frac{\exp(-y^2/2)}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx.$$

But

$$\frac{1}{\sqrt{2\pi(1-\rho^2)}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) dx = \rho y,$$

since this is the expected value for the $N(\rho y, 1 - \rho^2)$ distribution. Thus,

$$\mathbb{E}[XY] = \iint xyf(x, y) dx dy = \int y\rho y f_Y(y) dy = \rho \int y^2 f_Y(y) dy = \rho,$$

since the integral is the second moment of the standard normal, which is equal to one. We have established that $\text{Cov}(X, Y) = \rho$. Since the variances of X and Y are equal to unity, we obtain $\rho(X, Y) = \rho$. If X and Y are independent, then $\rho(X, Y) = 0$, implying that $\rho = 0$. Conversely, if $\rho = 0$, then

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right) = f_X(x)f_Y(y),$$

and therefore X and Y are independent. Note that the condition $\rho(X, Y) = 0$ implies independence, for the special case of the bivariate normal, whereas this implication is not always true, for general random variables.

(d) Let us now compute the conditional PDF. Using the expression for $f_Y(y)$

$$\begin{aligned} f_{X|Y}(x | y) &= \frac{f(x, y)}{f_Y(y)} \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) \sqrt{2\pi} \exp(y^2/2) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{x^2 - 2\rho xy + \rho^2 y^2}{2(1-\rho^2)}\right) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(x - \rho y)^2}{2(1-\rho^2)}\right), \end{aligned}$$

which we recognize as the $N(\rho y, 1 - \rho^2)$ PDF. □

We have discussed above the special case of a bivariate normal PDF, in which the means are zero and the variances are equal to one. More generally, the bivariate normal PDF is specified by five parameters, $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$, and is given by

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}Q(x, y)\right),$$

where

$$Q(x, y) = \frac{1}{1-\rho^2} \left[\frac{(x - \mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x - \mu_1)}{\sigma_1} \frac{(y - \mu_2)}{\sigma_2} + \frac{(y - \mu_2)^2}{\sigma_2^2} \right].$$

For this case, it can be verified that

$$\mathbb{E}[X] = \mu_1, \quad \text{var}(X) = \sigma_1^2, \quad \mathbb{E}[Y] = \mu_2, \quad \text{var}(Y) = \sigma_2^2, \quad \rho(X, Y) = \rho.$$

These properties can be derived by extending the tedious calculations in the preceding proof.

There is a further generalization to more than two random variables, resulting in the multivariate normal distribution. It will be carried out in a more elegant manner in a later lecture.

5 CONDITIONAL EXPECTATION

Recall that in the discrete case we defined $\mathbb{E}[X|Y = y] = \sum_{x \in \mathbb{R}} xp_{X|Y}(x|y)$. We have also defined $\mathbb{E}[X|Y]$ to be a random variable that takes the value $\mathbb{E}[X|Y = y]$, whenever $Y = y$ and $\mathbb{P}[Y = y] > 0$. The general case is more delicate:

Definition 3. Let X be integrable. A function $g(y)$ is a conditional expectation, denoted $\mathbb{E}[X|Y = y]$, of X given Y if

$$\mathbb{E}[f(Y)X] = \mathbb{E}[f(Y)g(Y)] \quad (12)$$

for every bounded measurable f . The random variable $g(Y)$, denoted $\mathbb{E}[X|Y]$, is also called a conditional expectation of Y given X . In the special case of $X = 1_B$ we write $\mathbb{P}[B|Y = y]$ or $\mathbb{P}[B|Y]$ to denote a conditional probability of B given Y .

Theorem 3. Let X, Y be random variables defined on a common probability space and X integrable.

(a) A conditional expectation $\mathbb{E}[X|Y]$ exists.

(b) If g_1 and g_2 are two conditional expectations of X given Y then

$$\mathbb{P}[g_1(Y) \neq g_2(Y)] = 0. \quad (13)$$

(c) If K is a regular branch of conditional probability of X given Y then

$$g(y) = \int_{\mathbb{R}} xK(y, dx) \quad (14)$$

is a conditional expectation $\mathbb{E}[X|Y = y]$. In particular, if a conditional PDF $f_{X|Y}$ exists then

$$g(y) = \int_{\mathbb{R}} xf_{X|Y}(x|y)dx.$$

Note: Conditional expectation is not unique. However as (13) shows – this non-uniqueness is immaterial in most cases. Nevertheless, it is a mistake (and a very common one!) to ask for the value of $\mathbb{P}[B|Y = 0]$, which can be set to anything unless $\mathbb{P}[Y = 0] > 0$. The correct question is to find a function $y \mapsto \mathbb{P}[B|Y = y]$ (defined upto almost-sure equivalence).

Proof. (a) Let $X = X^+ - X^-$ and define for any Borel set B

$$\nu^+(B) \triangleq \mathbb{E}[1_B(Y)X^+]$$

which evidently defines a finite ($\mathbb{E}[X^+] < \infty$) measure on $(\mathbb{R}, \mathcal{B})$. Furthermore, if $\mathbb{P}_Y(B) = 0$ then $\{Y \in B\}$ has probability 0 and thus $\nu \ll \mathbb{P}_Y$. By Radon-Nikodym theorem there exists a measurable function g^+ such that

$$\mathbb{E}[1_B(Y)X^+] = \mathbb{E}[g^+(Y)1_B(Y)].$$

Similarly, we may define $\nu^-(B)$ via X^- and apply Radon-Nikodym theorem to get g^- . Setting $g = g^+ - g^-$ we have for every Borel set B :

$$\mathbb{E}[1_B(Y)X] = \mathbb{E}[1_B(Y)g(Y)].$$

Thus, $g(Y)$ verifies (12) for all $f = 1_B$. By linearity of expectation (12) is also verified for all simple functions. The general case of bounded f follows by the DCT.

(b) If $g_1(Y)$ and $g_2(Y)$ are two conditional expectations then setting

$$f(Y) = 1\{g_1(Y) > g_2(Y)\} - 1\{g_1(Y) < g_2(Y)\}$$

from (12) we get

$$\mathbb{E}[|g_1(Y) - g_2(Y)|] = 0$$

implying $g_1 = g_2$ with \mathbb{P}_Y -probability 1.

(c) By definition if K is a regular branch of conditional expectation then $\mathbb{P}_{X,Y} = \mathbb{P}_Y \times K$, which by Theorem 1 implies

$$\int_{\mathbb{R}^2} \phi(x, y) \mathbb{P}_{X,Y}(dx dy) = \int_{\mathbb{R}} \mathbb{P}_Y(dy) \int_{\mathbb{R}} \phi(x, y) K(y, dx).$$

Taking $\phi(x, y) = xf(y)$ and using integrability of X property (12) follows by Fubini. \square

Example. One might expect that when X and Y are jointly continuous, then $\mathbb{E}[X | Y]$ is a continuous random variable, but this is not the case. To see this, suppose that X and Y are independent, in which case $\mathbb{E}[X | Y = y] = \mathbb{E}[X]$, which also implies that $\mathbb{E}[X | Y] = \mathbb{E}[X]$. Thus, $\mathbb{E}[X | Y]$ takes a constant value, and is therefore a trivial case of a discrete random variable.

Example. We have a stick of unit length $[0, 1]$, and break it at X , where X is uniformly distributed on $[0, 1]$. Given the value x of X , we let Y be uniformly distributed on $[0, x]$, and let Z be uniformly distributed on $[0, 1-x]$. We assume that conditioned on $X = x$, the random variables Y and Z are independent. We are interested in the distribution of Y and Z , their expected values, and the expected value of their product.

It is clear from symmetry that Y and Z have the same marginal distribution, so we focus on Y . Let us first find the joint distribution of Y and X . We have $f_X(x) = 1$, for $x \in [0, 1]$, and $f_{Y|X}(y | x) = 1/x$, for $y \in [0, x]$. Thus, the joint PDF is

$$f_{X,Y}(x, y) = f_{Y|X}(y | x) f_X(x) = \frac{1}{x} \cdot 1 = \frac{1}{x}, \quad 0 \leq y \leq x \leq 1.$$

We can now find the PDF of Y :

$$f_Y(y) = \int_0^1 f_{X,Y}(x, y) dx = \int_y^1 \frac{1}{x} dx = \log x \Big|_y^1 = \log(1/y).$$

(check that this indeed integrates to unity). Integrating by parts, we then obtain

$$\mathbb{E}[Y] = \int_0^1 y f_Y(y) dy = \int_0^1 y \log(1/y) dy = \frac{1}{4}.$$

The above calculation is more involved than necessary. For a simpler argument, simply observe that $\mathbb{E}[Y | X = x] = x/2$, since Y conditioned on $X = x$ is uniform on

$[0, x]$. In particular, $\mathbb{E}[Y | X] = X/2$. It follows that $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y | X]] = \mathbb{E}[X/2] = 1/4$.

For an alternative version of this argument, consider the random variable Y/X . Conditioned on the event $X = x$, this random variable takes values in the range $[0, 1]$, is uniformly distributed on that range, and has mean $1/2$. Thus, the conditional PDF of Y/X is not affected by the value x of X . This implies that Y/X is independent of X , and we have

$$\mathbb{E}[Y] = \mathbb{E}[(Y/X)X] = \mathbb{E}[Y/X] \cdot \mathbb{E}[X] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

To find $\mathbb{E}[YZ]$ we use the fact that, conditional on $X = x$, Y and Z are independent, and obtain

$$\begin{aligned}\mathbb{E}[YZ] &= \mathbb{E} \mathbb{E}[YZ | X] = \mathbb{E} \mathbb{E}[Y | X] \cdot \mathbb{E}[Z | X] \\ &= \mathbb{E} \left[\frac{X}{2} \cdot \frac{1-X}{2} \right] = \int_0^1 \frac{x(1-x)}{4} dx = \frac{1}{24}.\end{aligned}$$

Exercise 1. Find the joint PDF of Y and Z . Find the probability $\mathbb{P}(Y + Z \leq 1/3)$. Find $\mathbb{E}[X|Y]$, $\mathbb{E}[X|Z]$, and $\rho(Y, Z)$.

5.1 Other properties of $\mathbb{E}[\cdot | Y]$

We note that many properties of Lebesgue integration carry over without change to conditional expectation:

1. Monotonicity: $X \leq X' \Rightarrow \mathbb{E}[X|Y] \leq \mathbb{E}[X'|Y]$
2. Linearity: $\mathbb{E}[aX + bX'|Y] = a\mathbb{E}[X|Y] + b\mathbb{E}[X'|Y]$
3. MCT: $0 \leq X_n \nearrow X \Rightarrow \mathbb{E}[X_n|Y] \nearrow \mathbb{E}[X|Y]$
4. DCT: $|X_n| \leq Z$, Z -integrable, $X_n \rightarrow X$, then $\Rightarrow \mathbb{E}[X_n|Y] \rightarrow \mathbb{E}[X|Y]$
5. Fatou's lemma: $X_n \geq 0$, $\mathbb{E}[\liminf_n X_n|Y] \leq \liminf_n \mathbb{E}[X_n|Y]$
6. Jensen's inequality: f convex $\Rightarrow \mathbb{E}[f(X)|Y] \geq f(\mathbb{E}[X|Y])$

Caution: Right-hand sides of each of these implications only hold almost surely!

Proofs of all of these are simple: assume right-hand side is violated on a set E with $\mathbb{P}[Y \in E] > 0$, then using 1_E and (12) construct a counter-example to the unconditional version of the same property. As an application we prove disintegration theorem:

Proof of Theorem 2. Let $\{r_n\}_{=1}^\infty$ be enumeration of rational numbers \mathbb{Q} . Denote

$$g_n(y) \triangleq \mathbb{P}[X \in (-\infty; r_n] \mid Y = y].$$

By monotonicity property for any k and n such that $r_k \leq r_n$ we have

$$\mathbb{P}[g_k(Y) \leq g_n(Y)] = 1$$

Therefore the set

$$E_0 = \{y : g_k(y) \leq g_n(y) \quad \forall(k, n) : r_k \leq r_n\}$$

has \mathbb{P}_Y -measure 1. Similarly, sets

$$E_1 = \{y : \inf_n g_n(y) = 0\} \tag{15}$$

$$E_2 = \{y : \sup_n g_n(y) = 1\} \tag{16}$$

both also have \mathbb{P}_Y -measure 1. All together, for every y in the set

$$E \triangleq E_0 \cap E_1 \cap E_2$$

closure of the sequence of points $(r_n, g_n(y))$ on $\mathbb{R} \times [0, 1]$ is a graph of a CDF. Thus, we may define:

$$F_{X|Y}(x|y) = \begin{cases} 1\{x \geq 0\}, & y \notin E \\ \sup\{g_n(y) : r_n \leq x\}, & y \in E \end{cases}$$

Notice that

$$y \mapsto F_{X|Y}(x|y)$$

is measurable (as a countable supremum of measurable g_n 's). And the function

$$x \mapsto F_{X|Y}(x|y)$$

is right-continuous, monotonically non-decreasing and growing from 0 to 1 on \mathbb{R} . Thus by Proposition ?? function $F_{X|Y}$ is jointly measurable. Consequently, it satisfies all requirements of a conditional CDF and by Proposition 2.(b) there exists a Markov kernel $K(y, dx)$ satisfying (7). But then for every set $(-\infty, r_n] \times B$ we have

$$(\mathbb{P}_Y \times K)((-\infty, r_n] \times B) = \int_B \mathbb{P}_Y(dy)K(y, (-\infty, r_n]) \tag{17}$$

$$= \int_B g_n(y) \mathbb{P}_Y(dy) \tag{18}$$

$$= \mathbb{E}[1_{(-\infty, r_n]}(X)1_B(Y)] \tag{19}$$

$$= \mathbb{P}_{X,Y}((-\infty, r_n] \times B), \tag{20}$$

where first step is by (7), second by definition of $F_{X|Y}(r_n|y)$ and since $\mathbb{P}_Y(E) = 1$, third is by definition of g_n and (12), and fourth is just by definition of $\mathbb{P}_{X,Y}$.

Since sets $(-\infty, r_n] \times B$ form a generating p -system for $\mathcal{B} \times \mathcal{B}$ we conclude

$$\mathbb{P}_{X,Y} = \mathbb{P}_Y \times K$$

which proves the Theorem. \square

5.2 Optimality properties of conditional expectations

The conditional expectation $\mathbb{E}[X | Y]$ can be viewed as an estimate of X , based on the value of Y . In fact, it is an optimal estimate, in the sense that the mean square of the resulting estimation error, $X - \mathbb{E}[X | Y]$, is as small as possible.

Theorem 4. Suppose that $\mathbb{E}[X^2] < \infty$. Then, for any measurable function $g : \mathbb{R} \rightarrow \mathbb{R}$, we have

$$\mathbb{E}[(X - \mathbb{E}[X | Y])^2] \leq \mathbb{E}[(X - g(Y))^2].$$

Proof: We have

$$\begin{aligned} \mathbb{E}[(X - g(Y))^2] &= \mathbb{E}[(X - \mathbb{E}[X | Y])^2] + \mathbb{E}[(\mathbb{E}[X | Y] - g(Y))^2] \\ &\quad + 2\mathbb{E}[(X - \mathbb{E}[X | Y])(\mathbb{E}[X | Y] - g(Y))] \\ &\geq \mathbb{E}[(X - \mathbb{E}[X | Y])^2]. \end{aligned}$$

The inequality above is obtained by noticing that the term $\mathbb{E}[(X - g(Y))^2]$ is always nonnegative, and that the term $\mathbb{E}[(X - \mathbb{E}[X | Y])(\mathbb{E}[X | Y] - g(Y))]$ is of the form $\mathbb{E}[(X - \mathbb{E}[X | Y])\psi(Y)]$ for $\psi(Y) = \mathbb{E}[X | Y] - g(Y)$, and is therefore equal to zero, by Eq. (12). \square

Notice that the preceding proof only relies on the property (12). As we have discussed, we can view this as the defining property of conditional expectations, for general random variables. It follows that the preceding theorem is true for all kinds of random variables.

6 MIXED VERSIONS OF BAYES' RULE

Let X be an unobserved random variable, with known CDF, F_X . We observe the value of a related random variable, Y , whose distribution depends on the value of X . This dependence can be captured by a conditional CDF, $F_{Y|X}$. On the basis of the observed value y of Y , would like to make an inference on

the unknown value of X . While sometimes, this inference aims at a numerical estimate for X , the most complete answer, which includes everything that can be said about X , is the conditional distribution of X , given Y . This conditional distribution can be obtained by using an appropriate form of Bayes' rule.

When X and Y are both discrete, Bayes' rule takes the simple form

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)} = \frac{p_X(x)p_{Y|X}(y|x)}{\sum_{x'} p_X(x')p_{Y|X}(y|x')}.$$

When X and Y are both continuous, Bayes' rule takes a similar form,

$$f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)} = \frac{f_X(x)f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(x')f_{Y|X}(y|x')dx},$$

which follows readily from the definition of the conditional PDF.

It remains to consider the case where one random variable is discrete and the other continuous. Suppose that K is a discrete random variable and Z is a continuous random variable. We describe their joint distribution in terms of a function $f_{K,Z}(k,z)$ that satisfies

$$\mathbb{P}(K=k, Z \leq z) = \int_{-\infty}^z f_{K,Z}(k,t) dt.$$

We then have

$$p_K(k) = \mathbb{P}(K=k) = \int_{-\infty}^{\infty} f_{K,Z}(k,t) dt,$$

and³

$$F_Z(z) = \mathbb{P}(Z \leq z) = \sum_k \int_{-\infty}^z f_{K,Z}(k,t) dz = \int_{-\infty}^z \sum_k f_{K,Z}(k,t) dz,$$

which implies that

$$f_Z(z) = \sum_k f_{K,Z}(k,z),$$

is the PDF of Z .

Note that if $\mathbb{P}(K=k) > 0$, then

$$\mathbb{P}(Z \leq z | K=k) = \int_{-\infty}^z \frac{f_{K,Z}(k,t)}{p_K(k)} dt,$$

³The interchange of the summation and the integration can be rigorously justified, because the terms inside are nonnegative.

and therefore, it is reasonable to define

$$f_{Z|K}(z | k) = f_{K,Z}(k, z) / p_K(k).$$

Finally, for z such that $f_Z(z) > 0$, we define $p_{K|Z}(k | z) = f_{K,Z}(k, z) / f_Z(z)$, and interpret it as the conditional probability of the event $K = k$, given that $Z = z$. (Note that we are conditioning on a zero probability event; a more accurate interpretation is obtained by conditioning on the event $z \leq Z \leq z + \delta$, and let $\delta \rightarrow 0$.) With these definitions, we have

$$f_{K,Z}(k, z) = p_K(k) f_{Z|K}(z | k) = f_Z(z) p_{K|Z}(k | z),$$

for every (k, z) for which $f_{K,Z}(k, z) > 0$. By rearranging, we obtain two more versions of the Bayes' rule:

$$f_{Z|K}(z | k) = \frac{f_Z(z) p_{K|Z}(k | z)}{p_K(k)} = \frac{f_Z(z) p_{K|Z}(k | z)}{\int f_Z(z') p_{K|Z}(k | z') dz'},$$

and

$$p_{K|Z}(k | z) = \frac{p_K(k) f_{Z|K}(z | k)}{f_Z(z)} = \frac{p_K(k) f_{Z|K}(z | k)}{\sum_{k'} p_K(k') f_{Z|K}(z | k')}.$$

Note that all four versions of Bayes' rule take the exact same form; the only difference is that we use PMFs and summations for discrete random variables, as opposed to PDFs and integrals for continuous random variables.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

DERIVED DISTRIBUTIONS**Contents**

1. Functions of a single random variable
2. Multivariate transformations
3. A single function of multiple random variables
4. Maximum and minimum of random variables
5. Sum of independent random variables – Convolution

Given a random variable X with density f_X , and a measurable function g , we are often interested in the distribution (CDF, PDF, or PMF) of the random variable $Y = g(X)$. For the case of a discrete random variable X , this is straightforward:

$$p_Y(y) = \sum_{\{x \mid g(x)=y\}} p_X(x).$$

However, the case of continuous random variables is more complicated. Note that even if X is continuous, $g(X)$ is not necessarily a continuous random variable, e.g., if the range of the function g is discrete. However, in many cases, Y is continuous and its PDF can be found by following a systematic procedure.

1 FUNCTIONS OF A SINGLE RANDOM VARIABLE

The principal method for deriving the PDF $g(y)$ is the following two-step approach.

Calculation of the PDF of a Function $y = g(x)$ of a Continuous Random Variable X

(a) Calculate the CDF F_Y of Y using the formula

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \int_{\{x \mid g(x) \leq y\}} f_X(x) dx.$$

(b) Differentiate to obtain the PDF of Y :

$$f_Y(y) = \frac{dF_Y}{dy}(y).$$

Example. Let $Y = g(X) = X^2$, where X is a continuous random variable with known PDF. For any $y > 0$, we have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X^2 \leq y) \\ &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}), \end{aligned}$$

and therefore, by differentiating and using the chain rule,

$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}), \quad y > 0.$$

Example. Suppose that X is a nonnegative random variable and that $Y = \exp(X^2)$. Note that $F_Y(y) = 0$ for $y < 1$. For $y \geq 1$, we have

$$F_Y(y) = \mathbb{P}(e^{X^2} \leq y) = \mathbb{P}(X^2 \leq \log y) = \mathbb{P}(X \leq \sqrt{\log y}).$$

By differentiating and using the chain rule, we obtain

$$f_Y(y) = f_X(\sqrt{\log y}) \frac{1}{2y\sqrt{\log y}}, \quad y > 1.$$

1.1 The case of monotonic functions

The calculation in the last example can be generalized as follows. Assume that the range of the random variable X contains an open interval \mathbb{I} . Suppose that g is strictly monotone (say increasing), and also differentiable on the interval

A. Let B the set of values of $g(x)$, as x ranges over A . Let g^{-1} be the inverse function of g , so that $g(g^{-1}(y)) = y$, for $y \in B$. Then, for $y \in B$, and using the chain rule in the last step, we have

$$f_Y(y) = \frac{d}{dy} \mathbb{P}(g(X) \leq y) = \frac{d}{dy} \mathbb{P}(X \leq g^{-1}(y)) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}.$$

Recall from calculus that the derivative of an inverse function satisfies

$$\frac{dg^{-1}}{dy}(y) = \frac{1}{g'(g^{-1}(y))},$$

where g' is the derivative of g . Therefore,

$$f_Y(y) = f_X(g^{-1}(y)) \frac{1}{|g'(g^{-1}(y))|}.$$

When g is strictly monotone decreasing, the only change is a minus sign in front of $g'(g^{-1}(y))$. Thus, the two cases can be summarized in the single formula:

$$f_Y(y) = f_X(g^{-1}(y)) \frac{1}{|g'(g^{-1}(y))|}, \quad y \in B. \quad (1)$$

An easy mnemonic for remembering (and also understanding this formula) is

$$f_Y(y) |dy| = f_X(x) |dx|,$$

where x and y are related by $y = g(x)$, and since $dy = |g'(x)| \cdot |dx|$,

$$f_Y(y) |g'(x)| = f_X(x).$$

1.2 Linear functions

Consider now the special case where $g(x) = ax + b$, i.e., $Y = aX + b$. We assume that $a \neq 0$. Then, $g'(x) = a$ and $g^{-1}(y) = (y - b)/a$. We obtain

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right).$$

Example. (A linear function of a normal random variable is not normal) Suppose that $X \stackrel{d}{=} N(0, 1)$ and $Y = aX + b$. Then,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}|a|} e^{-\frac{(y-b)^2}{2a^2}},$$

so that Y is $N(b, a^2)$. More generally, if $X \stackrel{d}{=} N(\mu, \sigma)$, then the same argument shows that $Y = ax + b \stackrel{d}{=} N(a\mu + b, a^2\sigma^2)$. We conclude that a linear (more precisely, affine) function of a normal random variable is normal.

1.3 The general case

Definition 1. Consider two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ and a measurable function $g : \Omega_1 \rightarrow \Omega_2$. The **pushforward** $g_*\mu$ of measure μ on $(\Omega_1, \mathcal{F}_1)$ along g is a measure on $(\Omega_2, \mathcal{F}_2)$ defined as

$$g_*\mu(E) \triangleq \mu(g^{-1}(E)) \quad \forall E \in \mathcal{F}_2.$$

To relate this general notion to the discussion above, we notice that

$$Y = g(X) \Rightarrow \mathbb{P}_Y = g_*\mathbb{P}_X$$

To emphasize the fact that “carries” μ into $\rho = g_*\mu$ we may schematically denote

$$d\mu \xrightarrow{g} d\rho \tag{2}$$

We establish some simple facts of pushforwards:

Proposition 1. Let $\rho = g_*\mu$. Then

(i) (Change of variable formula) For any measurable $\Omega_2 \rightarrow \mathbb{R}$ we have

$$\int_{\Omega_2} f(\omega_2) d\rho = \int_{\Omega_1} f(g(\omega_1)) d\mu \tag{3}$$

and both integrals exist or do not exist simultaneously.

(ii) For arbitrary non-negative $\phi(\omega_2)$ measure $\phi(g(\omega_1)) d\mu$ pushes forward to $\phi(\omega_2) d\rho$. Schematically:

$$\phi \circ g d\mu \xrightarrow{g} \phi d\rho \tag{4}$$

Proof: (i) is easy: if $f = 1_E$ this follows from the definition of μ . Linearity of integration implies the statement for simple functions. Finally, splitting $f^+ - f^-$ and approximating both f^+ and f^- by simple functions we conclude by invoking the MCT.

(ii) follows from the application of (3) with $1_E \cdot \phi$:

$$\int_E \phi(\omega_2) d\rho = \int_{g^{-1}E} \phi \circ g(\omega_1) d\mu$$

□

Note: Consider a special case with $\Omega_1 = \Omega_2 = \mathbb{R}$ and $d\mu = dx$ – Lebesgue measure. Then whenever g – continuously differentiable with non-vanishing derivative on an open set we have

$$1_U(x)|g'(x)| dx \xrightarrow{g} 1_{g(U)}(y) dy \quad (5)$$

This result will be established in multiple dimensions below. For now -we apply (4) with $\phi(x) = \frac{f_X \circ g^{-1}}{|g' \circ g^{-1}|}$ to get:

$$f_X(x) dx \xrightarrow{g} \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|} dy$$

which exactly means the formula for given in (1).

2 MULTIVARIATE TRANSFORMATIONS

Suppose now that $X = (X_1, \dots, X_n)$ is a vector of random variables that are jointly continuous, with joint PDF $f_X(x) = f_X(x_1, \dots, x_n)$. Consider a function $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$, and the random vector $Y = (Y_1, \dots, Y_n) = g(X)$. Let g_i be the components of g , so that $Y_i = g_i(X) = g_i(X_1, \dots, X_n)$. Suppose that the function g is continuously differentiable on some open set $\Omega \subset \mathbb{R}^n$. Let $B = g(\Omega)$ be the image of Ω under g . Assume that the inverse function¹ is well-defined on B ; that is, for every $y \in B$, there is a unique $x \in \mathbb{R}^n$ such that $g(x) = y$.

The formula that we develop here is an extension of the formula $|g'(x)| = f_X(x)$ that we derived for the one-dimensional case, with the derivative being replaced by a matrix of partial derivatives, and with the absolute value being replaced by the absolute value of the determinant. It can be justified by appealing to the change of variables theorem from multivariate calculus, but we provide here a more transparent argument.

¹Here and below we abuse notation and write instead of $\text{Leb}(dx)$.

2.1 Linear functions

Let us first assume that x is a linear function, of the form $x = Mx$, for some $n \times n$ matrix M . Fix some $x \in A$ and some $\delta > 0$. Consider the cube $C = [x, x + \delta]^n$, and assume that δ is small enough so that $C \subset A$. The image $D = \{Mx \mid x \in C\}$ of the cube C under the mapping g is a parallelepiped. Furthermore, the volume of D is known to be equal $|M| \cdot \delta^n$, where we use $|\cdot|$ to denote the absolute value of the determinant of a matrix.

Having fixed x , let us also fix $y = Mx$. Assuming that $f_X(x)$ is continuous at x , we have

$$\mathbb{P}(X \in C) = \int_C f_X(t) dt = f_X(x)\delta^n + o(\delta^n) \approx f_X(x)\delta^n,$$

where $o(\delta^n)$ stands for a function such that $\lim_{\delta \downarrow 0} o(\delta^n)/\delta^n = 0$, and where the symbol \approx indicates that the difference between the two sides is δ^0 . Thus,

$$\begin{aligned} f_X(x) \cdot \delta^n &\approx \mathbb{P}(X \in C) \\ &= \mathbb{P}(g(X) \in g(C)) \\ &= \mathbb{P}(Y \in D) \\ &\approx f_Y(y) \cdot \text{vol}(D) \\ &= f_Y(y) \cdot |M| \cdot \delta^n. \end{aligned}$$

Dividing by δ^n , and taking the limit as $\delta \downarrow 0$, we obtain $f_X(x) = f_Y(y) \cdot |M|$. Let us now assume that the matrix M is invertible, so that its determinant is nonzero. Using the relation $y = Mx$ and the fact that $\det(M^{-1}) = 1/\det(M)$,

$$f_Y(y) = \frac{f_X(M^{-1}y)}{|M|} = f_X(M^{-1}y) \cdot |M^{-1}|.$$

Note that if M is not invertible, the random variable takes values in a proper subspace S of \mathbb{R}^n . Then, Y is not jointly continuous (cannot be described by a joint PDF). On the other hand, if we restrict our attention to S since S is isomorphic to \mathbb{R}^m for some $m < n$, we could describe the distribution of Y in terms of a joint PDF \mathbb{R}^m .

2.2 The general case: heuristic

Let us now generalize to the case where g is continuously differentiable at x . We define $M(x)$ as the Jacobian matrix $(\partial g_i / \partial x_j)(x)$, with entries $(\partial g_i / \partial x_j)(x)$.

The image $D = g(C)$ of the cube C is not a parallelepiped. However, from a first order Taylor series expansion, is approximately linear in the vicinity of x . It can then be shown that the ~~area~~ has volume $|M(x)| \cdot \delta^n + o(\delta^n)$. It then follows, as in the linear case, that

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|M(g^{-1}(y))|} = f_X(g^{-1}(y)) \cdot |M^{-1}(g^{-1}(y))|.$$

We note a useful fact from calculus that sometimes simplifies the application of the above formula. If we define $J(y)$ as the Jacobian (the matrix of partial derivatives) of the mapping $g^{-1}(y)$, and if some particular x and y are related by $y = g(x)$, then $J(y) = M^{-1}(x)$. Therefore,

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |J(y)|. \quad (6)$$

2.3 The general case: formal results

Our goal is to prove (6). First we state assumptions.

Assumption A. Let U be an open set \mathbb{R}^n and $g : U \rightarrow \mathbb{R}^n$ be continuously differentiable injection with non-vanishing Jacobian $\frac{\partial g}{\partial x} \neq 0$.

It is well known that if $U \rightarrow g(U)$ satisfies Assumption A, then (U) is open and the inverse map g^{-1} satisfies Assumption A on (U) .

We first remind the following important result from calculus: for any continuous $f : g(U) \rightarrow \mathbb{R}$ with compact support we have

$$\mathcal{R}\int_{g(U)} f(\mathbf{y}) d\mathbf{y} = \mathcal{R}\int_U f(g(\mathbf{x})) \left| \frac{\partial g}{\partial \mathbf{x}} \right| d\mathbf{x} \quad (7)$$

where $\mathcal{R}\int$ denotes the Riemann integral.

We now use this fact to establish the following:

Theorem 1(Jacobian formula) *If U and $g : U \mapsto g(U)$ satisfy Assumption A then*

$$1_U(\mathbf{x}) \left| \frac{\partial g}{\partial \mathbf{x}} \right| d\mathbf{x} \xrightarrow{g} 1_{g(U)}(\mathbf{y}) d\mathbf{y}$$

Applying Theorem 1 with (3) and (4) we can establish a number of consequences (for simplicity we assume $\mathbb{E} = \mathbb{R}^n$):

$$\left| \frac{\partial g}{\partial \mathbf{x}} \right| d\mathbf{x} \xrightarrow{g} d\mathbf{y} \quad (8)$$

$$d\mathbf{x} \xrightarrow{g} \left| \frac{\partial g^{-1}}{\partial \mathbf{y}} \right| d\mathbf{y} \quad (9)$$

$$f_X(\mathbf{x}) d\mathbf{x} \xrightarrow{g} f_X(g^{-1}(\mathbf{y})) \left| \frac{\partial g^{-1}}{\partial \mathbf{y}} \right| d\mathbf{y} \quad (10)$$

where we also used the fact that $\left| \frac{\partial g^{-1}}{\partial \mathbf{y}} \right| = \left| \frac{\partial g}{\partial \mathbf{x}} \right|^{-1}$ for $\mathbf{y} = g(\mathbf{x})$. In particular (10) implies (6). When g satisfies Assumption A on disjoint U_1 and U_2 then we have

$$1_{U_1 \cup U_2}(\mathbf{x}) \left| \frac{\partial g}{\partial \mathbf{x}} \right| d\mathbf{x} \xrightarrow{g} \{1_{g(U_1)}(\mathbf{y}) + 1_{g(U_2)}(\mathbf{y})\} d\mathbf{y}$$

which may be useful to find pushforwards along many-to-one maps.

Proof (optional). Let μ be a measure defined via

$$\mu(E) = \int_E 1_U(\mathbf{x}) \left| \frac{\partial g}{\partial \mathbf{x}} \right| d\mathbf{x}$$

and $\rho = g_*\mu$. We need to show that

$$\rho(E) = \text{Leb}(E \cap g(U))$$

for every Borel set E . Or, equivalently, that $\rho(E) = \text{Leb}(E)$ for every Borel $E \subset g(U)$. Under conditions of the theorem it is easy to see that $g(U)$ is an open set (since g is continuously invertible and hence every point in $g(U)$ has an open ball around itself contained in $g(U)$).

First, we have indicated before (and proved in dimension 1) that Riemann integral, when it exists, coincides with Lebesgue integral over Lebesgue measure. Thus from (7) we conclude that for every continuous function with compact support contained inside $g(U)$ we have

$$\int_{\mathbb{R}^n} f(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^n} f(g(\mathbf{y})) d\mu$$

where this time integrals are Lebesgue, and as before stands for $\text{Leb}(dy)$. By (3) this implies that for all such functions we have

$$\int_{\mathbb{R}^n} f(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^n} f(\mathbf{y}) d\rho$$

Now consider an arbitrary open ball contained in $g(U)$. Although the indicator $1_B(\mathbf{y})$ is not a continuous function, it can be easily approximated by a sequence continuous functions from below:

$$0 \leq f_n(\mathbf{y}) \nearrow 1_B(\mathbf{y})$$

Thus by the MCT we conclude that

$$\rho(B) = \text{Leb}(B)$$

for every open ball contained in $g(U)$. Now every open set $V \subset g(U)$ can be written as a union of countably many open balls. Thus by continuity of measure $\rho(V) = \text{Leb}(V)$. But two σ -finite measures coinciding on every open set must be identical (since open sets are a generating system). \square

2.4 The bivariate normal in polar coordinates

Let X and Y be independent standard normal random variables. Let g be the mapping that transforms Cartesian coordinates to polar coordinates, and let $(R, \Theta) = g(X, Y)$. The mapping g is either undefined or discontinuous at $(x, y) = (0, 0)$. So, strictly speaking, in order to apply the multivariate transformation formula (6), we should work with $\mathbb{R}^2 \setminus \{(0, 0)\}$. The inverse mapping g^{-1} is given by $(x, y) = g^{-1}(r, \theta) = (r \cos \theta, r \sin \theta)$. Its Jacobian matrix is of the form

$$\begin{bmatrix} \cos \theta & \sin \theta \\ -r \sin \theta & r \cos \theta \end{bmatrix},$$

and therefore, $|J(r, \theta)| = r \cos^2 \theta + r \sin^2 \theta = r$. From the multivariate transformation formula, we obtain

$$f_{R, \Theta}(r, \theta) = \frac{r}{2\pi} e^{-(r^2 \cos^2 \theta + r^2 \sin^2 \theta)/2} = \frac{1}{2\pi} r e^{-r^2/2}, \quad r > 0.$$

We observe that the joint PDF of R and Θ is of the form $f_{R, \Theta} = f_R f_\Theta$, where

$$f_R(r) = r e^{-r^2/2},$$

and

$$f_\Theta(\theta) = \frac{1}{2\pi}, \quad \theta \in [0, 2\pi].$$

In particular, R and Θ are independent. The random variable R is said to have a **Rayleigh** distribution.

We can also find the density $Zf = R^2$. For the mapping g defined by $g(r) = r^2$, we have $g^{-1}(z) = \sqrt{z}$, and $g'(r) = 2r$, which leads to $1/g'(g^{-1}(z)) = 1/2\sqrt{z}$. We conclude that

$$f_Z(z) = \sqrt{z}e^{-z/2} \frac{1}{2\sqrt{z}} = \frac{1}{2}e^{-z/2}, \quad z > 0$$

which we recognize as an exponential PDF with parameter $\lambda/2$.

An interesting consequence of the above results is that in order to simulate normal random variables, it suffices to generate two independent random variables, one uniform and one exponential. Furthermore, an exponential random variable is easy to generate using a nonlinear transformation of another independent uniform random variable.

3 A SINGLE FUNCTION OF MULTIPLE RANDOM VARIABLES

Suppose that $X = (X_1, \dots, X_n)$ is a vector of jointly continuous random variables with known PDF f_X . Consider a function $g_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ and the random variable $Y_1 = g_1(X)$. Note that the formulas from Section 2 cannot be used directly. In order to find the PDF of Y_1 , one possibility is to calculate the multi-dimensional integral

$$F_Y(y) = \mathbb{P}(g_1(X) \leq y) = \int_{\{x \mid g_1(x) \leq y\}} f_X(x) dx,$$

and then differentiate.

Another possibility is to introduce additional functions $g_2, \dots, g_n : \mathbb{R}^n \rightarrow \mathbb{R}$, and define $Y_i = g_i(X)$, for $i \geq 2$. As long as the resulting function $g = (g_1, \dots, g_n)$ is invertible, we can appeal to our earlier formula to find the joint PDF of Y , and then integrate to find the marginal PDF f_{Y_1} .

The simplest choice in the above described method is to let $X_i, i \neq 1$, so that $g(x) = (g_1(x), x_2, \dots, x_n)$. Let $h : \mathbb{R}^n \rightarrow \mathbb{R}$ be a function that corresponds to the first component g_1 . That is, if $y = g(x)$, then $x_1 = h(y)$. Then, the inverse mapping g^{-1} is of the form

$$g^{-1}(y) = (h(y), y_2, \dots, y_n),$$

and its Jacobian matrix is of the form

$$J(y) = \begin{bmatrix} \frac{\partial h}{\partial y_1}(y) & \frac{\partial h}{\partial y_2}(y) & \cdots & \frac{\partial h}{\partial y_n}(y) \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

It follows that $|J(y)| = \left| \frac{\partial h}{\partial y_1}(y) \right|$, and

$$f_Y(y) = f_X(h(y), y_2, \dots, y_n) \left| \frac{\partial h}{\partial y_1}(y) \right|.$$

Integrating, we obtain

$$f_{Y_1}(y_1) = \int f_X(h(y), y_2, \dots, y_n) \left| \frac{\partial h}{\partial y_1}(y) \right| dy_2 \cdots dy_n.$$

Example. Let X_1 and X_2 be positive, jointly continuous, random variables, and suppose that we wish to derive the PDF of $Y_1 = X_1 X_2$. We define $Y_2 = X_2$. From the relation $y_1 = y_1/x_2$ we see that $h(y_1, y_2) = y_1/y_2$. The partial derivative $\partial h/\partial y_1$ is $1/y_2$. We obtain

$$f_{Y_1}(y_1) = \int f_X(y_1/y_2, y_2) \frac{1}{y_2} dy_2 = \int f_X(y_1/x_2, x_2) \frac{1}{x_2} dx_2.$$

For a special case, suppose that $X_1, X_2 \stackrel{d}{=} U(0, 1)$ are independent. Their common PDF is $f_{X_i}(x_i) = 1$, for $x_i \in [0, 1]$. Note that $f_{Y_1}(y_1) = 0$ for $y_1 \notin (0, 1)$. Furthermore, $f_{X_1}(y_1/x_2)$ is positive (and equal to 1) only in the range $x_2 \geq y_1$. Also $f_{X_2}(x_2)$ is positive, and equal to 1, if $x_2 \in (0, 1)$. In particular,

$$f_X(y_1/x_2, x_2) = f_{X_1}(y_1/x_2) f_{X_2}(x_2) = 1, \quad \text{for } x_2 \geq y_1.$$

We then obtain

$$f_{Y_1}(y_1) = \int_{y_1}^1 \frac{1}{x_2} dx_2 = -\log y_1, \quad y_1 \in (0, 1).$$

The direct approach to this problem would first involve the calculation $F_{Y_1}(y_1) = \mathbb{P}(X_1 X_2 \leq y_1)$. It is actually easier to calculate

$$\begin{aligned} 1 - F_{Y_1}(y_1) &= \mathbb{P}(X_1 X_2 \geq y_1) = \int_{y_1}^1 \int_{y_1/x_1}^1 dx_2 dx_1 \\ &= \int_{y_1}^1 \left(1 - \frac{y_1}{x_1}\right) dx_1 \\ &= (x_1 - y_1 \log x_1) \Big|_{y_1}^1 = (1 - y_1) + y_1 \log y_1. \end{aligned}$$

Thus, $F_{Y_1}(y_1) = y_1 - y_1 \log y_1$. Differentiating, we find that $f_{Y_1}(y_1) = -\log y_1$.

An even easier solution for this particular problem (along the lines of the stick example in Lecture 9) is to realize that conditioned on x_1 , the random variable $Y_1 = X_1 X_2$ is uniform on $[0, x_1]$, and using the total probability theorem,

$$f_{Y_1}(y_1) = \int_{y_1}^1 f_{X_1}(x_1) f_{Y_1|X_1}(y_1 | x_1) dx_1 = \int_{y_1}^1 \frac{1}{x_1} dx_1 = -\log y_1.$$

4 MAXIMUM AND MINIMUM OF RANDOM VARIABLES

Let X_1, \dots, X_n be independent random variables, and let $X^{(1)} \leq X^{(2)} \leq \dots \leq X^{(n)}$ denote the corresponding *order statistics*. Namely, $X^{(1)} = \min_j X_j$, $X^{(2)}$ is the second smallest of the values X_1, \dots, X_n , and $X^{(n)} = \max_j X_j$. We would like to find the joint distribution of the order statistics and specifically the distribution of $\min_j X_j$ and $\max_j X_j$. Note that

$$\begin{aligned}\mathbb{P}(\max_j X_j \leq x) &= \mathbb{P}(X_1, \dots, X_n \leq x) = \mathbb{P}(X_1 \leq x) \cdots \mathbb{P}(X_n \leq x) \\ &= F_{X_1}(x) \cdots F_{X_n}(x).\end{aligned}$$

For the minimum, we have

$$\begin{aligned}\mathbb{P}(\min_j X_j \leq x) &= 1 - \mathbb{P}(\min_j X_j > x) \\ &= 1 - \mathbb{P}(X_1, \dots, X_n > x) \\ &= 1 - (1 - F_{X_1}(x)) \cdots (1 - F_{X_n}(x)).\end{aligned}$$

Let us consider the special case where X_1, \dots, X_n are i.i.d., with common CDF F and PDF f . For simplicity, assume that F is differentiable everywhere. Then,

$$\mathbb{P}(\max_j X_j \leq x) = F^n(x), \quad \mathbb{P}(\min_j X_j \leq x) = 1 - (1 - F(x))^n,$$

implying that

$$f_{\max_j X_j}(x) = nF^{n-1}(x)f(x), \quad f_{\min_j X_j}(x) = n(1 - F(x))^{n-1}f(x).$$

Exercise 1. Assuming that X_1, \dots, X_n are independent with common density function f , establish that the joint distribution of $X^{(1)}, \dots, X^{(n)}$ is given by

$$f_{X^{(1)}, \dots, X^{(n)}}(x_1, \dots, x_n) = n!f(x_1) \cdots f(x_n), \quad x_1 < x_2 < \cdots < x_n,$$

and $f_{X^{(1)}, \dots, X^{(n)}}(x_1, \dots, x_n) = 0$, otherwise. Use this to derive the densities for $\max_j X_j$ and $\min_j X_j$.

5 SUM OF INDEPENDENT RANDOM VARIABLES – CONVOLUTION

If X and Y are independent discrete random variables, the PMF of $Z = X + Y$ is easy to find:

$$\begin{aligned} p_{X+Y}(z) &= \mathbb{P}(X + Y = z) \\ &= \sum_{\{(x,y) | x+y=z\}} \mathbb{P}(X = x, Y = y) \\ &= \sum_x \mathbb{P}(X = x, Y = z - x) \\ &= \sum_x p_X(x)p_Y(z - x). \end{aligned}$$

When X and Y are independent and jointly continuous, an analogous formula can be expected to hold. We derive it in two different ways.

A first derivation involves plain calculus. Let $f_{X,Y}$ be the joint PDF of X and Y . Then,

$$\mathbb{P}(X + Y \leq z) = \int_{\{(x,y) | x+y \leq z\}} f_{X,Y}(x, y) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{X,Y}(x, y) dy dx.$$

Introduce the change of variables $t = x + y$. Then,

$$\mathbb{P}(X + Y \leq z) = \int_{-\infty}^{\infty} \int_{-\infty}^z f_{X,Y}(x, t - x) dt dx,$$

which gives

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x) dx.$$

In the special case where X and Y are independent, we have $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, resulting in

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x) dx.$$

If we were to use instead our general tools, we could proceed as follows. Consider the linear function that maps (X, Y) to $(X, X + Y)$. It is easily seen that the associated determinant is equal to 1. Thus, with $Z = X + Y$, we have

$$f_{X,Z}(x, z) = f_{X,Y}(x, z - x) = f_X(x)f_Y(z - x).$$

We then integrate over all x to obtain the marginal PDF of Z :

Exercise 2. Suppose that $X \stackrel{d}{=} N(\mu_1, \sigma_1^2)$, $X_2 \stackrel{d}{=} N(\mu_2, \sigma_2^2)$ and independent. Establish that $X_1 + X_2 \stackrel{d}{=} N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Exercise 3. Establish the semigroup properties for Gamma and Cauchy distributions mentioned in Lecture 10.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Fall 2018

Lecture 13

MOMENT GENERATING FUNCTIONS

Contents

1. Moment generating functions
2. Sum of a random number of random variables
3. Transforms associated with joint distributions

Moment generating functions, and their close relatives (probability generating functions and characteristic functions) provide an alternative way of representing a probability distribution by means of a certain function of a single variable.

These functions turn out to be useful in many different ways:

- (a) They provide an easy way ~~calculating the moments~~ of a distribution.
- (b) They provide some powerful tools for addressing ~~certain~~ counting and combinatorial problems.
- (c) They provide an easy way of characterizing the distribution ~~of~~ of independent random variables.
- (d) They provide tools for dealing with the distribution ~~of~~ of a random number of independent random variables.
- (e) They play a central role in the study ~~branching processes~~
- (f) They play a key role ~~large deviations~~ theory, that is, in studying the asymptotics of tail probabilities of the form $\Pr(X \geq c)$, when c is a large number.
- (g) They provide a bridge between complex analysis and probability, so that complex analysis methods can be brought to bear on probability problems.
- (h) They provide powerful tools for providing limit theorems such as laws of large numbers and the central limit theorem.

1 MOMENT GENERATING FUNCTIONS

1.1 Definition

Definition 1. The moment generating function associated with a random variable X is a function $M_X : \mathbb{R} \rightarrow [0, \infty]$ defined by

$$M_X(s) = \mathbb{E}[e^{sX}].$$

The domain D_X of M_X is defined as the set $D_X = \{s \mid M_X(s) < \infty\}$.

If X is a discrete random variable, with PMF, then

$$M_X(s) = \sum_x e^{sx} p_X(x).$$

If X is a continuous random variable with PDF, then

$$M_X(s) = \int e^{sx} f_X(x) dx.$$

Note that this is essentially the same as the definition of the Laplace transform of a function f_X , except that we are using instead of $-s$ in the exponent.

1.2 The domain of the moment generating function

Note that $0 \in D_X$, because $M_X(0) = \mathbb{E}[e^{0X}] = 1$. For a discrete random variable that takes only a finite number of different values, we have \mathbb{R} . For example, if X takes the values 1, 2, and 3, with probabilities $1/3$, and $1/6$, respectively, then

$$M_X(s) = \frac{1}{2}e^s + \frac{1}{3}e^{2s} + \frac{1}{6}e^{3s}, \quad (1)$$

which is finite for every $s \in \mathbb{R}$. On the other hand, for the Cauchy distribution, $f_X(x) = 1/(\pi(1+x^2))$, for all x , it is easily seen that $M_X(s) = \infty$, for all $s \neq 0$.

In general, D_X is an interval (possibly infinite or semi-infinite) that contains zero.

Exercise 1. Suppose that $M_X(s) < \infty$ for some $s > 0$. Show that $M_X(t) < \infty$ for all $t \in [0, s]$. Similarly, suppose that $M_X(s) < \infty$ for some $s < 0$. Show that $M_X(t) < \infty$ for all $t \in [s, 0]$.

Exercise 2. Suppose that

$$\limsup_{x \rightarrow \infty} \frac{\log \mathbb{P}(X > x)}{x} \triangleq -\nu < 0.$$

Establish that $M_X(s) < \infty$ for all $s \in [0, \nu)$.

1.3 Inversion of transforms

By inspection of the formula $f_M(s)$ in Eq. (1), it is clear that the distribution of X is readily determined. The various powers indicate the possible values of the random variable X , and the associated coefficients provide the corresponding probabilities.

At the other extreme, if we are told $M_X(s) = \infty$ for every $s \neq 0$, this is certainly not enough information to determine the distribution of X .

On this subject, there is the following fundamental result. It is intimately related to the inversion properties of Laplace transforms. Its proof requires sophisticated analytical machinery and is omitted.

Theorem 1. Inversion theorem Suppose that $M_X(s)$ is finite for all s in an interval of the form $[-a, a]$, where a is a positive number. Then M_X determines uniquely the CDF of the random variable X .

In particular, if $M_X(s) = M_Y(s) < \infty$, for all $s \in [-a, a]$, where a is a positive number, then the random variables X and Y have the same CDF.

There are explicit formulas that allow us to recover the PMF or PDF of a random variable starting from the associated transform, but they are quite difficult to use (e.g., involving “contour integrals”). In practice, transforms are usually inverted by “pattern matching,” based on tables of known distribution-transform pairs.

1.4 Moment generating properties

There is a reason why M_X is called a moment generating function. Let us consider the derivatives of M_X at zero. Assuming for a moment we can interchange the order of integration and differentiation, we obtain

$$\begin{aligned} \left. \frac{dM_X(s)}{ds} \right|_{s=0} &= \left. \frac{d}{ds} \mathbb{E}[e^{sX}] \right|_{s=0} = \mathbb{E}[X e^{sX}] \Big|_{s=0} = \mathbb{E}[X], \\ \left. \frac{d^m M_X(s)}{ds^m} \right|_{s=0} &= \left. \frac{d^m}{ds^m} \mathbb{E}[e^{sX}] \right|_{s=0} = \mathbb{E}[X^m e^{sX}] \Big|_{s=0} = \mathbb{E}[X^m] \end{aligned}$$

Thus, knowledge of the transform M_X allows for an easy calculation of the moments of a random variable X .

Justifying the interchange of the expectation and the differentiation does require some work. The steps are outlined in the following exercise. For simplicity, we restrict to the case of nonnegative random variables.

Exercise 3. Suppose that X is a nonnegative random variable and that $M_X(s) < \infty$ for all $s \in (-\infty, a]$, where a is a positive number.

- (a) Show that $\mathbb{E}[X^k] < \infty$, for every k .
- (b) Show that $\mathbb{E}[X^k e^{sX}] < \infty$, for every $s < a$.
- (c) Show that $(e^{hX} - 1)/h \leq X e^{hX}$.
- (d) Use the DCT to argue that

$$\mathbb{E}[X] = \mathbb{E}\left[\lim_{h \downarrow 0} \frac{e^{hX} - 1}{h}\right] = \lim_{h \downarrow 0} \frac{\mathbb{E}[e^{hX}] - 1}{h}.$$

1.5 The probability generating function

For discrete random variables, the following **probability generating function** is sometimes useful. It is defined by

$$g_X(s) = \mathbb{E}[s^X],$$

with s usually restricted to positive values. It is of course closely related to the moment generating function in that, for $s > 0$, we have $g_X(s) = M_X(\log s)$.

One difference is that when X is a positive random variable, we can define $g_X(s)$, as well as its derivatives, for $s = 0$. So, suppose that X has a PMF $p_X(m)$, for $m = 1, \dots$. Then,

$$g_X(s) = \sum_{m=1}^{\infty} s^m p_X(m),$$

resulting in

$$\left. \frac{d^m}{ds^m} g_X(s) \right|_{s=0} = m! p_X(m).$$

(The interchange of the summation and the differentiation needs justification, but is indeed legitimate for small m .) Thus, we can use g_X to easily recover the PMF p_X , when X is a positive integer random variable.

At the same time

$$\left. \frac{d}{ds} g_X(s) \right|_{s=1} = \sum_{m \geq 1} m p_X(m) = \mathbb{E}[X].$$

1.6 Examples

Example : $X \stackrel{d}{=} \text{Exp}(\lambda)$. Then,

$$M_X(s) = \int_0^\infty e^{sx} \lambda e^{-\lambda x} dx = \begin{cases} \frac{\lambda}{\lambda-s}, & s < \lambda; \\ \infty, & \text{otherwise.} \end{cases}$$

Example : $X \stackrel{d}{=} \text{Ge}(p)$

$$M_X(s) = \sum_{m=1}^{\infty} e^{sm} p(1-p)^{m-1} \begin{cases} \frac{e^s p}{1-(1-p)e^s}, & e^s < 1/(1-p); \\ \infty, & \text{otherwise.} \end{cases}$$

In this case, we also find $(s) = ps/(1 - (1 - p)s)$, $s < 1/(1 - p)$ and $g_X(s) = \infty$, otherwise.

Example : $X \stackrel{d}{=} N(0, 1)$. Then,

$$\begin{aligned} M_X(s) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(sx) \exp(-\frac{x^2}{2}) dx \\ &= \frac{\exp(s^2/2)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\frac{x^2 + 2sx - s^2}{2}) dx \\ &= \exp(s^2/2). \end{aligned}$$

1.7 Properties of moment generating functions

We record some useful properties of moment generating functions.

Theorem 2.

- (a) If $Y = aX + b$, then $M_Y(s) = e^{sb} M_X(as)$.
- (b) If X and Y are independent, then $M_{X+Y}(s) = M_X(s)M_Y(s)$.
- (c) Let X and Y be independent random variables. Let Z be equal to X , with probability p , and equal to Y , with probability $1 - p$. Then,

$$M_Z(s) = pM_X(s) + (1 - p)M_Y(s).$$

Proof: For part (a), we have

$$M_X(aX + b) = \mathbb{E}[\exp(saX + sb)] = \exp(sb)\mathbb{E}[\exp(saX)] = \exp(sb)M_X(as).$$

For part (b), we have

$$M_{X+Y}(s) = \mathbb{E}[\exp(sX + sY)] = \mathbb{E}[\exp(sX)]\mathbb{E}[\exp(sY)] = M_X(s)M_Y(s).$$

For part (c), by conditioning on the random choice between X and Y , we have

$$M_Z(s) = \mathbb{E}[e^{sZ}] = p\mathbb{E}[e^{sX}] + (1-p)\mathbb{E}[e^{sY}] = pM_X(s) + (1-p)M_Y(s).$$

□

Example : (Normal random variables)

- (a) Let X be a standard normal random variable, and let $Y = \sigma X + \mu$, which we know to have a $N(\mu, \sigma^2)$ distribution. We then find that $M_Y(s) = \exp(s\mu + \frac{1}{2}s^2\sigma^2)$.
- (b) Let $X \stackrel{d}{=} N(\mu_1, \sigma_1^2)$ and $Y = N(\mu_2, \sigma_2^2)$. Then,

$$M_{X+Y}(s) = \exp\left(s(\mu_1 + \mu_2) + \frac{1}{2}s^2(\sigma_1^2 + \sigma_2^2)\right).$$

Using the inversion property of transforms, we conclude that $Y \stackrel{d}{=} N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$, thus corroborating a result we first obtained using convolutions.

2 SUM OF A RANDOM NUMBER OF INDEPENDENT RANDOM VARIABLES

Let X_1, X_2, \dots be a sequence of i.i.d. random variables, with mean and variance μ and σ^2 . Let N be another independent random variable that takes nonnegative integer values. Let $Y = \sum_{i=1}^N X_i$. Let us derive the mean, variance, and moment generating function of Y .

We have

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y | N]] = \mathbb{E}[N\mu] = \mathbb{E}[N]\mathbb{E}[X].$$

Furthermore, using the law of total variance,

$$\begin{aligned} \text{var}(Y) &= \mathbb{E}[\text{var}(Y | N)] + \text{var}(\mathbb{E}[Y | N]) \\ &= \mathbb{E}[N\sigma^2] + \text{var}(N\mu) \\ &= \mathbb{E}[N]\sigma^2 + \mu^2\text{var}(N). \end{aligned}$$

Finally, note that

$$\mathbb{E}[\exp(sY) | N = n] = M_X^n(s) = \exp(n \log M_X(s)),$$

implying that

$$M_Y(s) = \sum_{n=1}^{\infty} \exp(n \log M_X(s)) \mathbb{P}(N = n) = M_N(\log M_X(s)).$$

The reader is encouraged to take the derivative of the above expression, and evaluate it at $s = 0$, to recover the formula $\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X]$.

Example : Suppose that each X_i is exponentially distributed, with parameter λ , and that N is geometrically distributed, with parameter $p \in (0, 1)$. We find that

$$M_Y(s) = \frac{e^{\log M_X(s)} p}{1 - e^{\log M_X(s)}(1-p)} = \frac{p\lambda/(\lambda-s)}{1 - \lambda(1-p)/(\lambda-s)} = \frac{\lambda p}{\lambda p - s}$$

which we recognize as a moment generating function of an exponential random variable with parameter λp . Using the inversion theorem, we conclude that Y is exponentially distributed. In view of the fact that the sum of a fixed number of exponential random variables is far from exponential, this result is rather surprising. An intuitive explanation will be provided later in terms of the Poisson process.

3 TRANSFORMS ASSOCIATED WITH JOINT DISTRIBUTIONS

If two random variables X and Y are described by some joint distribution (e.g., a joint PDF), then each one is associated with a transform $M_X(s)$ or $M_Y(s)$. These are the transforms of the marginal distributions and do not convey information on the dependence between the two random variables. Such information is contained in a multivariate transform, which we now define.

Consider n random variables X_1, \dots, X_n related to the same experiment. Let s_1, \dots, s_n be real parameters. The associated **multivariate transform** is a function of these parameters and is defined by

$$M_{X_1, \dots, X_n}(s_1, \dots, s_n) = \mathbb{E}[e^{s_1 X_1 + \dots + s_n X_n}].$$

The inversion property of transforms discussed earlier extends to the n -multivariate case. That is, if Y_1, \dots, Y_n is another set of random variables and $M_{X_1, \dots, X_n}(s_1, \dots, s_n), M_{Y_1, \dots, Y_n}(s_1, \dots, s_n)$ are the same functions of s_1, \dots, s_n , in a neighborhood of the origin, then the joint distribution of X_1, \dots, X_n is the same as the joint distribution of Y_1, \dots, Y_n .

Remarks:

- (a) Consider two random variables X and Y . Their joint transform is

$$M_{X,Y}(s,t) = \mathbb{E}[e^{sX}e^{tY}] = \mathbb{E}[e^{sX+tY}] = M_Z(1),$$

where $Z = sX + tY$. Thus, calculating a multivariate transform essentially amounts to calculating the univariate transform associated with a single random variable that is a linear combination of the original random variables.

- (b) If X and Y are independent, then $M_{X,Y}(s,t) = M_X(s)M_Y(t)$.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Fall 2018

Lecture 14

MULTIVARIATE NORMAL DISTRIBUTIONS

Contents

1. Background on positive definite matrices
2. Definition of the multivariate normal distribution
3. Means and covariances of vector random variables
4. Key properties of the multivariate normal

In an earlier lecture, we worked through the bivariate normal distribution and its properties, relying mostly on algebraic manipulation and integration of normal PDFs. Here, we revisit the subject in more generality (in dimensions), while using more elegant tools. First, some background.

1 BACKGROUND ON POSITIVE DEFINITE MATRICES.

Definition 1. Let A be a square $n \times n$ **symmetric** matrix.

- (a) We say that A is **positive definite** and write $A > 0$, if $x^T A x > 0$, for every nonzero $x \in \mathbb{R}^n$.
- (b) We say that A is **nonnegative definite** and write $A \geq 0$, if $x^T A x \geq 0$, for every $x \in \mathbb{R}^n$.

It is known (e.g., see any basic linear algebra text) that:

- (a) A symmetric matrix has real eigenvalues.
- (b) A positive definite matrix has real and positive eigenvalues.
- (c) A nonnegative definite matrix has real and nonnegative eigenvalues.

- (d) To each eigenvalue of a symmetric matrix, we can associate a real eigenvector. Eigenvectors associated with distinct eigenvalues are orthogonal; eigenvectors associated with repeated eigenvalues can always be taken to be orthogonal. Without loss of generality, all these eigenvectors can be normalized so that they have unit length, resulting in an orthonormal basis.
- (e) The above essentially states that a symmetric definite matrix becomes diagonal after a suitable orthogonal change of basis.

A concise summary of the above discussion is the following **spectral decomposition** formula: Every symmetric matrix A can be expressed in the form

$$A = \sum_{i=1}^n \lambda_i \mathbf{z}_i \mathbf{z}_i^T,$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A , and $\mathbf{z}_1, \dots, \mathbf{z}_n$ is an associated collection of orthonormal eigenvectors. (Note here that \mathbf{z}_i^T is an $n \times 1$ matrix, of rank 1.)

For nonnegative definite matrices, we have $\lambda_i \geq 0$, which allows us to take square roots and define

$$B = \sum_{i=1}^n \sqrt{\lambda_i} \mathbf{z}_i \mathbf{z}_i^T.$$

We then observe that:

- (a) The matrix B is symmetric.
- (b) We have $B^2 = A$ (this is an easy calculation). Thus B is a **symmetric square root** of A .
- (c) The matrix B has eigenvalues $\sqrt{\lambda_i}$. Therefore, it is positive (respectively, nonnegative) definite if and only if A is positive (respectively, nonnegative) definite.

Finally, if A is positive definite, then each λ_i is positive, and we can define the matrix

$$C = \sum_{i=1}^n \frac{1}{\lambda_i} \mathbf{z}_i \mathbf{z}_i^T.$$

An easy calculation shows that $A = AC = I$, so that $C = A^{-1}$.

2 DEFINITION OF THE MULTIVARIATE NORMAL DISTRIBUTION

Our interest in positive definite matrices stems from the following. When positive definite, the quadratic form $q(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$ goes to infinity as $\|\mathbf{x}\| \rightarrow \infty$, so that $e^{-q(\mathbf{x})}$ decays to zero, as $\|\mathbf{x}\| \rightarrow \infty$, and therefore can be used to define a multivariate PDF.

There are multiple ways of defining multivariate normal distributions. We will present three, and will eventually show that they are consistent with each other.

The first generalizes our definition of the bivariate normal. It is the most explicit and transparent; on the downside it can lead to unpleasant algebraic manipulations. Recall that $|V|$ stands for the absolute value of the determinant of a square matrix V .

Definition 2. A random vector \mathbf{X} has a **nondegenerate (multivariate) normal distribution** if it has a joint PDF of the form

$$f_X(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp \left\{ -\frac{(\mathbf{x} - \mu)^T V^{-1} (\mathbf{x} - \mu)}{2} \right\},$$

for some real vector μ and some positive definite matrix V .

The second definition is constructive, which makes it operationally useful.

Definition 3. A random vector \mathbf{X} has a **(multivariate) normal distribution** if it can be expressed in the form

$$\mathbf{X} = D\mathbf{W} + \mu,$$

for some matrix D and some real vector μ , where \mathbf{W} is a random vector whose components are independent $N(0, 1)$ random variables.

The last definition is possibly the hardest to penetrate, but in the eyes of some, it is the most elegant.

Definition 4. A random vector \mathbf{X} has a **(multivariate) normal distribution** if for every real vector \mathbf{a} , the random variable $\mathbf{a}^T \mathbf{X}$ is normal.

A brief remark on the use of the word “nondegenerate” in Definition 2. Under Definition 2 $f_X(\mathbf{x}) > 0$ for all \mathbf{x} . On the other hand, consider the following example. Let $X_1 \sim N(0, 1)$ and let $X_2 = 0$. The random vector $\mathbf{X} = (X_1, X_2)$ is normal according to Definitions 3 or 4, but cannot be described by a joint PDF (all of the probability is concentrated on the horizontal axis, a set of zero area). This is an example of a degenerate normal distribution: the distribution is concentrated on a proper subspace \mathbb{R}^1 . The most extreme example is a one-dimensional random variable, which is identically equal to zero. This qualifies as normal under Definitions 3 and 4. One may question the wisdom of calling the number “zero” a “normal random variable;” the reason for doing so is that it allows us to state results such as “a linear function of a normal random variable is normal”, etc., without having to worry about exceptions and special conditions that will prevent degeneracy.

3 MEANS AND COVARIANCES OF VECTOR RANDOM VARIABLES

Let us first introduce a bit more notation. If (X_1, \dots, X_n) is a random vector, we define

$$\mathbf{E}[\mathbf{X}] = (\mathbf{E}[X_1], \dots, \mathbf{E}[X_n]),$$

which we treat as a column vector. Similarly if \mathbf{A} is a random matrix (a matrix with each entry being a random variable A_{ij}), we use the notation $\mathbf{E}[\mathbf{A}]$, to denote the matrix whose entries are $\mathbf{E}[A_{ij}]$.

Given two random vectors $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$, we can consider all the possible covariances

$$\text{Cov}(X_i, Y_j) = \mathbf{E}[(X_i - \mathbf{E}[X_i])(Y_j - \mathbf{E}[Y_j])],$$

and we can arrange them in an $m \times m$ covariance matrix

$$\text{Cov}(\mathbf{X}, \mathbf{Y})$$

whose (i, j) th entry is $\text{Cov}(X_i, Y_j)$. It is easily checked that

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{E}[(\mathbf{X} - \mathbf{E}[\mathbf{X}])(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])^T].$$

Notice also that $\text{Cov}(\mathbf{X}, \mathbf{X})$ is an $n \times n$ symmetric matrix.

Exercise 1. Prove that $\text{Cov}(\mathbf{X}, \mathbf{X})$ is nonnegative definite.

4 KEY PROPERTIES OF THE MULTIVARIATE NORMAL

The theorem below includes almost everything useful there is to know about multivariate normals. We will prove and state the theorem, while working mostly with Definition 3. The proof of equivalence of the three definitions will be completed in the next lecture, together with some additional observations.

Theorem 1. Suppose that $\mathbf{X} = (X_1, \dots, X_n)$ is multivariate normal, in the sense of Definition 3, and let i be the i th component of \mathbf{X} .

- (a) For every i , X_i is normal, with mean μ_i .
- (b) We have $\text{Cov}(\mathbf{X}, \mathbf{X}) = DD^T$.
- (c) If C is an $m \times n$ matrix and d is a vector in \mathbb{R}^m , then $\mathbf{Y} = C\mathbf{X} + d$ is multivariate normal in the sense of Definition 3, with mean $C\mu + d$ and covariance matrix CDD^TC^T .
- (d) If $|D| \neq 0$, then \mathbf{X} is a nondegenerate multivariate normal in the sense of Definition 2, with $V = DD^T = \text{Cov}(\mathbf{X}, \mathbf{X})$.
- (e) The joint CDF F_X of \mathbf{X} is completely determined by the mean and covariance of \mathbf{X} .
- (f) The components of \mathbf{X} are uncorrelated (i.e., the covariance matrix is diagonal) if and only if they are independent.
- (g) If

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix} \right),$$
 and $V_{YY} > 0$, then:
 - (i) $\mathbf{E}[\mathbf{X} | \mathbf{Y}] = \mu_X + V_{XY}V_{YY}^{-1}(\mathbf{Y} - \mu_Y)$.
 - (ii) Let $\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{E}[\mathbf{X} | \mathbf{Y}]$. Then, $\tilde{\mathbf{X}}$ is independent of \mathbf{Y} , and independent of $\mathbf{E}[\mathbf{X} | \mathbf{Y}]$.
 - (iii) $\text{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}} | \mathbf{Y}) = \text{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}) = V_{XX} - V_{XY}V_{YY}^{-1}V_{YX}$.

Proof:

(a) Under definition 3, X_i is a linear function of independent normal random variables, hence normal. Since $\mathbf{E}[\mathbf{W}] = 0$, we have $\mathbf{E}[X_i] = \mu_i$.

(b) For simplicity, let us just consider the zero mean case. We have

$$\text{Cov}(\mathbf{X}, \mathbf{X}) = \mathbf{E}[\mathbf{X}\mathbf{X}^T] = \mathbf{E}[D\mathbf{W}\mathbf{W}^TD^T] = D\mathbf{E}[\mathbf{W}\mathbf{W}^T]D^T = DD^T,$$

where the last equality follows because the components \mathbf{W} are independent (hence the covariance matrix is diagonal), with unit variance (hence the diagonal entries are all equal to 1).

- (c) We have $\mathbf{Y} = C\mathbf{X} + d = C(D\mathbf{W} + \mu) + d$, which is itself a linear function of independent standard normal random variables. This is multivariate normal. The formula $f_{\mathbf{Y}}(\mathbf{y})$ is immediate. The formula for the covariance matrix follows from part (b), \mathbf{V} being replaced by (CD) .
- (d) This is an exercise in derived distributions. Let us again just consider the case of $\mu = 0$. We already know (Lecture 10) that $\mathbf{w} \sim D\mathbf{V}$, with D invertible, then

$$f_{\mathbf{w}}(\mathbf{w}) = \frac{f_{\mathbf{V}}(D^{-1}\mathbf{w})}{|\det D|}.$$

In our case, since the V_i are i.i.d. $N(0,1)$, we have

$$f_{\mathbf{V}}(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^n}} \exp \left\{ -\frac{1}{2} \mathbf{v}^T \mathbf{v} \right\},$$

leading to

$$f_{\mathbf{w}}(\mathbf{w}) = \frac{1}{\sqrt{(2\pi)^n |DD^T|}} \exp \left\{ -\frac{1}{2} \mathbf{w}^T (D^{-1})^T D^{-1} \mathbf{w} \right\}.$$

This is of the form given in Definition 2, with DD^T . In conjunction with part (b), we also have $\text{Cov}(\mathbf{X}, \mathbf{X}) = V$. The argument for the non zero mean case is essentially the same.

- (e) Using part (d), the joint PDF of \mathbf{X} is completely determined by the matrix V , which happens to be equal to $\text{Cov}(\mathbf{X}, \mathbf{X})$, together with the vector μ . The degenerate case is a little harder, because of the absence of a convenient closed form formula. One could think of a limiting argument that involves injecting a tiny bit of noise in all directions, to make the distribution nondegenerate, and then taking the limit. This type of argument can be made to work, but will involve tedious technicalities. Instead, we will take a shortcut, based on the inversion property of transforms. This argument is simpler, but relies on the heavy machinery behind the proof of the inversion property.

Let us find the multivariate transform $M_{\mathbf{X}}(\mathbf{s}) = \mathbf{E}[e^{\mathbf{s}^T \mathbf{X}}]$. We note that $\mathbf{s}^T \mathbf{X}$ is normal with mean $\mathbf{s}^T \mu$. Letting $\tilde{\mathbf{X}} = \mathbf{X} - \mu$, the variance of $\tilde{\mathbf{X}}$

is

$$\text{var}(\mathbf{s}^T \mathbf{X}) = \mathbf{E}[\mathbf{s}^T \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T \mathbf{s}] = \mathbf{s}^T \mathbf{E}[\tilde{\mathbf{X}} \tilde{\mathbf{X}}^T] \mathbf{s} = \mathbf{s}^T \text{Cov}(\mathbf{X}, \mathbf{X}) \mathbf{s} = \mathbf{s}^T V \mathbf{s}.$$

Using the formula for the transform of a single normal random variable ($\mathbf{s}^T \mathbf{X}$ in this case), we have

$$M_{\mathbf{X}}(\mathbf{s}) = \mathbf{E}[e^{\mathbf{s}^T \mathbf{X}}] = M_{\mathbf{s}^T \mathbf{X}}(1) = e^{\mathbf{s}^T \mu} e^{\mathbf{s}^T V \mathbf{s}/2}.$$

Thus, μ and V completely determine the transform \mathbf{oX} . By the inversion property of transforms, μ and V completely determine the distribution (e.g., the CDF) oX .

- (f) If the components oX are independent they are of course uncorrelated. For the converse, suppose that the components oX are uncorrelated, i.e., the matrix V is a diagonal. Consider another random vector \mathbf{Y} that has the same mean and oX , whose components are independent normal, and such that the variance df_i is the same as the variance Xf_i . Then, \mathbf{X} and \mathbf{Y} have the same mean and covariance. By part (d), \mathbf{X} and \mathbf{Y} have the same distribution. Since the components Yf_i are independent, it follows that the components oX are also independent.

For the special case when V is invertible, we could alternatively use part (d) which provides an explicit formula for the joint PDF of \mathbf{X} and \mathbf{Y} . When V is diagonal we see that the joint PDF is the product of its marginal PDFs. Namely $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}_1}(x_1) \cdots f_{\mathbf{X}_n}(x_n)$ and thus the components oX are independent.

- (g) Once more, to simplify notation, let us just deal with the zero-mean case. Let us define

$$\hat{\mathbf{X}} = V_{XY} V_{YY}^{-1} \mathbf{Y}.$$

We then have

$$\mathbf{E}[\hat{\mathbf{X}} \mathbf{Y}^T] = V_{XY} V_{YY}^{-1} \mathbf{E}[\mathbf{Y} \mathbf{Y}^T] = V_{XY} = \mathbf{E}[\mathbf{X} \mathbf{Y}^T].$$

This proves that $\mathbf{X} - \hat{\mathbf{X}}$ is uncorrelated with \mathbf{Y} . Note that $(\mathbf{X} - \hat{\mathbf{X}}, \mathbf{Y})$ is a linear function of (\mathbf{X}, \mathbf{Y}) , so, by part (c), it is also multivariate normal. Using an argument similar to the one in the proof of part (f), we conclude that $\mathbf{X} - \hat{\mathbf{X}}$ is independent of \mathbf{Y} , and therefore independent from any function of \mathbf{Y} . Recall now the abstract definition of conditional expectations. The relation $\mathbf{E}[(\mathbf{X} - \hat{\mathbf{X}})g(\mathbf{Y})] = 0$, for every function g , implies that $\hat{\mathbf{X}} = \mathbf{E}[\mathbf{X} | \mathbf{Y}]$, which proves part (i).

For part (ii), note that we already proved $\tilde{\mathbf{X}} = \mathbf{X} - \hat{\mathbf{X}} = \mathbf{X} - \mathbf{E}[\mathbf{X} | \mathbf{Y}]$ is independent of \mathbf{Y} . Since $\mathbf{E}[\mathbf{X} | \mathbf{Y}]$ is a function of \mathbf{Y} , it follows that $\tilde{\mathbf{X}}$ is independent of $\mathbf{E}[\mathbf{X} | \mathbf{Y}]$.

For part (iii), note that $\tilde{\mathbf{X}}$ is independent of \mathbf{Y} , which implies that $\text{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}} | \mathbf{Y}) = \text{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}})$. Finally,

$$\begin{aligned}\text{Cov}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}) &= \mathbf{E}[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T] = \mathbf{E}[(\mathbf{X} - \hat{\mathbf{X}})(\mathbf{X} - \hat{\mathbf{X}})^T] = \mathbf{E}[(\mathbf{X} - \hat{\mathbf{X}})\mathbf{X}^T] \\ &= V_{XX} - \mathbf{E}[V_{XY}V_{YY}^{-1}\mathbf{Y}\mathbf{X}^T] = V_{XX} - V_{XY}V_{YY}^{-1}\mathbf{E}[\mathbf{Y}\mathbf{X}^T] \\ &= V_{XX} - V_{XY}V_{YY}^{-1}V_{YX}.\end{aligned}$$

□

Note that in the case of the bivariate normal, we have $\text{Cov}(X, Y) = \rho\sigma_X\sigma_Y$, $V_{XX} = \sigma_X^2$, $V_{YY} = \sigma_Y^2$. Then, part (g) of the preceding theorem, for the zero-mean case, reduces to

$$\mathbf{E}[X | Y] = \rho \frac{\sigma_X}{\sigma_Y} Y, \quad \text{var}(\tilde{X}) = \sigma_X^2(1 - \rho^2),$$

which agrees with the formula we derived through elementary means in Lecture 9, for the special case of unit variances.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Fall 2018

Lecture 15

MULTIVARIATE NORMAL DISTRIBUTIONS (CTD.); CHARACTERISTIC FUNCTIONS

Contents

1. Equivalence of the three definitions of the multivariate normal
2. Proof of equivalence
3. Whitening of a sequence of normal random variables
4. Characteristic functions

1 EQUIVALENCE OF THE THREE DEFINITIONS OF THE MULTIVARIATE NORMAL DISTRIBUTION

1.1 The definitions

Recall the following three definitions from the previous lecture.

Definition 1. A random vector \mathbf{X} has a **nondegenerate (multivariate) normal distribution** if it has a joint PDF of the form

$$f_X(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |V|}} \exp \left\{ -\frac{(\mathbf{x} - \boldsymbol{\mu}) V^{-1} (\mathbf{x} - \boldsymbol{\mu})^T}{2} \right\},$$

for some real vector $\boldsymbol{\mu}$ and some positive definite matrix V .

Definition 2. A random vector \mathbf{X} has a **(multivariate) normal distribution** if it can be expressed in the form

$$\mathbf{X} = D\mathbf{W} + \boldsymbol{\mu},$$

for some matrix D and some real vector $\boldsymbol{\mu}$, where \mathbf{W} is a random vector whose components are independent $N(0, 1)$ random variables.

Definition 3. A random vector \mathbf{X} has a **(multivariate) normal distribution** if for every real vector \mathbf{a} , the random variable $\mathbf{a}^T \mathbf{X}$ is normal.

2 PROOF OF EQUIVALENCE

In the course of the proof of Theorem 1 in the previous lecture, we argued that if \mathbf{X} is multivariate normal, in the sense of Definition 2, then:

- (a) It also satisfies Definition 3: $\mathbf{X} = D\mathbf{W} + \boldsymbol{\mu}$, where the W_i are independent, then $\mathbf{a}^T \mathbf{X}$ is a linear function of independent normals, hence normal.
- (b) As long as the matrix D is nonsingular (equivalently, if $\text{Cov}(\mathbf{X}, \mathbf{X}) = DD^T$ is nonsingular), \mathbf{X} also satisfies Definition 1. (We used the derived distributions formula.)

We complete the proof of equivalence by establishing converses of the above two statements.

Theorem 1.

- (a) If \mathbf{X} satisfies Definition 1, then it also satisfies Definition 2.
- (b) If \mathbf{X} satisfies Definition 3, then it also satisfies Definition 2.

Proof:

- (a) Suppose that \mathbf{X} satisfies Definition 1, so in particular, the matrix V is positive definite. Let D be a symmetric matrix such that $D^2 = V$. Since

$$(\det(D))^2 = \det(D^2) = \det(V) > 0,$$

we see that D is nonsingular, and therefore invertible. Let

$$\mathbf{W} = D^{-1}(\mathbf{X} - \mu).$$

Note that $\mathbf{E}[\mathbf{W}] = 0$. Furthermore,

$$\begin{aligned}\text{Cov}(\mathbf{W}, \mathbf{W}) &= \mathbf{E}[D^{-1}(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T D^{-1}] \\ &= D^{-1}\mathbf{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T]D^{-1} \\ &= D^{-1}V D^{-1} = I.\end{aligned}$$

We have shown thus far that \mathbf{W} are normal and uncorrelated. We now proceed to show that they are independent. Using the formula for the PDF of \mathbf{X} and the change of variables formula, we find that the PDF of \mathbf{W} is of the form

$$c \cdot \exp\{-\mathbf{w}^T \mathbf{w}/2\} = c \cdot \exp\{(w_1^2 + \dots + w_n^2)/2\},$$

for some normalizing constant, which is the joint PDF of a vector of independent normal random variables. It follows that $\mathbf{X} = D\mathbf{W} + \mu$ is a multivariate normal in the sense of Definition 2.

- (b) Suppose that \mathbf{X} satisfies Definition 3, i.e., any linear function \mathbf{X} is normal. Let $V = \text{Cov}(\mathbf{X}, \mathbf{X})$, and let D be a symmetric matrix such that $D^2 = V$. We first give the proof for the easier case where D is invertible.

Let $\mathbf{W} = D^{-1}(\mathbf{X} - \mu)$. As before $\mathbf{E}[\mathbf{W}] = 0$, and $\text{Cov}(\mathbf{W}, \mathbf{W}) = I$. Fix a vectors, Then, $\mathbf{s}^T \mathbf{W}$ is a linear function of \mathbf{W} , and is therefore normal. Note that

$$\text{var}(\mathbf{s}^T \mathbf{W}) = \mathbf{E}[\mathbf{s}^T \mathbf{W} \mathbf{W}^T \mathbf{s}] = \mathbf{s}^T \text{Cov}(\mathbf{W}, \mathbf{W}) \mathbf{s} = \mathbf{s}^T \mathbf{s}.$$

Since $\mathbf{s}^T \mathbf{W}$ is a scalar, zero mean, normal random variable, we know that

$$M_{\mathbf{W}}(\mathbf{s}) = \mathbf{E}[\exp\{\mathbf{s}^T \mathbf{W}\}] = M_{\mathbf{s}^T \mathbf{W}}(1) = \exp\{\text{var}(\mathbf{s}^T \mathbf{W})/2\} = \exp\{\mathbf{s}^T \mathbf{s}/2\}.$$

We recognize that this is the transform associated with a vector of independent standard normal random variables. By the inversion property of transforms, it follows that \mathbf{W} is a vector of independent standard normal random variables. Therefore, $\mathbf{X} = D\mathbf{W} + \mu$ is multivariate normal in the sense of Definition 2.

(b)' Suppose now that V is singular (as opposed to positive definite). For simplicity, we will assume that the mean \mathbf{x}_0 is zero. Then, there exists some $\mathbf{a} \neq 0$, such that $V\mathbf{a} = 0$, and $\mathbf{a}^T V \mathbf{a} = 0$. Note that

$$\mathbf{a}^T V \mathbf{a} = \mathbf{E}[(\mathbf{a}^T \mathbf{X})^2].$$

This implies that $\mathbf{a}^T \mathbf{X} = 0$, with probability 1. Consequently, some component of \mathbf{X} is a deterministic linear function of the remaining components.

By possibly rearranging the components of \mathbf{X} , let us assume that \mathbf{x}_n is a linear function of $(\mathbf{x}_1, \dots, \mathbf{x}_{n-1})$. If the covariance matrix $(\mathbf{x}_1, \dots, \mathbf{x}_{n-1})$ is also singular, we repeat the same argument, until eventually a nonsingular covariance matrix is obtained. At that point we have reached the situation where \mathbf{X} is partitioned as $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$, with $\text{Cov}(\mathbf{Y}, \mathbf{Y}) > 0$, and with \mathbf{Z} a linear function of \mathbf{Y} (i.e., $\mathbf{Z} = A\mathbf{Y}$, for some matrix A , with probability 1).

The vector \mathbf{Y} also satisfies Definition 3. Since its covariance matrix is non singular, the previous part of the proof shows that it also satisfies Definition 2. Let k be the dimension of \mathbf{Y} . Then, $\mathbf{Y} = D\mathbf{W}$, where \mathbf{W} consists of k independent standard normals, and D is a $k \times k$ matrix. Let $\overline{\mathbf{W}}$ be a vector of $n - k$ independent standard normals. Then, we can write

$$\mathbf{X} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} = \begin{bmatrix} D & 0 \\ AD & 0 \end{bmatrix} \begin{bmatrix} \mathbf{W} \\ \overline{\mathbf{W}} \end{bmatrix},$$

which shows that \mathbf{X} satisfies Definition 2.

We should also consider the extreme possibility that in the process of eliminating components of \mathbf{X} , a nonsingular covariance matrix is never obtained. But in that case, we have $\mathbf{X} = 0$, which also satisfies Definition 2, with $D = 0$. (This is the most degenerate case of a multivariate normal.)

□

3 WHITENING OF A SEQUENCE OF NORMAL RANDOM VARIABLES

The last part of the proof in the previous section provides some interesting intuition. Given a multivariate normal vector \mathbf{X} , we can always perform a change of coordinates, and obtain a representation of that vector in terms of independent normal random variables. Our process of going from \mathbf{X} to \mathbf{W} involved factoring the covariance matrix V of \mathbf{X} in the form $V = D^2$, where D was a symmetric square root of V . However, other factorizations are also possible. The most useful one is described below.

Let

$$\begin{aligned} W_1 &= X_1, \\ W_2 &= X_2 - \mathbf{E}[X_2 | X_1], \\ W_3 &= X_3 - \mathbf{E}[X_3 | X_1, X_2], \\ &\vdots \quad \vdots \\ W_n &= X_n - \mathbf{E}[X_n | X_1, \dots, X_{n-1}]. \end{aligned}$$

- (a) Each W_i can be interpreted as the new information provided by given the past, X_1, \dots, X_{i-1} . The W_i are sometimes called **innovations**.
- (b) When we deal with multivariate normals, conditional expectations are linear functions of the conditioning variables. Thus, W are linear functions of the X_i . Furthermore, we have $\mathbf{W} = L\mathbf{X}$, where L is a lower triangular matrix (all entries above the diagonal are zero). The diagonal entries are all equal to 1, so L is invertible. The inverse of L is also lower triangular. This means that the transformation from \mathbf{X} to \mathbf{W} is **causal** (W_i can be determined from X_1, \dots, X_i) and **causally invertible** (X_i can be determined from W_1, \dots, W_i). Engineers sometimes call this a “causal and causally invertible whitening filter.”
- (c) The W_i are independent of each other. This is a consequence of the general fact $\mathbf{E}[(X - \mathbf{E}[X | Y])Y] = 0$, which shows that W_i is uncorrelated with X_1, \dots, X_{i-1} , hence uncorrelated with W_1, \dots, W_{i-1} . For multivariate normals, we know that zero correlation implies independence. As long as the W_i have nonzero variance, we can also normalize them so that their variance is equal to 1.
- (d) The covariance matrix of \mathbf{W} , call it B , is diagonal. An easy calculation shows that $\text{Cov}(X, X) = L^{-1}B(L^{-1})^T$. This kind of factorization into a product of a lower triangular $L^{-1}B^{1/2}$ and upper triangular $B^{1/2}(L^{-1})^T$ matrix is called **Cholesky factorization**.

4 INTRODUCTION TO CHARACTERISTIC FUNCTIONS

We have defined the moment generating function $M_X(s)$, for real values of s , and noted that it may be infinite for some values of s . In particular, if $M_X(s) = \infty$ for every $s \neq 0$, then the moment generating function does not provide enough information to determine the distribution of X . (As an example,

consider a PDF of the form $f_X(x) = c/(1 + x^2)$, where c is a suitable normalizing constant.) A way out of this difficulty is to consider complex values of s , and in particular, the case where a purely imaginary number $s = it$, where $i = \sqrt{-1}$, and $t \in \mathbb{R}$. The resulting function is called the **characteristic function**, formally defined by

$$\phi_X(t) = \mathbf{E}[e^{itX}].$$

For example, when X is a continuous random variable with PDF we have

$$\phi_X(t) = \int e^{ixt} f(x) dx,$$

which is very similar to the Fourier transform of f (except for the absence of a minus sign in the exponent). Thus, the relation between moment generating functions and characteristic functions is of the same kind as the relation between Laplace and Fourier transforms.

Note that e^{itX} is a **complex-valued** random variable, a new concept for us. However, using the relation $e^{itX} = \cos(tX) + i \sin(tX)$, defining its expectation is straightforward:

$$\phi_X(t) = \mathbf{E}[\cos(tX)] + i\mathbf{E}[\sin(tX)].$$

We make a few key observations:

- (a) Because $|e^{itX}| \leq 1$ for every t , its expectation $\phi_X(t)$ is well-defined and finite for every $t \in \mathbb{R}$. In fact, $|\phi_X(t)| \leq 1$, for every t .
- (b) The key properties of moment generating functions (cf. Lecture 14) are also valid for characteristic functions (same proof).

Theorem 2.

- (a) If $Y = aX + b$, then $\phi_Y(t) = e^{itb} \phi_X(at)$.
- (b) If X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.
- (c) Let X and Y be independent random variables. Let Z be equal to X , with probability p , and equal to Y , with probability $1 - p$. Then,

$$\phi_Z(t) = p\phi_X(t) + (1 - p)\phi_Y(t).$$

- (c) **Inversion theorem:** If two random variables have the same characteristic function, then their distributions are the same. We prove this result below.

- (d) The above inversion theorem remains valid for multivariate characteristic functions, defined by $\phi_{\mathbf{X}}(\mathbf{t}) = \mathbf{E}[e^{it^T \mathbf{X}}]$.
- (e) For the univariate case, X is a continuous random variable with PDF, there is an explicit inversion formula, namely

$$f_X(x) = \frac{1}{2\pi} \lim_{T \rightarrow \infty} \int_{-T}^T e^{-itx} \phi_X(t) dt,$$

for every x at which f_X is differentiable. (Note the similarity with inversion formulas for Fourier transforms.)

- (f) The dominated convergence theorem can be applied to complex random variables (simply apply the DCT separately to the complex and imaginary parts). Thus, if $\lim_{n \rightarrow \infty} X_n = X$, a.s., then, for every $t \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \phi_{X_n}(t) = \lim_{n \rightarrow \infty} \mathbf{E}[e^{itX_n}] = \mathbf{E}\left[\lim_{n \rightarrow \infty} e^{itX_n}\right] = \mathbf{E}[e^{itX}] = \phi_X(t).$$

The DCT applies here, because the random variables $|e^{itX_n}|$ are bounded by 1.

- (g) If $\mathbf{E}[|X|^k] < \infty$, then $\phi_X(t)$ is k -times continuously differentiable and also

$$\frac{d^k}{dt^k} \phi_X(t) \Big|_{t=0} = i^k \mathbf{E}[X^k].$$

(This is plausible, by moving the differentiation inside the expectation, but a formal justification is needed.)

- (h) If $\mathbf{E}[e^{\epsilon|X|}] < \infty$ for some $\epsilon > 0$ (equivalently if MGF of X exists in a neighborhood of zero) then $\phi_X(t)$ is analytic function of t , which extends to all complex z inside a strip $\{z : -\epsilon < \text{Im } z < \epsilon\}$.

Two useful characteristic functions:

- (a) **Exponential:** If $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$, then

$$\phi_X(y) = \frac{\lambda}{\lambda - iy}.$$

Note that this is the same as starting $M_X(s) = \lambda / (\lambda - s)$ and replacing s by iy ; however, this is not a valid proof. One must either use tools from complex analysis (contour integration), or evaluate separately $\mathbf{E}[\cos(tX)]$, $\mathbf{E}[\sin(tX)]$, which can be done using integration by parts.

- (b) **Normal (scalar):** If $X \stackrel{d}{=} N(\mu, \sigma^2)$, then

$$\phi_X(t) = e^{it\mu} e^{-t^2 \sigma^2 / 2}.$$

4.1 Inversion theorem

Theorem 3(Inversion theorem) *Let X and Y have the same characteristic functions. Then $\mathbb{P}_X = \mathbb{P}_Y$.*

Proof. Let $a > 1$ and consider the following “trapezoidal function”

$$f_a(x) = \begin{cases} 0, & |x| \geq a \\ \frac{1}{a-1}(x+a), & -a < x < -1 \\ 1, & -1 \leq x \leq 1 \\ -\frac{1}{a-1}(x-a), & 1 < x < a \end{cases}$$

Note that

$$\lim_{a \rightarrow 1^+} f_a(x) = 1_{[-1,1]}(x) \quad (1)$$

Furthermore, there is an identity

$$f_a(x) = \frac{4}{(a-1)\sqrt{2\pi}} \int_{\mathbb{R}} e^{-itx} \frac{1}{t^2} \left[\frac{1}{a} \sin^2 \frac{ta}{2} - \sin^2 \frac{t}{2} \right] dt \quad (2)$$

To show this you may either compute the integral directly or use Fourier inversion and the observation that $f_a = \frac{1}{a-1}(g * g - h * h)$, where $g = 1_{[-a/2, a/2]}$, $h = 1_{[-1,1]}$ and $*$ is convolution.

Note that the integral in (2) is absolutely convergent since the absolute value of the integrand

$$\frac{1}{t^2} \left| \frac{1}{a} \sin^2 \frac{ta}{2} - \sin^2 \frac{t}{2} \right|$$

is continuous at 0 and integrable \Rightarrow . Thus, by Fubini we have

$$\mathbb{E}[f_a(X)] = \frac{4}{(a-1)\sqrt{2\pi}} \int_{\mathbb{R}} \phi_X(-t) \frac{1}{t^2} \left[\frac{1}{a} \sin^2 \frac{ta}{2} - \sin^2 \frac{t}{2} \right] dt$$

Since $\phi_X = \phi_Y$ we have

$$\mathbb{E}[f_a(X)] = \mathbb{E}[f_a(Y)]$$

for every $a > 1$. Taking limit as $a \searrow 1$ and applying the BCT to (1) we get

$$\mathbb{P}_X([-1, 1]) = \mathbb{P}_Y([-1, 1])$$

A similar argument (with shifted and scaled) shows that \mathbb{P}_X and \mathbb{P}_Y coincide on every closed interval. Since the collection of closed intervals is a generating p -system, we have $\mathbb{P}_X = \mathbb{P}_Y$. \square

4.2 Vector-valued random variables

A very useful extension is to define characteristic function for vector-valued random variable $\mathbf{X} = (X_1, \dots, X_d)^T \in \mathbb{R}^d$. In this case characteristic function is defined on \mathbb{R}^d as follows:

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E} \left[e^{i\mathbf{t}^T \mathbf{X}} \right], \quad \mathbf{t} = (t_1, \dots, t_d)^T \in \mathbb{R}^d$$

where $\mathbf{t}^T \mathbf{X} = \sum_{j=1}^d t_j X_j$ denotes a standard scalar product \mathbb{R}^d .

Most of the properties and results above (including inversion theorem) carry over to the vector case. This leads to numerous useful implications, of which we discuss two:

1. Checking independence: If $\mathbf{X} = (X_1, \dots, X_d)^T$, then X_j are independent if and only if

$$\phi_{\mathbf{X}}(\mathbf{t}) = \prod_{j=1}^d \phi_{X_j}(t_j) \quad (3)$$

This easily follows from the inversion theorem, since right-hand side represents the characteristic function of distribution $\prod_{j=1}^d \mathbb{P}_{X_j}$.

2. Fourth definition of multivariate normal. It is not hard to show that for (degenerate or non-degenerate) multivariate normal \mathbf{X} we have

$$\phi_{\mathbf{X}}(\mathbf{t}) = e^{i\mu^T \mathbf{t} - \frac{1}{2}\mathbf{t}^T V \mathbf{t}} \quad (4)$$

where $\mu = \mathbf{E}[\mathbf{X}]$ and $V = \text{Cov}(\mathbf{X}, \mathbf{X})$. Since ϕ uniquely determines the distribution, property (4) is frequently taken as *definition* of a multivariate normal. Most properties then follow immediately. For example, “uncorrelated implies independent” is just a consequence of (3).

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Fall 2018

Lecture 16

CONVERGENCE OF RANDOM VARIABLES

Contents

1. Definitions
2. Convergence in distribution
3. The hierarchy of convergence concepts

1 DEFINITIONS

1.1 Almost sure convergence

Definition 1. We say that X_n converges to X **almost surely** (a.s.), and write $X_n \xrightarrow{\text{a.s.}} X$, if there is a (measurable) $A \subset \Omega$ such that:

- (a) $\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)$, for all $\omega \in A$;
- (b) $\mathbb{P}(A) = 1$.

Note that for a.s. convergence to be relevant, all random variables need to be defined on the same probability space (one experiment). Furthermore, the different random variables X_n are generally highly dependent.

Two common cases where a.s. convergence arises are the following.

- (a) The probabilistic experiment runs over time. To each *time* associate a nonnegative random variable Z_n (e.g., income on day n). Let $X_n = \sum_{k=1}^n Z_k$ be the income on the first n days. Let $X = \sum_{k=1}^{\infty} Z_k$ be the lifetime income. Note that X is well defined (as an extended real number) for every $\omega \in \Omega$, because of our assumption that $Z_k \geq 0$, and $X_n \xrightarrow{\text{a.s.}} X$.

- (b) The various random variables are defined as different functions of a single underlying random variable. More precisely, suppose that a random variable, and let $g_n : \mathbb{R} \rightarrow \mathbb{R}$ be measurable functions. Let $X_n = g_n(Y)$ [which really means, $X_n(\omega) = g_n(Y(\omega))$, for all ω]. Suppose that $\lim_{n \rightarrow \infty} g_n(y) = g(y)$ for every y . Then, $X_n \xrightarrow{\text{a.s.}} X$. For example, let $g_n(y) = y + y^2/n$, which converges to $g(y) = y$. We then have $Y + Y^2/n \xrightarrow{\text{a.s.}} Y$.

When $X_n \xrightarrow{\text{a.s.}} X$, we always have

$$\phi_{X_n}(t) \rightarrow \phi_X(t), \quad \forall t,$$

by the dominated convergence theorem. On the other hand, the relation

$$\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$$

is not always true; sufficient conditions are provided by the monotone and dominated convergence theorems. For an example, where convergence of expectations fails to hold, consider a random variable which is uniform on $[0, 1]$, and let:

$$X_n = \begin{cases} n, & \text{if } U \leq 1/n, \\ 0, & \text{if } U > 1/n. \end{cases} \quad (1)$$

We have

$$\lim_{n \rightarrow \infty} \mathbb{E}[X_n] = \lim_{n \rightarrow \infty} \mathbb{P}(U \leq 1/n) = 1.$$

On the other hand, for any outcome ω for which $U(\omega) > 0$ (which happens with probability one), $X_n(\omega)$ converges to zero. Thus $X_n \xrightarrow{\text{a.s.}} 0$, but $\mathbb{E}[X_n]$ does not converge to zero.

1.2 Convergence in distribution

Definition 2. Let X and X_n , $n \in \mathbb{N}$, be random variables with CDFs F and F_n , respectively. We say that the sequence X_n converges to X in distribution, and write $X_n \xrightarrow{d} X$, if

$$\lim_{n \rightarrow \infty} F_n(x) = F(x),$$

for every $x \in \mathbb{R}$ at which F is continuous.

- (a) Recall that CDFs have discontinuities (“jumps”) only at the points that have positive probability mass. More precisely, F is continuous at x if and only if $\mathbb{P}(X = x) = 0$.
- (b) Let $X_n = 1/n$, and $X = 0$, with probability 1. Note that $F_{X_n}(0) = \mathbb{P}(X_n \leq 0) = 0$ for every n , but $F_X(0) = 1$. Still, because of the exception in the above definition, we have $X_n \xrightarrow{d} X$. More generally, if $X_n = a_n$ and $X = a$, with probability 1, and $a_n \rightarrow a$, then $X_n \xrightarrow{d} X$. Thus, convergence in distribution is consistent with the definition of convergence of real numbers. This would not have been the case if the definition required the condition $\lim_{n \rightarrow \infty} F_n(x) = F(x)$ to hold at every x .
- (c) Note that this definition just involves the marginal distributions of the random variables involved. These random variables may even be defined on different probability spaces.
- (d) Let Y be a random variable whose PDF is symmetric around 0. Namely, for every real value t , $\mathbb{P}(Y \leq t) = \mathbb{P}(Y \geq -t)$. Let $X_n = (-1)^n Y$. Then, every X_n has the same distribution, so, trivially, X_n converges to Y in distribution. However, for almost all the sequence $X_n(\omega)$ does not converge.
- (e) If we are dealing with random variables whose distribution is in a parametric class, (e.g., if every X_n is exponential with parameter λ_n), and the parameters converge (e.g., if $\lambda_n \rightarrow \lambda > 0$ and X is exponential with parameter λ), then we usually have convergence X_n to X , in distribution. Check this for the case of exponential distributions.
- (f) It is possible for a sequence of discrete random variables to converge in distribution to a continuous one. For example, if Y_n is uniform on $\{1, \dots, n\}$ and $X_n = Y_n/n$, then X_n converges in distribution to a random variable which is uniform on $[0, 1]$ (exercise).
- (g) Similarly, it is possible for a sequence of continuous random variables to converge in distribution to a discrete one. For example, if X_n is uniform on $[0, 1/n]$, then X_n converges in distribution to a discrete random variable which is identically equal to zero (exercise).
- (h) If X and all X_n are continuous, convergence in distribution does not imply convergence of the corresponding PDFs. (Exercise. Find an example, by emulating the example in (f).)
- (i) If X and all X_n are integer-valued, convergence in distribution turns out to be equivalent to convergence of the corresponding PMFs: $p_{X_n}(k) \rightarrow p_X(k)$, for all k . (exercise).

1.3 Convergence in probability

Definition 3. (a) We say that a sequence of random variables (not necessarily defined on the same probability space) converges in probability to a real number c , and write $X_n \xrightarrow{\text{i.p.}} c$, if

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - c| \geq \epsilon) = 0, \quad \forall \epsilon > 0.$$

(b) Suppose that X and X_n , $n \in \mathbb{N}$ are all defined on the same probability space. We say that the sequence X_n converges to X , in probability, and write $X_n \xrightarrow{\text{i.p.}} X$, if $X_n - X$ converges to zero, in probability, i.e.,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0, \quad \forall \epsilon > 0.$$

- (a) When X in part (b) of the definition is deterministic, say equal to some constant c , then the two parts of the above definition are consistent with each other.
- (b) As we will see below convergence $X_n \xrightarrow{\text{i.p.}} c$ is equivalent to $X_n \xrightarrow{\text{d.}} c$.
- (c) The intuitive content of the statement $X_n \xrightarrow{\text{i.p.}} c$ is that in the limit n increases, almost all of the probability mass becomes concentrated in a small interval around c , no matter how small this interval is. On the other hand, for any fixed n , there can be a small probability mass outside this interval, with a slowly decaying tail. Such a tail can have a strong impact on expected values. For this reason, convergence in probability does not have any implications on expected values. See for instance the example in Eq. (1). We have $X_n \xrightarrow{\text{i.p.}} X$, but $\mathbb{E}[X_n]$ does not converge to $\mathbb{E}[X]$.
- (d) If $X_n \xrightarrow{\text{i.p.}} X$ and $Y_n \xrightarrow{\text{i.p.}} Y$, and all random variables are defined on the same probability space, then $(X_n + Y_n) \xrightarrow{\text{i.p.}} (X + Y)$ (exercise).

The following is a convenient characterization, showing that convergence in probability is very closely related to almost sure convergence.

Proposition 1. $X_n \xrightarrow{\text{i.p.}} X$ iff for every subsequence X_{n_k} there exists a subsubsequence $X_{n_{k_s}} \xrightarrow{\text{a.s.}} X$.

2 CONVERGENCE IN DISTRIBUTION

The following result provides insights into the meaning of convergence -in distribution.

Recall that the boundary ∂E of a set E is a set of simultaneous limit points of E and E^c : $\partial E \triangleq [E] \cap [E^c]$, where $[.]$ denotes the closure. Also recall that quantile function q of the CDF F is a right-continuous inverse of the CDF:

$$q(s) \triangleq \inf\{x : F(x) > s\}$$

Theorem 1. Let X_n and X be random variables, \mathbb{P}_n and \mathbb{P} their distributions and q_n, q their quantile functions. The following are equivalent:

- (i) $X_n \xrightarrow{d} X$
- (ii) Quantile functions $q_n(u) \rightarrow q(u)$ for every continuity point u of q .
- (iii) $\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)]$ for every bounded continuous f .
- (iv) $\mathbb{P}_n[E] \rightarrow \mathbb{P}[E]$ for every Borel E with $\mathbb{P}[\partial E] = 0$
- (v) $\lim \sup_{n \rightarrow \infty} \mathbb{P}_n[F] \leq \mathbb{P}[F]$ for every closed F
- (vi) $\lim \inf_{n \rightarrow \infty} \mathbb{P}_n[U] \geq \mathbb{P}[U]$ for every open U .

Note: The last four statements remain equivalent for a general metric space, in which case any of them is usually taken as the definition of weak convergence of measures.

Proof. Equivalence of the first two follows by definition of quantiles. Indeed, in the case where F is continuous and strictly monotonically increasing this is clear. The general case follows from carefully analyzing the inclusions:

$$\{(x, y) : y < F(x)\} \subseteq \{(x, y) : q(y) \leq x\} \subseteq \{(x, y) : y \leq F(x)\}$$

valid for any pair of a CDF and its quantile.

Next, (ii) implies (iii), (v) and (vi) by the Theorem to follow next (Skorokhod representation), since $Y_n \xrightarrow{\text{a.s.}} Y$ implies $f(Y_n) \xrightarrow{\text{a.s.}} f(Y)$ for continuous functions and $\lim \sup_{n \rightarrow \infty} 1_F(Y_n) \not\downarrow 1_F(Y)$. The statement of (iii) and (v) then follows by the BCT and Fatou's lemma respectively.

Furthermore, (v) and (vi) are equivalent by taking complements. To show (v) and (vi) imply (iv) let $E = [E]$ and $U = \text{int}E = [E^c]^c$. Then $\partial E = F \setminus U$.

Then since $U \subseteq E \subseteq F$ we have

$$\mathbb{P}[U] \leq \liminf_{n \rightarrow \infty} \mathbb{P}_n[U] \leq \liminf_{n \rightarrow \infty} \mathbb{P}_n[E] \quad (2)$$

$$\leq \limsup_{n \rightarrow \infty} \mathbb{P}_n[E] \leq \limsup_{n \rightarrow \infty} \mathbb{P}_n[F] \leq \mathbb{P}[F] \quad (3)$$

Thus when $\mathbb{P}[\partial E] = 0$ we have $\mathbb{P}[F] = \mathbb{P}[U]$ and thus $\mathbb{P}_n[E] \rightarrow \mathbb{P}[E]$.

On the other hand, (iv) implies (i) by taking $(-\infty, x]$ for any x - point of continuity of F . So overall we have shown:

$$(i) \iff (ii) \Rightarrow (v) \iff (vi) \Rightarrow (iv) \Rightarrow (i) \Rightarrow (iii)$$

It only remains to show that (iii) implies any of the other ones. For example, we can show (iii) \Rightarrow (v). To that end take

$$f_\epsilon(x) = 1 - \frac{1}{\epsilon} \min(d(x, F), \epsilon),$$

where $d(x, F) = \inf_{y \in F} |x - y|$ is the minimum distance between x and F . It is easy to see that $d(x, F)$ is a continuous function of x which is equal to zero only on F itself. Furthermore, $f_\epsilon \searrow 1_F$ as $\epsilon \rightarrow 0$. So we have:

$$\inf_{\epsilon > 0} f_\epsilon(X_n) = 1_F(X_n) \quad (4)$$

and by the BCT

$$\inf_{\epsilon > 0} \mathbb{E}[f_\epsilon(X_n)] = \mathbb{P}_n[F] \quad (5)$$

From here consider the following:

$$\limsup_{n \rightarrow \infty} \mathbb{P}_n[F] \leq \inf_{\epsilon > 0} \limsup_{n \rightarrow \infty} \mathbb{E}[f_\epsilon(X_n)] \quad (6)$$

$$= \inf_{\epsilon > 0} \mathbb{E}[f_\epsilon(X)] \quad (7)$$

$$= \mathbb{P}[F] \quad (8)$$

where (6) follows from (5) by taking the limsup and using the usual inequality $\limsup \inf \leq \inf \limsup$. (7) is by the assumption (iii) and (8) by (4)-(5) applied to \mathbb{P} instead of \mathbb{P}_n . \square

The following result shows a close relation with almost sure convergence.

Theorem 2 (Skorokhod representation) Suppose that $X_n \xrightarrow{d} X$. Then, there exists a probability space and random variables Y_n defined on that space with the following properties:

- (a) For every n , the random variables X_n and Y_n have the same CDF; similarly, X and Y have the same CDF.
- (b) $Y_n \xrightarrow{\text{a.s.}} Y$.

For convergence in distribution, it makes no difference whether the random variables X_n are independent or not; they do not even need to be defined on the same probability space. On the other hand, almost sure convergence implies a strong form of dependence between the random variables involved. The idea in the preceding theorem is to preserve the marginal distributions, but introduce a particular form of dependence between the X_n , which then results in almost sure convergence. This dependence is introduced by generating random variables and Y with the desired distributions, using a common random number generator, e.g., a single random variable, uniformly distributed on $(0, 1)$.

Proof. Recall that if q_n is the quantile function of X_n then $q_n(U) \sim X_n$, where U is uniform on $(0, 1)$. Take $Y_n = q_n(U)$ and apply Theorem 1(ii). \square

2.1 Convergence to subprobability measures: Helly's theorem

It frequently turns out to be convenient to extend the concept of convergence in distribution to cases when the limiting measure is not a probability measure. For example, we may say that $\mu_n = \delta_n$ converges in distribution to $\mu = 0$, since the sequence of corresponding CDFs $F_n(x) = 1_{[n, \infty)}(x)$ converges to $F_0(x) = 0$ at every point of continuity. Similar to Theorem 1 we have the following equivalent representations:

Proposition 2. Let \mathbb{P}_n and μ be measures on \mathbb{R} with CDFs F_n and F , respectively. The following are equivalent:

1. For every a, b —points of continuity of F :

$$F_n(b) - F_n(a) \rightarrow F(b) - F(a)$$

2. For every continuous f possessing limits at infinity ($f(-\infty) = f(+\infty) = 0$):

$$\int f d\mathbb{P}_n \rightarrow \int f d\mu$$

3. For every bounded Borel E with $\mathbb{P}[\partial E] = 0$:

$$\mathbb{P}_n[E] \rightarrow \mathbb{P}[E]$$

In this case we say \mathbb{P}_n converges to μ (weakly, or in the vague topology) and write $\mathbb{P}_n \rightarrow \mu$.

Note: In the case where μ is a probability measure the above definition coincides with convergence in distribution. Note however, that $\mathbb{P}_n \rightarrow \mu$ does not imply $F_n(b) \rightarrow F(b)$ or even that this limit exists. As an example consider

$$\frac{1}{2}\delta_{n(-1)^n} + \frac{1}{2}\delta_0 \rightarrow \frac{1}{2}\delta_0$$

Theorem 3 (Helly). Any (infinite) collection of probability measures on $(\mathbb{R}, \mathcal{B})$ contains a sequence converging in distribution to a measure with $\mu^*(\mathbb{R}) \leq 1$.

Caution: Theorem does not imply that a sequence of probability measures contains a subsequence converging to a probability measure. Necessary and sufficient conditions for the latter will be discussed in the next Section.

Proof. Let $\{r_j, j = 1, \dots\}$ be enumeration of rationals in \mathbb{R} . Let $\{\mu_s, s \in S\}$ be the collection of probability measures and the respective CDFs. For each r_j the values taken by $F_s(r_j)$ belong to $[0, 1]$. By compactness of $[0, 1]$ it follows that for every j there is a sequence $s_{j,n}$ indexed by n such that

$$F_{s_{j,n}}(r_j) \rightarrow F(r_j).$$

Furthermore, we may arrange the choice so that $s_{j,n}$ is a subsequence of $s_{j-1,n}$, etc. Then define

$$F_n \triangleq F_{s_{n,n}}$$

(Cantor's diagonal process). Since $s_{n,n}$ is a subsequence of $s_{j,n}$ for every j we have

$$F_n(r_j) \rightarrow F(r_j) \quad \forall r_j \in \mathbb{Q}.$$

Finally, define

$$F^*(x) = \inf_{r > x} F(r).$$

One easily verifies that F^* is a right-continuous, non-decreasing function on \mathbb{R} with

$$0 \leq F^*(-\infty) \leq F(+\infty) \leq 1.$$

Thus there is a unique measure on $(\mathbb{R}, \mathcal{B})$ so that

$$\mu^*((a, b]) = F^*(b) - F^*(a).$$

The proof completes by showing that $F_n(x) \rightarrow F^*(x)$ at every point of continuity of F^* .

First, notice that for every rational we have

$$F^*(r) \geq F(r) \triangleq \lim_{n \rightarrow \infty} F_n(r).$$

Thus for every $x > r$ we have by monotonicity of F_n :

$$F^*(r) \geq \lim_{n \rightarrow \infty} F_n(r) \geq \lim_{n \rightarrow \infty} \sup F_n(x)$$

Taking limit as $\searrow x$ and using right-continuity of F^* we obtain

$$F^*(x) \geq \lim_{n \rightarrow \infty} \sup F_n(x) \quad \forall x \in \mathbb{R} \quad (9)$$

Conversely, for every $x_1 < x$ and some rational between them we have

$$F^*(x_1) \leq \lim_{n \rightarrow \infty} F_n(r) \leq \lim_{n \rightarrow \infty} \inf F_n(x)$$

by monotonicity of F_n . Thus, taking the limit as $\nearrow x$ we get:

$$F^*(x-) \leq \lim_{n \rightarrow \infty} \inf F_n(x). \quad (10)$$

Together (9) and (10) establish convergence at the points of continuity since $F^*(x-) = F^*(x)$. \square

2.2 Convergence to probability measures: tightness

Definition 4. A collection of probability measures $\{\mathbb{P}_s, s \in S\}$ on $(\mathbb{R}, \mathcal{B})$ is called **tight** if for every $\epsilon > 0$ there exists a compact set $K = [-A, A]$ such that

$$\sup_{s \in S} \mathbb{P}_s(K^c) \leq \epsilon.$$

In words, a collection is tight if there is no “escaping of mass to infinity”, similar to the case $\mathbb{P}_n = \delta_n$.

Theorem 4 (Prokhorov’s criterion) A collection of probability measures $\{\mathbb{P}_s, s \in S\}$ on $(\mathbb{R}, \mathcal{B})$ is tight if and only if every sequence contains a subsequence converging to a probability measure.

Proof. If collection is tight, then every sequence contains a convergent subsequence by Helly's theorem $\mathbb{P}_n \rightarrow \mu_*$. Assuming without loss of generality that $\mu^*(\{n\} \cup \{-n\}) = 0$ (otherwise just shift these slightly) and since $\mathbb{P}_n([-n, n]^c) \leq \epsilon$ we have

$$\mu_*([-n, n]^c) = \lim_{k \rightarrow \infty} \mathbb{P}_k([-n, n]^c) \leq \epsilon$$

for each n and thus $\mu_*(\mathbb{R}) = 1$. Conversely, if collection is not tight, then there exist $\epsilon_0 > 0$ and measures \mathbb{P}_n such that

$$\mathbb{P}_n([-n, n]^c) \geq \epsilon_0 > 0$$

for all n . If there is a subsequence $\mathbb{P}_{n_k} \rightarrow \mu_*$ then $\mu^*(\mathbb{R}) \leq 1 - \epsilon_0$ and cannot be a probability measure. \square

3 THE HIERARCHY OF CONVERGENCE CONCEPTS

Theorem 5. We have

$$[X_n \xrightarrow{\text{a.s.}} X] \Rightarrow [X_n \xrightarrow{\text{i.p.}} X] \Rightarrow [X_n \xrightarrow{\text{d}} X] \iff [\phi_{X_n}(t) \rightarrow \phi_X(t), \forall t].$$

(The first two implications assume that all random variables be defined on the same probability space.)

Proof:

(a) $[X_n \xrightarrow{\text{a.s.}} X] \Rightarrow [X_n \xrightarrow{\text{i.p.}} X]$:

We give a short proof, based on the DCT, but more elementary proofs are also possible. Fix some $\epsilon > 0$. Let

$$Y_n = I_{\{|X_n - X| \geq \epsilon\}}.$$

If $X_n \xrightarrow{\text{a.s.}} X$, then $Y_n \xrightarrow{\text{a.s.}} 0$. By the DCT $\mathbb{E}[Y_n] \rightarrow 0$. On the other hand,

$$\mathbb{E}[Y_n] = \mathbb{P}(|X_n - X| \geq \epsilon).$$

This implies that $\mathbb{P}(|X_n - X| \geq \epsilon) \rightarrow 0$, and therefore $X_n \xrightarrow{\text{i.p.}} X$.

(b) $[X_n \xrightarrow{\text{i.p.}} X] \Rightarrow [X_n \xrightarrow{\text{d}} X]$:

Since the magnitude of derivative of the function $\cos(ta)$ is bounded by t , we have that

$$|\cos(ta) - \cos(tb)| \leq t|a - b|.$$

Notice, however, that when $|a - b| > 2$ this bound is not good, so overall we get:

$$|\cos(ta) - \cos(tb)| \leq \begin{cases} t\epsilon, & |a - b| \leq 2\epsilon/t, \\ 2, & |a - b| > 2\epsilon/t \end{cases}$$

Using this with $a = X_n$ and $b = X$ and taking the expectation we get

$$\mathbb{E}[\cos(tX_n) - \cos(tX)] \leq t\epsilon\mathbb{P}[|X_n - X| \leq \epsilon/t] + 2\mathbb{P}[|X_n - X| > 2\epsilon/t].$$

The second term converges to zero as $\epsilon \rightarrow 0$ for any t, ϵ , whereas the first term is bounded by $t\epsilon$. Thus, first taking $\lim_{n \rightarrow \infty}$ and then $\lim_{\epsilon \rightarrow 0}$ we obtain

$$\mathbb{E}[\cos(tX_n)] \rightarrow \mathbb{E}[\cos(tX)]$$

for every $t \in \mathbb{R}$. Similar proof shows

$$\mathbb{E}[\sin(tX_n)] \rightarrow \mathbb{E}[\sin(tX)].$$

Thus, characteristic functions $\phi_{X_n} \rightarrow \phi_X$ and from the last part we get the claimed result.

(c) $[X_n \xrightarrow{d} X] \Rightarrow [\phi_{X_n}(t) \rightarrow \phi_X(t), \forall t]$:

Suppose that $X_n \xrightarrow{d} X$. Let Y_n and Y be as in Theorem 2, so that $Y_n \xrightarrow{\text{a.s.}} Y$.

Then, for any $t \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \phi_{X_n}(t) = \lim_{n \rightarrow \infty} \phi_{Y_n}(t) = \lim_{n \rightarrow \infty} \mathbb{E}[e^{itY_n}] = \mathbb{E}[e^{itY}] = \phi_Y(t) = \phi_X(t),$$

where we have made use of the facts $Y_n \xrightarrow{\text{a.s.}} Y$, $e^{itY_n} \xrightarrow{\text{a.s.}} e^{itY}$, and the DCT.

Finally, the converse direction will be established in the next lecture \square

Exercise 1 (Smoothing method) Show that for every \mathbb{P}_X on $(\mathbb{R}, \mathcal{B})$ there exist a sequence $\mathbb{P}_{X_n} \xrightarrow{d} \mathbb{P}_X$ such that every \mathbb{P}_{X_n} has continuous, bounded, infinitely-differentiable PDF. Steps:

1. Show $X_\epsilon = X + \epsilon Z \xrightarrow{d} X$ as $\epsilon \rightarrow 0$.
2. Let $X \perp\!\!\!\perp Z \sim \mathcal{N}(0, 1)$ and show that CDF of X_ϵ is continuous (*Hint: BCT*) and differentiable (*Hint: Fubini*) with derivative

$$f_{X_\epsilon}(a) = \mathbb{E} \left[f_Z \left(\frac{a - X}{\epsilon} \right) \frac{1}{\epsilon} \right]$$

3. Show that $a \mapsto f_{X_\epsilon}(a)$ is continuous.
4. Conclude the proof (*Hint: derivatives of f_Z are uniformly bounded on \mathbb{R} .*)

At this point, it is natural to ask whether the converses of the implications in Theorem 5 hold. For the first two, the answer is, in general, “no”, although we will also note some exceptions.

3.1 Convergence almost surely versus in probability

$[X_n \xrightarrow{\text{i.p.}} X]$ **does not imply** $[X_n \xrightarrow{\text{a.s.}} X]$:

Let X_n be equal to 1, with probability $\frac{1}{n}$, and equal to zero otherwise. Suppose that the X_n are independent. We have $X_n \xrightarrow{\text{i.p.}} 0$. On the other hand, by the Borel-Cantelli lemma, the event $\{X_n = 1, \text{ i.o.}\}$ has probability 1 (check this). Thus, for almost all ω , the sequence $X_n(\omega)$ does not converge to zero.

Nevertheless, a weaker form of the converse implication turns out to be true.

If $X_n \xrightarrow{\text{i.p.}} X$, then there exists an increasing (deterministic) sequence of integers, such that $\lim_{k \rightarrow \infty} X_{n_k} = X$, a.s. (We omit the proof.)

For an illustration of the last statement in action, consider the preceding counterexample. If we let $n_k = k^2$, then we note that $\mathbb{P}(X_{n_k} \neq 0) = 1/k^2$, which is summable. By the Borel-Cantelli lemma, the event $\{X_{n_k} \neq 0\}$ will occur for only finitely many k , with probability 1. Therefore X_{n_k} converges, a.s., to the zero random variable.

3.2 Convergence in probability versus in distribution

The converse turns out to be false in general, but true when the limit is deterministic.

$[X_n \xrightarrow{d} X]$ **does not imply** $[X_n \xrightarrow{\text{i.p.}} X]$:

Let the random variables X, X_n be i.i.d. and nonconstant random variables, in which case we have (trivially) $X_n \xrightarrow{d} X$. Fix some ϵ . Then, $\mathbb{P}(|X_n - X| \geq \epsilon)$ is positive and the same for all n , which shows that X_n does not converge to X , in probability.

$[X_n \xrightarrow{d} c]$ **implies** $[X_n \xrightarrow{\text{i.p.}} c]$:

The proof is very simple: by definition we have

$$\mathbb{P}[X_n \leq c - \epsilon] \rightarrow 0, \quad \mathbb{P}[X_n > c + \epsilon] \rightarrow 0$$

for any $\epsilon > 0$. Thus

$$\mathbb{P}[|X_n - c| > \epsilon] \rightarrow 0$$

for any ϵ as well.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Lecture 17

Fall 2018

LAWS OF LARGE NUMBERS AND CENTRAL LIMIT THEOREM

Contents

1. Convergence in distribution and characteristic functions
2. Useful inequalities
3. The weak law of large numbers
4. The central limit theorem
5. Berry-Esseen theorem

1 USEFUL INEQUALITIES

Markov inequality: If X is a nonnegative random variable, then $\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a$.

Proof: Let I be the indicator function of the event $\{X \geq a\}$. Then, $aI \leq X$. Taking expectations of both sides, we obtain the claimed result. \square

Chebyshev inequality: $\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \text{var}(X)/\epsilon^2$.

Proof: Apply the Markov inequality, to the random variable $|X - \mathbb{E}[X]|^2$, and with $a = \epsilon^2$. \square

2 CONVERGENCE IN DISTRIBUTION vs CHARACTERISTIC FUNCTIONS

We know that equality of two characteristic functions implies equality of the corresponding distributions. It is then plausible to hope that “near-equality” of characteristic functions implies “near equality” of corresponding distributions. This would be essentially a statement that the mapping from characteristic functions to distributions is a continuous one.

Theorem 1. Continuity of inverse transform Let X and X_n be random variables with given CDFs and corresponding characteristic functions. We have

$$[\phi_{X_n}(t) \rightarrow \phi_X(t), \forall t] \Rightarrow [X_n \xrightarrow{d} X].$$

Proof. First, suppose that we are in the special situation that $|\phi_{X_n}(t)| \leq g(t)$ where $g(t)$ is positive and integrable (on \mathbb{R}) function. Then, the inverse Fourier transform exists and we conclude that each X_n and X in such a case must possess a pdf (i.e. X_n 's and X are all continuous random variables) given by

$$f_{X_n}(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \phi_{X_n}(t) dt$$

and similarly for f_X . By the DCT we conclude that

$$f_{X_n}(x) \rightarrow f_X(x)$$

for every x . It will be shown later (in the lecture on uniform integrability) that convergence of pdfs implies convergence in distribution.

Second, to reduce to a special case proven above, notice the following: If Z_ϵ is a collection of random variables (independent of X) such that $\mathbb{P}[|Z_\epsilon| \leq \epsilon] = 1$ then

$$\forall \epsilon > 0 \quad X_n + Z_\epsilon \xrightarrow{d} X_n \iff X_n \xrightarrow{d} X. \quad (1)$$

Finally, notice that if we take Z_ϵ to have triangular pdf

$$f_{Z_\epsilon}(x) = \begin{cases} \frac{1}{\epsilon^2}(x + \epsilon), & x \in (-\epsilon, 0] \\ \frac{1}{\epsilon^2}(\epsilon - x), & x \in (0, \epsilon) \\ 0, & \text{o/w} \end{cases}$$

then $\phi_{Z_\epsilon}(t) = \frac{4\sin^2(t\epsilon/2)}{t^2\epsilon^2} \leq \frac{\text{const}}{1+\epsilon^2t^2}$ (a calculation). Since $\phi_{X_n+Z_\epsilon} = \phi_{X_n}\phi_{Z_\epsilon}$ we see that sequence of random variables $X_n + Z_\epsilon$ satisfies conditions of the special case above. Application of (1) completes the proof. \square

The preceding theorem involves two separate conditions: (i) the sequence of characteristic functions ϕ_{X_n} converges (pointwise), and (ii) the limit is the characteristic function associated with some other random variable. If we are only given the first condition (pointwise convergence), how can we tell if the limit is indeed a legitimate characteristic function associated with some random

variable? One way is to check for various properties that every legitimate characteristic function must possess. One such property is continuity: if t^* , then (using dominated convergence),

$$\lim_{t \rightarrow t^*} \phi_X(t) = \lim_{t \rightarrow t^*} \mathbb{E}[e^{itX}] = \mathbb{E}[e^{it^*X}] = \phi_X(t^*).$$

It turns out that continuity at zero is all that needs to be checked.

Theorem 2. Continuity of inverse transform Let X_n be random variables with characteristic functions ϕ_{X_n} , and suppose that the limit $\phi(t) = \lim_{n \rightarrow \infty} \phi_{X_n}(t)$ exists for every. Then, either

- (i) The function ϕ is discontinuous at zero (in this case does not converge in distribution); or
- (ii) The function ϕ is continuous at zero, there exists a random variable whose characteristic function is ϕ , and $X_n \xrightarrow{d} X$.

To illustrate the two possibilities in Theorem 2, consider a sequence, and assume that X_n is exponential with parameter λ_n , so that $\phi_{X_n}(t) = \lambda_n / (\lambda_n - it)$.

- (a) Suppose that λ_n converges to a positive number. Then, the sequence of characteristic functions ϕ_{X_n} converges to the function ϕ defined by $\phi(t) = \lambda / (\lambda - it)$. We recognize this as the characteristic function of an exponential distribution with parameter λ . In particular, we conclude that X_n converges in distribution to an exponential random variable with parameter
- (b) Suppose now that λ_n converges to zero. Then,

$$\lim_{n \rightarrow \infty} \phi_{X_n}(t) = \lim_{n \rightarrow \infty} \frac{\lambda_n}{\lambda_n - it} = \lim_{\lambda \downarrow 0} \frac{\lambda}{\lambda - it} = \begin{cases} 1, & \text{if } t = 0, \\ 0, & \text{if } t \neq 0. \end{cases}$$

Thus, the limit of the characteristic functions is discontinuous at 0, and X_n does not converge in distribution. Intuitively, this is because the distribution of X_n keeps spreading in a manner that does not yield a limiting distribution.

Proof. We only need to show (ii). The main step is to show that if continuous at zero, then collection of measures $\mathbb{P}_{X_n}, n = 1, 2, \dots$ is tight. Indeed, from tightness and Prokhorov's criterion we conclude that there exists a convergent subsequence $\mathbb{P}_{X_{n_k}} \rightarrow \mathbb{P}_X$ and since $\phi_{n_k} \rightarrow \phi$ the characteristic function of \mathbb{P}_X is precisely ϕ , and thus \mathbb{P}_X is identified uniquely. A short argument (Exercise!) shows that then we must have $\mathbb{P}_{X_n} \rightarrow \mathbb{P}_X$.

Showing that continuity of ϕ implies tightness requires the following (Fourier analytic) trick: Tails of the distribution can be read off the small-neighborhood averages of ϕ around 0. Formally, we have

Lemma 1. *Let Y have characteristic function ϕ_Y then for all $a > 0$:*

$$\mathbb{P}\left[|Y| \geq \frac{1}{a}\right] \leq \frac{7}{a} \int_0^a [1 - \operatorname{Re} \phi_Y(t)] dt$$

Lemma indeed implies tightness: From continuity of ϕ for every $\epsilon > 0$ there exists small enough $a > 0$ such that

$$\frac{1}{a} \int_0^a (1 - \operatorname{Re} \phi(t)) dt < \frac{\epsilon}{2}$$

and from the DCT there is also such that for all $n \geq n_0$ we have

$$\frac{1}{a} \int_0^a (1 - \operatorname{Re} \phi_n(t)) dt \leq \frac{1}{a} \int_0^a (1 - \operatorname{Re} \phi(t)) + \frac{\epsilon}{2} \leq \epsilon.$$

Finally, we may take $n \geq n_0$ such that

$$\sup_{n \leq n_0} \mathbb{P}[|X_n| \geq A] \leq \epsilon$$

to conclude the tightness of the whole $\{B_{X_n}\}$.

It remains to prove the Lemma. Roughly, the idea is the following. Let have PDF f_Y with mass $\delta > 0$ outside $[-A, A]$. Then ϕ_Y is a Fourier transform of f_Y . It is well-known that multiplication of functions corresponds to convolution of Fourier transforms, and vice-versa. Thus, we conclude that $\frac{1}{\epsilon} \cdot \frac{1}{\epsilon} * 1_{(-\epsilon, \epsilon)}$ is a Fourier transform of $f_Y(x) \cdot \frac{\sin \epsilon x}{\epsilon x}$. However, note that $\frac{\sin \epsilon x}{\epsilon x}$ kills the tails of f_Y and hence the Fourier transform of the product evaluated at zero should be around $1 - \frac{\delta}{\epsilon A}$.

Rigorously, from

$$1 - \operatorname{Re} \phi_Y(t) = \mathbb{E}[1 - \cos(tY)]$$

by Fubini we have

$$\frac{1}{a} \int_0^a [1 - \operatorname{Re} \phi_Y(t)] dt = \mathbb{E} \frac{1}{a} \int_0^a [1 - \cos tY] dt \quad (2)$$

$$= \mathbb{E} \left[1 - \frac{\sin aY}{aY} \right] \quad (3)$$

$$\geq (1 - \sin 1) \mathbb{P}\left[|Y| \geq \frac{1}{a}\right], \quad (4)$$

where in the last step we used the fact that $\frac{\sin u}{u}$ is a non-negative function, exceeding $(1 - \sin 1)$ for $|u| > 1$. From (4) lemma follows by noting $(1 - \sin 1) > \frac{1}{7}$. This concludes the proof of Lemma and Theorem. \square

3 THE WEAK LAW OF LARGE NUMBERS

Intuitively, an expectation can be thought of as the average of the outcomes over an infinite repetition of the same experiment. If so, the observed average in a finite number of repetitions (which is called **sample mean**) should approach the expectation, as the number of repetitions increases. This is a vague statement, which is made more precise by so-called laws of large numbers.

Theorem 3. (Weak law of large numbers) Let X_n be a sequence of i.i.d. random variables, and assume that $\mathbb{E}[|X_1|] < \infty$. Let $S_n = X_1 + \dots + X_n$. Then,

$$\frac{S_n}{n} \xrightarrow{\text{i.p.}} \mathbb{E}[X_1].$$

This is called the “weak law” in order to distinguish it from the “strong law” of large numbers, which asserts, under the same assumptions, that $\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[X_1]$. Of course, since almost sure convergence implies convergence in probability, the strong law implies the weak law. On the other hand, the weak law can be easier to prove, especially in the presence of additional assumptions. Indeed, in the special case where the X_i have mean μ and **finite variance**, Chebyshev’s inequality yields, for every $\epsilon > 0$,

$$\mathbb{P}(|(S_n/n) - \mu| \geq \epsilon) \leq \frac{\text{var}(S_n/n)}{\epsilon^2} = \frac{\text{var}(X_1)}{n\epsilon^2}, \quad (5)$$

which converges to zero, as $n \rightarrow \infty$, thus establishing convergence in probability.

Historical note: WLLN has been one of the focal points of the development of the probability theory. Reader is welcome to muse upon the mathematical progress made since 1713, when J. Bernoulli proved WLLN for iid $X_j \sim \text{Bern}(p)$. It took him 20 years (his own account) and he referred to it as his “Golden Theorem”. The simple proof (5) under finite variance only appeared in Chebyshev’s work in 1867 (who used an inequality due to Bienaymé, which we now call Chebyshev’s). In 1913 A. Markov organized a big celebration on the occasion of 200’th anniversary of LLN. The final form of the WLLN as given in Theorem 3 was obtained by Khintchine in 1929. For more history see [3].

Before we proceed to the proof for the general case, we note two important facts that we will use.

- (a) **First-order Taylor series expansion** Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a function that has a derivative at zero, denoted by L be a function that represents the error in a first order Taylor series approximation:

$$g(\epsilon) = g(0) + L\epsilon + h(\epsilon).$$

By the definition of the derivative, we have

$$L = \lim_{\epsilon \rightarrow 0} \frac{g(\epsilon) - g(0)}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{d\epsilon + h(\epsilon)}{\epsilon} = d + \lim_{\epsilon \rightarrow 0} \frac{h(\epsilon)}{\epsilon}.$$

Thus, $h(\epsilon)/\epsilon$ converges to zero, as $\epsilon \rightarrow 0$. A function h with this property is often written as $o(\epsilon)$. This discussion also applies to complex-valued functions, by considering separately the real and imaginary parts.

- (b) **A classical sequence** Recall the well known fact

$$\lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n = e^a, \quad a \in \mathbb{R}. \quad (6)$$

We note (without proof) that this fact remains true even when a is a complex number. Furthermore, with little additional work, it can be shown that if $\{a_n\}$ is a sequence of complex numbers that converges, then,

$$\lim_{n \rightarrow \infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Proof of Theorem 3 Let $\mu = \mathbb{E}[X_1]$. Fix some $t \in \mathbb{R}$. Using the assumption that the X_i are independent, and the fact that the derivative of ϕ_X at $t = 0$ equals $i\mu$, the characteristic function of n/n is of the form

$$\phi_n(t) = (\mathbb{E}[e^{itX_1/n}])^n = (\phi_{X_1}(t/n))^n = \left(1 + \frac{\mu it}{n} + o(t/n)\right)^n,$$

where the function satisfies $\lim_{\epsilon \rightarrow 0} o(\epsilon)/\epsilon = 0$. Therefore,

$$\lim_{n \rightarrow \infty} \phi_{X_n}(t) = e^{i\mu t}, \quad \forall t.$$

We recognize $e^{i\mu t}$ as the characteristic function associated with a random variable which is equal to μ with probability one.

Applying Theorem 1 from the previous lecture (continuity of inverse-transforms), we conclude that ξ_n/n converges to μ , in distribution. Furthermore, as mentioned in the previous lecture, convergence in distribution to a constant implies convergence in probability. \square

Remark: It turns out that the assumption $\mathbb{E}[|X_1|] < \infty$ can be relaxed, although not by much. Suppose that the distribution of X_1 is symmetric around zero. It is known that $S_n/n \rightarrow 0$, in probability, if and only if $\lim_{n \rightarrow \infty} n\mathbb{P}(|X_1| > n) = 0$. There exist distributions that satisfy this condition, while $\mathbb{E}[|X_1|] = \infty$. On the other hand, it can be shown that any such distribution satisfies $\mathbb{E}[|X_1|^{1-\epsilon}] < \infty$, for every $\epsilon > 0$, so the condition $\lim_{n \rightarrow \infty} n\mathbb{P}(|X_1| > n) = 0$ is not much weaker than the assumption of a finite mean.

4 THE CENTRAL LIMIT THEOREM

Suppose that X_1, X_2, \dots are i.i.d. with common (and finite) mean and variance σ^2 . Let $S_n = X_1 + \dots + X_n$. The central limit theorem (CLT) asserts that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to a standard normal random variable. For a discussion of the uses of the central limit theorem, see the handout from [BT] (pages 388–394).

Proof of the CLT For simplicity, suppose that the random variables have zero mean and unit variance. Finiteness of the first two moments implies that $\phi_{X_1}(t)$ is twice differentiable at zero. The first derivative is the mean (assumed zero), and the second derivative is $\mathbb{E}[X_1^2]$ (assumed equal to one), and we can write

$$\phi_X(t) = 1 - t^2/2 + o(t^2),$$

where $o(t^2)$ indicates a function such that $t^2/o(t^2) \rightarrow 0$ as $t \rightarrow 0$. The characteristic function of S_n/\sqrt{n} is of the form

$$(\phi_X(t/\sqrt{n}))^n = \left(1 - \frac{t^2}{2n} + o(t^2/n)\right)^n.$$

For any fixed t , the limit as $n \rightarrow \infty$ is $e^{-t^2/2}$, which is the characteristic function ϕ_Z of a standard normal random variable. Since $\phi_{S_n/\sqrt{n}}(t) \rightarrow \phi_Z(t)$ for every t , we conclude that S_n/\sqrt{n} converges to Z , in distribution. \square

The central limit theorem, as stated above, does not give any information on the PDF or PMF of S_n . However, some further refinements are possible, under some additional assumptions. We state, without proof, two such results.

- (a) Suppose that $\int |\phi_{X_1}(t)|^r dt < \infty$, for some positive integer r . Then, S_n is a continuous random variable for every $r \geq r$, and the PDF of $(S_n -$

$\mu_n)/(\sigma\sqrt{n})$ converges pointwise to the standard normal PDF:

$$\lim_{n \rightarrow \infty} f_n(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad \forall z.$$

In fact, convergence is uniform over all

$$\lim_{n \rightarrow \infty} \sup_z |f_n(z) - \frac{1}{\sqrt{2\pi}} e^{-z^2/2}| = 0.$$

- (b) Suppose that X_i is a discrete random variable that takes values of the form $a + kh$, where a and h are constants, and k ranges over the integers. Suppose furthermore that X has zero mean and unit variance. Then, for any z of the form $z = (na + kh)/\sqrt{n}$ (these are the possible values of S_h/\sqrt{n}), we have

$$\lim_{n \rightarrow \infty} \frac{\sqrt{n}}{h} \mathbb{P}(S_n = z) = \frac{1}{2\pi} e^{-z^2/2}.$$

4.1 Berry-Esseen theorem

It turns out that CDF of normalized sums approaches the CDF of standard normal *uniformly* on all of \mathbb{R} with speed $\frac{1}{\sqrt{n}}$:

$$\mathbb{P}\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \lambda\right] = \Phi(\lambda) \pm \frac{\text{const}}{\sqrt{n}} \quad \forall \lambda.$$

The following is a precise version. Just like for the CLT there are great many refinements and extensions. For proof see e.g. Theorem 2, Chapter XVI.5 in [1].

Theorem 4(Berry-Esseen). *Let X_k , $k = 1, \dots, n$ be independent (possibly not identically distributed) with*

$$\mu_k = \mathbb{E}[X_k], \tag{7}$$

$$\sigma_k^2 = \text{var}[X_k], \tag{8}$$

$$t_k = \mathbb{E}[|X_k - \mu_k|^3], \tag{9}$$

$$\sigma^2 = \sum_{k=1}^n \sigma_k^2, \tag{10}$$

$$T = \sum_{k=1}^n t_k. \tag{11}$$

Then for any ¹ $-\infty < \lambda < \infty$

$$\mathbb{P} \left[\sum_{k=1}^n (X_k - \mu_k) \leq \lambda \sigma \right] - \Phi(\lambda) \leq \frac{6T}{\sigma^3}, \quad (12)$$

where Φ is the CDF of $\mathcal{N}(0, 1)$.

References

- [1] W. Feller, *An Introduction to Probability Theory and Its Applications, Volume II*, Second edition, John Wiley & Sons, Inc., New York, 1971.
- [2] P. Van Beeck, “An application of Fourier methods to the problem of sharpening the Berry-Esseen inequality *Z. Wahrscheinlichkeitstheorie und Verw. Geb.*, vol. 23, 187-196, 1972.
- [3] E. Seneta, “A Tricentenary history of the Law of Large Numbers”, *Bernoulli*, vol. 19, no. 4, pp.1088–1121, 2013.

¹Note that for i.i.d X_k it is known [2] that the factor 6 in (12) can be replaced by 7975.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

LAWS OF LARGE NUMBERS – II**Contents**

1. The strong law of large numbers
2. The Chernoff bound

1 THE STRONG LAW OF LARGE NUMBERS

While the weak law of large numbers establishes convergence of the sample mean, in probability, the strong law establishes almost sure convergence.

Before we proceed, we point out two common methods for proving almost sure convergence.

Proposition 1. *Let $\{X_n\}$ be a sequence of random variables, not necessarily independent.*

- (i) *If $\sum_{n=1}^{\infty} \mathbb{E}[|X_n|^s] < \infty$, and $s > 0$, then $X_n \xrightarrow{\text{a.s.}} 0$.*
- (ii) *If $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > \epsilon) < \infty$, for every $\epsilon > 0$, then $X_n \xrightarrow{\text{a.s.}} 0$.*
- (iii) *$X_n \xrightarrow{\text{a.s.}} 0$ iff for every $\epsilon > 0$ we have $\mathbb{P}[\sup_{m \geq n} |X_m| > \epsilon] \rightarrow 0$ as $n \rightarrow \infty$.*

Proof. (i) By the monotone convergence theorem, we obtain $\sum_{n=1}^{\infty} \mathbb{E}[|X_n|^s] < \infty$, which implies that the random variable $\sum_{n=1}^{\infty} |X_n|^s$ is finite, with probability 1. Therefore $|X_n|^s \xrightarrow{\text{a.s.}} 0$, which also implies that $X_n \xrightarrow{\text{a.s.}} 0$.

(ii) Setting $\epsilon = 1/k$, for any positive integer k , the Borel-Cantelli Lemma shows that the event $\{|X_n| > 1/k\}$ occurs only a finite number of times, with probability 1. Thus $\mathbb{P}(\limsup_{n \rightarrow \infty} X_n > 1/k) = 0$, for every positive integer k .

Note that the sequence of events $\{\limsup_{n \rightarrow \infty} |X_n| > 1/k\}$ is monotone and converges to the event $\{\limsup_{n \rightarrow \infty} |X_n| > 0\}$. The continuity of probability measures implies that $\mathbb{P}(\limsup_{n \rightarrow \infty} |X_n| > 0) = 0$. This establishes that $X_n \xrightarrow{\text{a.s.}} 0$.

(iii) This follows since

$$\{\omega : X_n(\omega) \not\nearrow 0\} = \bigcup_{\epsilon > 0} \bigcap_{n \geq 1} \{\omega : \sup_{m \geq n} |X_m(\omega)| > \epsilon\}$$

□

Theorem 1: Let X, X_1, X_2, \dots be i.i.d. random variables, and assume that $\mathbb{E}[|X|] < \infty$. Let $S_n = X_1 + \dots + X_n$. Then, S_n/n converges to a finite constant c almost surely if and only if $\mathbb{E}[|X|] < \infty$ and $c = \mathbb{E}[X]$.

Proof of necessity of $\mathbb{E}[|X|] < \infty$. Note that

$$\frac{1}{n} \sum_{k=1}^n a_k \rightarrow 0 \quad \Rightarrow \quad \frac{a_n}{n} \rightarrow 0$$

(just write $a_n = \sum_{k=1}^n a_k - \sum_{k=1}^{n-1} a_k$). Thus, we have

$$\frac{X_n - c}{n} \xrightarrow{\text{a.s.}} 0$$

And by Borel-Cantelli this implies

$$\sum_{n=1}^{\infty} \mathbb{P}[|X - c| > n] < \infty.$$

On the other hand, from $\mathbb{E}[|Y|] = \int_0^\infty \mathbb{P}[|Y| > t] dt$ we derive

$$\mathbb{E}[|X - c|] \leq 1 + \sum_{n=1}^{\infty} \mathbb{P}[|X - c| > n] < \infty,$$

which implies $\mathbb{E}[|X|] < \infty$. By what is to be shown, whenever $\mathbb{E}[|X|] < \infty$ we should have

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[X]$$

which implies $c = \mathbb{E}[X]$.

Proof of convergence, assuming $\mathbb{E}[X^2] < \infty$. We now consider the case where we assume additionally that X has a finite second moment $\mathbb{E}[X^2]$. We have

$$\mathbb{E} \left[\left(\frac{S_n}{n} - \mu \right)^2 \right] = \frac{\text{var}(X)}{n}.$$

If we only consider values n that are perfect squares, we obtain

$$\sum_{i=1}^{\infty} \mathbb{E} \left[\left(\frac{S_{i^2}}{i^2} - \mu \right)^2 \right] = \sum_{i=1}^{\infty} \frac{\text{var}(X)}{i^2} < \infty,$$

which implies that $((S_{i^2}/i^2) - \mathbb{E}[X])^2$ converges to zero, with probability 1. Therefore, S_{i^2}/i^2 converges to $\mathbb{E}[X]$, with probability 1.

Suppose that the random variable X_i are nonnegative. Consider some i such that $i^2 \leq n < (i+1)^2$. We then have $S_{i^2} \leq S_n \leq S_{(i+1)^2}$. It follows that

$$\frac{S_{i^2}}{(i+1)^2} \leq \frac{S_n}{n} \leq \frac{S_{(i+1)^2}}{i^2},$$

or

$$\frac{i^2}{(i+1)^2} \cdot \frac{S_{i^2}}{i^2} \leq \frac{S_n}{n} \leq \frac{(i+1)^2}{i^2} \cdot \frac{S_{(i+1)^2}}{(i+1)^2}.$$

As $n \rightarrow \infty$, we also have $i \rightarrow \infty$. Since $i/(i+1) \rightarrow 1$, and since $S_{i^2} \cdot i^2$ converges to $\mathbb{E}[X]$, with probability 1, we see that for almost all sample points, S_n/n is sandwiched between two sequences that converge to $\mathbb{E}[X]$. This proves that $S_n/n \rightarrow \mathbb{E}[X]$, with probability 1.

For a general random variable X , we write it in the form $X = X^+ - X^-$, where X^+ and X^- are nonnegative. The strong law applied to X^+ and X^- separately, implies the strong law for X as well. \square

Proof of convergence (general case). The proof for the most general case (finite mean, but possibly infinite variance) is conceptually simple: We truncate the distribution of X and apply previous argument to $Y = X \cdot 1\{|X| < c\}$, so that the second moment of the latter is finite. Technically, this involves showing that difference $Y - X$, although potentially of infinite variance, cannot contribute much to the limiting value. The method is based upon what is called “maximal ergodic lemma”, or “weak L_1 ” estimate of the maximal function, see Lemma 1 below.

Without loss of generality we assume $\mathbb{E}[X] = 0$. Then by Proposition 1.(iii) it suffices to show for every $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P} \left[\sup_{m \geq n} \frac{|S_m|}{m} > \epsilon \right] = 0. \quad (1)$$

To that end, fix, choose (very large) $c > 0$ and define

$$Y_n = X_n \mathbf{1}\{|X_n| \leq c\} \quad (2)$$

$$Z_n = X_n \mathbf{1}\{|X_n| > c\} \quad (3)$$

$$T_n = \frac{Y_1 + \cdots + Y_n}{n} \quad (4)$$

$$Z^* = \sup_{n \geq 1} \frac{Z_1 + \cdots + Z_n}{n} \quad (5)$$

Note that Y_j and Z_j are iid. By dominated convergence theorem, as $n \rightarrow \infty$ we have $\mathbb{E}[Y] \rightarrow 0$ and $\mathbb{E}[|Z|] \rightarrow 0$. Therefore for any $\epsilon > 0$ it is possible to find $c > 0$ such that

$$|\mathbb{E}[Y]| \leq \frac{\epsilon}{4} \quad (6)$$

$$\mathbb{E}[|Z|] \leq \delta \frac{\epsilon}{2} \quad (7)$$

By the proof above, we have

$$T_n \xrightarrow{\text{a.s.}} \mathbb{E}[Y] \quad (8)$$

and therefore we have

$$\mathbb{P}\left[\sup_{m \geq n} \frac{|S_m|}{m} > \epsilon\right] \leq \mathbb{P}\left[\sup_{m \geq n} \frac{|T_m|}{m} + |Z^*| > \epsilon\right] \quad (9)$$

$$\leq \mathbb{P}\left[\sup_{m \geq n} \frac{|T_m|}{m} > \frac{\epsilon}{2}\right] + \mathbb{P}\left[|Z^*| > \frac{\epsilon}{2}\right] \quad (10)$$

$$\leq \mathbb{P}\left[\sup_{m \geq n} \frac{|T_m - \mathbb{E}[Y]|}{m} > \frac{\epsilon}{4}\right] + \mathbb{P}\left[|Z^*| > \frac{\epsilon}{2}\right] \quad (11)$$

where (9) is because $|S_m - T_m| \leq |Z^*|$, (10) follows from the union-bound applied to non-negative A, B :

$$\mathbb{P}[A + B > 2\epsilon] \leq \mathbb{P}[A > \epsilon] + \mathbb{P}[B > \epsilon]$$

and (11) is because of (6) and $|T_m - \mathbb{E}[Y]| \leq |T_m - \mathbb{E}[Y]| + \mathbb{E}[Y]$.

Taking limit of (11) as $n \rightarrow \infty$ the first term disappears due to (8) and Proposition 1.(iii). The Lemma 1 to follow bounds the second term as

$$\mathbb{P}\left[|Z^*| > \frac{\epsilon}{2}\right] \leq \frac{2\mathbb{E}[|Z|]}{\epsilon} \leq \delta.$$

Alltogether, we have shown

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left[\sup_{m \geq n} \frac{|S_m|}{m} > \epsilon \right] \leq \delta$$

for every $\delta > 0$, which proves (1) and the Theorem. \square

Lemma 1 (Estimate for the maximum of averages) *Let Z_n be iid with $\mathbb{E}[|Z|] < \infty$ then*

$$\mathbb{P} \left[\sup_{n \geq 1} \frac{|Z_1 + \dots + Z_n|}{n} > a \right] \leq \frac{\mathbb{E}[|Z|]}{a} \quad \forall a > 0$$

Proof. The argument for this Lemma has originally been quite involved, until a dramatically simple proof (below) was found by A. Garcia. We note that the result applies to arbitrary stationary processes $\{Z_n, n = 1, \dots\}$, although we only need an iid version here.

Define

$$S_n = \sum_{k=1}^n Z_k \tag{12}$$

$$L_n = \max\{0, Z_1, \dots, Z_1 + \dots + Z_n\} \tag{13}$$

$$M_n = \max\{0, Z_2, Z_2 + Z_3, \dots, Z_2 + \dots + Z_n\} \tag{14}$$

$$Z^* = \sup_{n \geq 1} \frac{S_n}{n} \tag{15}$$

It is sufficient to show that

$$\mathbb{E}[Z_1 1_{\{Z^* > 0\}}] \geq 0. \tag{16}$$

Indeed, applying (16) to $\tilde{Z}_1 = Z_1 - a$ and noticing that $\tilde{Z}^* = Z^* - a$ we obtain

$$\mathbb{E}[Z_1 1_{\{Z^* > a\}}] \geq a \mathbb{P}[Z^* > a],$$

from which Lemma follows by upper-bounding the left-hand side $\mathbb{E}[Z_1]$.

In order to show (16) we first notice $\{Z_1 > 0\} \nearrow \{Z^* > 0\}$. Next we notice that

$$Z_1 + M_n = \max\{S_1, \dots, S_n\}$$

and furthermore

$$Z_1 + M_n = L_n \quad \text{on } \{L_n > 0\}$$

Thus, we have

$$Z_1 1_{\{L_n > 0\}} = L_n - M_n 1_{\{L_n > 0\}}$$

where we do not need indicator in the first term since 0 on $\{L_n > 0\}^c$. Taking expectation we get

$$\mathbb{E}[Z_1 1_{\{L_n > 0\}}] = \mathbb{E}[L_n] - \mathbb{E}[M_n 1_{\{L_n > 0\}}] \quad (17)$$

$$\geq \mathbb{E}[L_n] - \mathbb{E}[M_n] \quad (18)$$

$$= \mathbb{E}[L_n] - \mathbb{E}[L_{n-1}] = \mathbb{E}[L_n - L_{n-1}] \geq 0, \quad (19)$$

where we used $M_n \geq 0$, the fact that M_n has the same distribution as L_{n-1} , and $L_n \geq L_{n-1}$, respectively. Taking limit as $n \rightarrow \infty$ in (19) we obtain (16). \square

2 THE CHERNOFF BOUND

Let again X, X_1, \dots be i.i.d., and $S_n = X_1 + \dots + X_n$. Let us assume, for simplicity, that $\mathbb{E}[X] = 0$. According to the weak law of large numbers, we know that $\mathbb{P}(S_n \geq na) \rightarrow 0$, for every $a > 0$. We are interested in a more detailed estimate of $\mathbb{P}(S_n \geq na)$, involving the rate at which this probability converges to zero. It turns out that if the moment generating function is finite on some interval $[0, c]$ (where $c > 0$), then $\mathbb{P}(S_n \geq na)$ decays exponentially with n , and much is known about the precise rate of exponential decay.

2.1 Upper bound

Let $M(s) = \mathbb{E}[e^{sX}]$, and assume that $M(s) < \infty$, for $s \in [0, c]$, where $c > 0$. Recall that $M_{S_n}(s) = \mathbb{E}[e^{s(X_1 + \dots + X_n)}] = (M(s))^n$. For any $s > 0$, the Markov inequality yields

$$\mathbb{P}(S_n \geq na) = \mathbb{P}(e^{sS_n} \geq e^{nsa}) \leq e^{-nsa} \mathbb{E}[e^{sS_n}] = e^{-nsa} (M(s))^n.$$

Every nonnegative value of s gives us a particular bound $\mathbb{P}(S_n \geq a)$. To obtain the tightest possible bound, we minimize over s and obtain the following result.

Theorem 2. (Chernoff upper bound) Suppose that $\mathbb{E}[e^{sX}] < \infty$ for some $s > 0$, and that $a > 0$. Then,

$$\mathbb{P}(S_n \geq na) \leq e^{-n\phi(a)},$$

where

$$\phi(a) = \sup_{s \geq 0} (sa - \log M(s)).$$

For $s = 0$, we have

$$sa - \log M(s) = 0 - \log 1 = 0,$$

where we have used the generic property $M(0) = 1$. Furthermore,

$$\frac{d}{ds}(sa - \log M(s))_{s=0} = a - \frac{1}{M(s)} \cdot \frac{d}{ds}M(s)_{s=0} = a - 1 \cdot \mathbb{E}[X] > 0.$$

Since the function $sa - \log M(s)$ is zero and has a positive derivative at 0, it must be positive when s is positive and small. It follows that the supremum $\phi(a)$ of the function $sa - \log M(s)$ over all $s \geq 0$ is also positive. In particular, for any fixed $a > 0$, the probability $\mathbb{P}(S_n \geq na)$ decays at least exponentially fast with n .

Example: For a standard normal random variable X , we have $M(s) = e^{s^2/2}$. Therefore, $sa - \log M(s) = sa - s^2/2$. To maximize this expression over all $s \geq 0$, we form the derivative, which is $-s$, and set it to zero, resulting in $s = a$. Thus, $\phi(a) = a^2/2$, which leads to the bound

$$\mathbb{P}(X \geq na) \leq e^{-a^2 n/2}.$$

2.2 Lower bound

Remarkably, it turns out that the estimate $\phi(a)$ of the decay rate is tight, under minimal assumptions. To keep the argument simple, we introduce some simplifying assumptions.

Assumption 1.

- (i) $M(s) = \mathbb{E}[e^{sX}] < \infty$, for all $s \in \mathbb{R}$.
- (ii) The random variable X is continuous, with PDF f_X .
- (iii) The random variable X does not admit finite upper and lower bounds.
(Formally, $0 < F_X(x) < 1$, for all $x \in \mathbb{R}$.)

We then have the following lower bound.

Theorem 2. (Chernoff lower bound) Under Assumption 1, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na) = -\phi(a), \quad (20)$$

for every $a > 0$.

We note two consequences of our assumptions, whose proof is left as an exercise:

- (a) $\lim_{s \rightarrow \infty} \frac{\log M(s)}{s} = \infty$;
- (b) $M(s)$ is differentiable at every $s > 0$.

The first property guarantees that for any $a > 0$ we have $\lim_{s \rightarrow \infty} (\log M(s) - sa) = \infty$. Since $M(s) > 0$ for all s , and since $M(s)$ is differentiable, it follows that $\log M(s)$ is also differentiable and that there exists some $s^* \geq 0$ at which $\log M(s) - sa$ is minimized over all $s \geq 0$. Taking derivatives, we see that such a s^* satisfies $a = M'(s^*)/M(s^*)$, where M' stands for the derivative of M . In particular,

$$\phi(a) = s^*a - \log M(s^*). \quad (21)$$

Let us introduce a new PDF

$$f_Y(x) = \frac{e^{s^*x}}{M(s^*)} f_X(x).$$

This is a legitimate PDF because

$$\int f_Y(x) dx = \frac{1}{M(s^*)} \int e^{s^*x} f_X(x) dx = \frac{1}{M(s^*)} \cdot M(s^*) = 1.$$

The moment generating function associated with the new PDF is

$$M_Y(s) = \frac{1}{M(s^*)} \int e^{sx} e^{s^*x} f_X(x) dx = \frac{M(s + s^*)}{M(s^*)}.$$

Thus,

$$\mathbb{E}[Y] = \frac{1}{M(s^*)} \cdot \frac{d}{ds} M(s + s^*) \Big|_{s=0} = \frac{M'(s^*)}{M(s^*)} = a,$$

where the last equality follows from our definition of M . The distribution of Y is called a “tilted” version of the distribution of X .

Let Y_1, \dots, Y_n be i.i.d. random variables with PDF f_Y . Because of the close relation between f_X and f_Y , approximate probabilities of events involving Y_1, \dots, Y_n can be used to obtain approximate probabilities of events involving X_1, \dots, X_n .

We keep assuming that $a > 0$, and fix some $\delta > 0$. Let

$$B = \left\{ (x_1, \dots, x_n) \mid a - \delta \leq \frac{1}{n} \sum_{i=1}^n x_i \leq a + \delta \right\} \subset \mathbb{R}^n.$$

Let $S_n = X_1 + \dots + X_n$ and $T_n = Y_1 + \dots + Y_n$. We have

$$\begin{aligned}
\mathbb{P}(S_n \geq n(a - \delta)) &\geq \mathbb{P}(n(a - \delta) \leq S_n \leq n(a + \delta)) \\
&= \int_{(x_1, \dots, x_n) \in B} f_X(x_1) \cdots f_X(x_n) dx_1 \cdots dx_n \\
&= \int_{(x_1, \dots, x_n) \in B} (M(s^*))^n e^{-s^* x_1} f_Y(x_1) \cdots e^{-s^* x_n} f_Y(x_n) dx_1 \cdots dx_n \\
&\geq (M(s^*))^n e^{-ns^*(a+\delta)} \int_{(x_1, \dots, x_n) \in B} f_Y(x_1) \cdots f_Y(x_n) dx_1 \cdots dx_n \\
&= (M(s^*))^n e^{-ns^*(a+\delta)} \mathbb{P}(T_n \in B).
\end{aligned} \tag{22}$$

The second inequality above was obtained because for every $(x_1, \dots, x_n) \in B$, we have $x_1 + \dots + x_n \leq n(a + \delta)$, so that $e^{-s^* x_1} \cdots e^{-s^* x_n} \geq e^{-ns^*(a+\delta)}$.

By the weak law of large numbers, we have

$$\mathbb{P}(T_n \in B) = \mathbb{P}\left(\frac{Y_1 + \dots + Y_n}{n} \in [na - n\delta, na + n\delta]\right) \rightarrow 1,$$

as $n \rightarrow \infty$. Taking logarithms, dividing by, and then taking the limit of the two sides of Eq. (22), and finally using Eq. (21), we obtain

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na - n\delta) \geq \log M(s^*) - s^* a - s^* \delta = -\phi(a) - s^* \delta.$$

This inequality is true for every $a > 0$ and $\delta > 0$. By replacing a with $a + \delta$, we have

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na) \geq -\phi(a + \delta) - s^* \delta.$$

The proof of the lower bound in Eq. (20) is completed by verifying that the function ϕ is continuous (the proof is omitted and is left as an exercise) and letting $\delta \downarrow 0$.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Lecture 19

Fall 2018

Uniform integrability, convergence of series

Contents

1. L_1 convergence (aka convergence in mean LLN).
2. Uniform integrability
3. Convergence of series of independent summands

1 CONVERGENCE IN L_1

Definition 1(Convergence in mean) A sequence of integrable random variables X_j is said to converge in L_1 to X (also known as “convergence in mean”), denoted $X_j \xrightarrow{L_1} X$ if

$$\mathbb{E}[|X_j - X|] \rightarrow 0 \quad j \rightarrow \infty.$$

For $p > 1$ we define $X_j \xrightarrow{L_p} X$ if $\mathbb{E}[|X_j|^p] < \infty$ and $\mathbb{E}[|X_j - X|^p] \rightarrow 0$.

Some simple properties are given below:

Proposition 1. (i) $X_n \xrightarrow{L_1} X$ implies $\mathbb{E}[|X|] < \infty$.

(ii) $X_j \xrightarrow{L_1} X$ implies $X_j \xrightarrow{\text{i.p.}} X$

(iii) $X_j \xrightarrow{L_1} X$ does not imply and is not implied by $X_j \xrightarrow{\text{a.s.}} X$.

(iv) The space of integrable random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ modulo almost-sure equivalence is a Banach space, denoted as $L_1(\Omega, \mathcal{F}, \mathbb{P})$ with norm $\|X\|_1 \triangleq \mathbb{E}[|X|]$. Similarly for $L_p(\Omega, \mathcal{F}, \mathbb{P})$.

Proof: (i) Follows from taking expectation in the triangle inequality:

$$|X| \leq |X_n - X| + |X_n| \quad (1)$$

(ii) and (iii) is an exercise. (iv) is outside the scope of this class. \square

Our goal is to show the following the following (third!) variation of the LLN:

Proposition 2(L_1 -LLN). *Let X_j be iid random variables with finite expectation, then*

$$\frac{1}{n} \sum_{j=1}^n X_j \xrightarrow{L_1} \mathbb{E}[X]$$

The proof of this proposition follows by Theorem 1 and Corollary 1 below.

2 UNIFORM INTEGRABILITY

Definition 2. *A collection of random variables $X_\alpha, \alpha \in S$ is uniformly integrable if*

$$k(b) \triangleq \sup_{\alpha} \mathbb{E}[|X_\alpha| \mathbf{1}\{|X_\alpha| > b\}] \rightarrow 0 \quad b \rightarrow \infty. \quad (2)$$

Some useful criteria for checking u.i.:

Proposition 3. *The following hold:*

(i) *If $\mathbb{E}[|X|] < \infty$ then $\{X\}$ is u.i.*

(ii) *When $X_\alpha \stackrel{d}{=} Y_\alpha$ then $\{X_\alpha\}$ -u.i. iff $\{Y_\alpha\}$ -u.i.*

(iii) *$\{X_\alpha\}$ -u.i. iff $\{X_\alpha\}$ is L_1 -bounded and uniformly continuous:*

$$\sup_{\alpha} \mathbb{E}[|X_\alpha|] < \infty \quad (3)$$

$$\sup_{\alpha} \mathbb{E}[|X_\alpha| \mathbf{1}_E] \rightarrow 0 \quad \text{as } \mathbb{P}[E] \rightarrow 0 \quad (4)$$

(iv) *If $G : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is such that $\frac{G(t)}{t} \rightarrow \infty$ as t grows without bound¹ then*

$$\sup_{\alpha} \mathbb{E}[G(|X_\alpha|)] < \infty \quad \Rightarrow \quad \{X_\alpha\}-\text{u.i.}$$

¹Some typical choices are $G(t) = t^2$, $|t|^{1+\epsilon}$ and $|t \log t|$.

(v) If $X_n \xrightarrow{\text{a.s.}} X$ and $\mathbb{E}[|X_n|] \rightarrow \mathbb{E}[|X|]$ then $\{X_n\}$ is u.i.

Proof: (i) follows from the MCT, (ii) is obvious from the definition, (v) is part of the homework.

For (iii), first notice that $\mathbb{E}[|X_\alpha|] \leq k(b) + b$ for every $b > 0$ and thus (3) holds. Similarly, notice that for every

$$\mathbb{E}[|X_\alpha|1_E] \leq \mathbb{E}[|X_\alpha|1\{|X_\alpha| > b\}] + b\mathbb{P}[E] \leq k(b) + b\mathbb{P}[E]$$

and thus by taking $\mathbb{P}[E] \rightarrow 0$ and $b \rightarrow \infty$ we prove (4). Conversely, if (3) and (4) hold, but $\{X_\alpha\}$ is not uniformly integrable then for some sequence and $\epsilon_0 > 0$ we have

$$\mathbb{E}[|X_{\alpha_k}|1\{|X_{\alpha_k}| > k\}] \geq \epsilon_0 > 0 \quad (5)$$

On the other hand, by (3) and Markov inequality $\mathbb{P}[|X_{\alpha_k}| > k] \rightarrow 0$. Consequently, (5) contradicts (4).

Finally, to see (iv) just notice that for every $a > 0$ there exists $b > 0$ such that $\frac{G(t)}{t} \geq a$ for all $t > b$. Then,

$$G(|X_\alpha|) \geq a|X_\alpha|1\{|X_\alpha| > b\}$$

and taking expectation here we obtain:

$$k(b) \leq \frac{1}{a} \sup_{\alpha} \mathbb{E}[G(|X_\alpha|)]$$

from which (2) follows by taking $\rightarrow \infty$. □

As a simple consequence of the Proposition we get:

Corollary 1. Let X_j be identically distributed (not necessarily independent!) and integrable. Then collection of normalized sums $\left\{ \frac{1}{n} \sum_{j=1}^n X_j, n = 1, \dots \right\}$ is uniformly integrable.

Proof: Indeed, by Proposition 3(i) and (ii) we get $\{X_j, j = 1, \dots\}$ is uniformly integrable. Now defining $Y_n = \frac{1}{n} \sum_{j=1}^n X_j$ we have

$$\sup_n \mathbb{E}[|Y_n|] \leq \mathbb{E}[|X|] < \infty$$

and on the other hand

$$\sup_n \mathbb{E}[|Y_n|1_E] \leq \sup_n \mathbb{E}[|X_n|1_E] \rightarrow 0 \quad \text{as } \mathbb{P}[E] \rightarrow 0,$$

where the last step follows by (4) applied to $\{X_j\}$. Uniform integrability of $\{Y_j\}$ then follows from Proposition 3 (iii). \square

The main value of studying uniform integrability is the following:

Theorem 1. *We have*

$$X_n \xrightarrow{L_1} X \iff X_n \xrightarrow{\text{i.p.}} X \text{ and } \{X_n\} \text{- u.i.}$$

Proof: The \Rightarrow direction follows from Markov's inequality and Proposition 3(iii). Indeed, by (1) we have the inequality

$$\mathbb{E}[|X_n|1_E] \leq \mathbb{E}[|X|1_E] + \mathbb{E}[|X_n - X|1_E] \quad (6)$$

For very large $n \geq n_0$ the second term is smaller than and hence

$$\lim_{\mathbb{P}[E] \rightarrow 0} \sup_n \left(\mathbb{E}[|X_n|1_E] \right) \leq \epsilon + \lim_{\mathbb{P}[E] \rightarrow 0} \sup_n \left(\mathbb{E}[|X|1_E] + \max_{1 \leq n \leq n_0} \mathbb{E}[|X_n - X|1_E] \right)$$

where the second term is zero by (4) because $|X_1 - X|, |X_2 - X|, \dots, |X_{n_0} - X|$ is a uniformly integrable collection. Consequently, taking $\mathbb{P}[E] \rightarrow 0$ we have shown

$$\sup_n \mathbb{E}[X_n 1_E] \rightarrow 0 \quad \mathbb{P}[E] \rightarrow 0$$

Setting $E = \Omega$ in (6) we verify (4). Thus Proposition 3(iii) implies that the infinite collection $\{X_n\}$ is also u.i.

For the converse direction, we first notice that by characterization of convergence in probability there must exist a subsequence $X_{n_k} \xrightarrow{\text{a.s.}} X$. Then by Fatou's lemma and (3) we have

$$\mathbb{E}[|X|] = \mathbb{E}[\liminf_k |X_{n_k}|] \leq \liminf_k \mathbb{E}[|X_{n_k}|] < \infty$$

Thus, the limit random variable is integrable and consequently (Exercise!) collection of nonnegative random variables $\{Y_n, n = 1, \dots\}$ is u.i. and $Y_n \xrightarrow{\text{i.p.}} 0$, where

$$Y_n \triangleq |X_n - X|.$$

Then, we have for every $\epsilon > 0$

$$\mathbb{E}[Y_n] = \mathbb{E}[Y_n 1\{Y_n > \epsilon\}] + \mathbb{E}[Y_n 1\{Y_n \leq \epsilon\}] \quad (7)$$

$$\leq \epsilon + \mathbb{E}[Y_n 1\{Y_n > \epsilon\}]. \quad (8)$$

Since $Y_n \xrightarrow{\text{i.p.}} 0$ we have $\mathbb{P}[Y_n > \epsilon] \rightarrow 0$. Then by (4) the second term converges to zero as $n \rightarrow \infty$. Hence, for all $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \sup \mathbb{E}[Y_n] \leq \epsilon,$$

which shows $Y_n \xrightarrow{L_1} 0$. \square

As a corollary we obtain a result we assumed before (in proving that convergence of characteristic functions implies convergence in distribution).

Corollary 2. *Let $f_n(x) \rightarrow f(x) \quad \forall x \in \mathbb{R}$ be a pointwise convergent sequence of pdfs. If $X_n \sim f_n$ and $X \sim f$ then $X_n \xrightarrow{d} X$.*

Proof. Let $\phi(x) = \frac{1}{2}f(x) + \sum_{n=1}^{\infty} 2^{-n-1}f_n(x)$. It is clear that $\phi(x)$ is another pdf. Let $(\phi, \mathcal{F}, \mathbb{P})$ be defined as $\Omega = \mathbb{R}, \mathcal{F} = \mathcal{B}, \mathbb{P}(dx) = \phi(x)dx$ and define random variables $Y_n(x) = f_n(x)/\phi(x)$. Note that as a consequence of our definition of ϕ this ratio is well-defined almost everywhere $\phi(x) = 0$ implies $f_n(x) = 0$ and $\mathbb{P}(\{x : \phi(x) = 0\}) = 0$. Similarly, define $Y(x) = f(x)/\phi(x)$. We have $Y_n \xrightarrow{\text{a.s.}} Y$.

Furthermore, by construction $\mathbb{E}[|Y_n|] = 1 = \mathbb{E}[|Y|]$. Thus, by Prop. 3(v) the collection $\{Y_n\}$ is u.i.. Consequently, from the previous Theorem we have $\mathbb{E}[|Y_n - Y|] \rightarrow 0$. Rewriting this last statement explicitly, we have shown

$$\int_{\mathbb{R}} |f_n(x) - f(x)| dx \rightarrow 0 \quad n \rightarrow \infty. \quad (9)$$

In particular, for any $E \in \mathcal{B}$ we have

$$P_{X_n}[E] = \int_E f_n(x) dx \rightarrow \int_E f(x) dx = P_X[E],$$

and taking $E = (-\infty, a]$ shows convergence of CDFs of X_n to the CDF of X at every point.

(In fact, (9) is usually stated as “distribution of X_n converges to the distribution of X in total-variation”. This is a stronger mode of convergence than convergence in distribution.) \square

3 SUMS OF INDEPENDENT RANDOM VARIABLES

A classical topic tightly related to the SLLN is convergence of sums

$$S_n = \sum_{j=1}^n X_j$$

when X_j are independent (and not identically distributed, of course). There is a great deal results about properties of S_n , and here we will only mention a core principle: *Convergence behavior of S_n , its central moments, concentration properties, etc are largely encoded in the behavior of $\sum_{j=1}^n \text{var}[X_j]$.*

We start with an example. Consider two independent sequences:

$$\mathbb{P}[X_n = \pm 1] = \frac{1}{2n}, \quad \mathbb{P}[X_n = 0] = -\frac{1}{n}, \quad (10)$$

$$\mathbb{P}[Y_n = \pm \frac{1}{n}] = \frac{1}{2}. \quad (11)$$

First, we notice that $\mathbb{E}[|X_n|] = \mathbb{E}[|Y_n|] = \frac{1}{2n}$, so that sum of first moments diverges at the same speed. Furthermore, both series $\sum X_n$ and $\sum Y_n$ do not absolutely converge:

$$\mathbb{P}\left[\sum_n |X_n| = +\infty\right] = \mathbb{P}\left[\sum_n |Y_n| = +\infty\right] = 1.$$

Indeed, for Y_n this is obvious as $|Y_n| = \frac{1}{n}$, while for X_n it follows from Borel Cantelli that: $\mathbb{P}[|X_n| = 1 \text{ i.o.}] = 1$.

So far, we see that $\sum X_n$ and $\sum Y_n$ behave quite similarly. However, as we will see next it turns out that

$$\mathbb{P}\left[\sum_n X_n \text{ converges}\right] = 0 \quad \sum_n \text{var}[X_n] = +\infty \quad (12)$$

$$\mathbb{P}\left[\sum_n Y_n \text{ converges}\right] = 1 \quad \sum_n \text{var}[Y_n] < +\infty \quad (13)$$

The explanation of this phenomena is the following: While both series diverge absolutely, the rapidly decreasing variances of terms allows for “sign cancellation” effect to kick in making the series Y_n converge (similar to convergence of $\sum_n \frac{(-1)^n}{n}$).

Theorem 2(Kolmogorov, Khintchine) *Let X_j be independent and*

$$\mu \triangleq \sum_{j=1}^{\infty} \mathbb{E}[X_j], \quad |\mu| < \infty, \quad (14)$$

$$\sigma^2 \triangleq \sum_{j=1}^{\infty} \text{var}[X_j] < \infty \quad (15)$$

then

$$S_n = \sum_{j=1}^n X_j$$

converges almost surely and in L_2 to a limit S with $\mathbb{E}[S] = \mu$, $\text{var}[S] = \sigma^2$.

Conversely, if $|X_j| \leq c$ for some constant c and $S_n \xrightarrow{\text{a.s.}} S$ with real-valued S , then conditions (14)-(15) hold.

Proof: We prove the direct part. Without loss of generality we assume $\mathbb{E}[X_j] = 0$. As we have shown in the homework (Cauchy criterion of almost sure convergence) it is sufficient to show that

$$\mathbb{P}[\sup_{k \geq 1} |S_{n+k} - S_n| > \epsilon] \rightarrow 0 \quad n \rightarrow \infty \quad (16)$$

Kolmogorov's inequality (see Theorem following the proof) shows that

$$\mathbb{E}[\sup_{k \geq 1} |S_{n+k} - S_n|^2] \leq 2 \sum_n \text{var}[X_j]. \quad (17)$$

By Chebyshev's inequality we obtain then

$$\mathbb{P}[\sup_{k \geq 1} |S_{n+k} - S_n| > \epsilon] \leq \frac{2}{\epsilon^2} \sum_n \text{var}[X_j] \quad (18)$$

Since sum of variances converges, the left-hand side of (18) decreases to 0 as $n \rightarrow \infty$ and thus (16) is shown. The proof of $S \xrightarrow{\text{a.s.}} S$ is complete.

Notice that (17) with $\epsilon = 0$ shows that “life-time maximum”

$$M_\infty \triangleq \sup_{n \geq 1} |S_n|$$

has finite second moment. Since

$$|S_n - S| \leq 2M_\infty,$$

by the DCT it follows that

$$\mathbb{E}[|S_n - S|^2] \rightarrow 0$$

and similarly for $\mathbb{E}[S_n] \rightarrow \mathbb{E}[S]$, $\mathbb{E}[S_n^2] \rightarrow \mathbb{E}[S]$.

We proceed to proving the converse. First, assume $\mathbb{E}[X_j] = 0$ and suppose $S_n \xrightarrow{\text{a.s.}} S$ but

$$D_n = \sum_{j=1}^n \text{var}[X_j] \nearrow \infty$$

Then, notice that

$$\mathbb{E}[|X_j|^3] \leq \mathbb{E}[|X_j|^2]c$$

and thus

$$\sum_{j=1}^n \mathbb{E}[|X_j|^3] \leq D_n c$$

Consequently, by the CLT for non-identically distributed random variables we have

$$\frac{1}{\sqrt{D_n}} S_n \xrightarrow{\text{d}} Z \sim \mathcal{N}(0, 1).$$

On the other hand, we have for every $s > 0$ and for all large enough

$$\mathbb{P}[S_n > t] \geq \mathbb{P}[S_n > s\sqrt{D_n}]. \quad (19)$$

Since $S_n \xrightarrow{\text{d}} S$ we also have for all such that $\mathbb{P}[S = t] = 0$ that

$$\mathbb{P}[S_n > t] \rightarrow \mathbb{P}[S > t].$$

However, upon taking the limit in (19) as $n \rightarrow \infty$ we get for all $s > 0$:

$$\mathbb{P}[S > t] \geq \mathbb{P}[Z > s]$$

Taking $s \rightarrow 0$ we get

$$\mathbb{P}[S > t] \geq \frac{1}{2} \quad \forall t \in \mathbb{R}$$

which is a contradiction, as no distribution can satisfy such inequality.

Next, if $\mathbb{E}[X_j] = \mu_j$, then let $Y_j = X_j - \bar{X}'_j$, where \bar{X}'_j is an independent copy of X_j . In this way $\mathbb{E}[Y_j] = 0$, $\text{var}[Y_j] = 2\text{var}[X_j]$ and

$$\sum_{j=1}^n Y_j = S_n - S'_n \xrightarrow{\text{a.s.}} S - S'.$$

Hence, by the previous argument we have

$$\sum_{j=1}^n \text{var}[X_j] < \infty$$

and by the direct part of the theorem

$$\sum_{j=1}^n (X_j - \mu_j) = S_n - \sum_{j=1}^n \mu_j$$

converges almost surely. Since S_n converges by assumption, so must $\sum \mu_j$. \square

Remark: Conditions (14) are necessary for convergence $S_n \xrightarrow{\text{a.s.}} S$ only under assumption of the boundedness X_j (see homework). We also mention that instead of relying on the CLT in the proof of the converse direction, we may have followed a more conventional route, based on the inequality

$$\mathbb{P}\left[\max_{1 \leq k \leq n} |S_k| > a\right] \geq 1 - \frac{(a+c)^2}{\text{var}[S_n]}.$$

The proof of this inequality, however, would appear rather unnatural without mentioning stopping times. Either method, however, really just shows that condition $|X_j| \leq c$ guarantees the width of the distribution S_n has the same order as $\sqrt{\text{var}[S_n]}$. (For unbounded X_j , rare large jumps may significantly increase the variance, while having very little effect on the bulk of the distribution of

Theorem 3(Kolmogorov). Let X_j be independent, zero-mean with finite second moments and let

$$M_n = \sup_{1 \leq k \leq n} \sum_{j=1}^k X_j, \quad 1 \leq n \leq \infty.$$

Then we have for any $1 \leq n \leq \infty$

$$\mathbb{E}[|M_n|^2] \leq 2 \sum_{j=1}^n \mathbb{E}[X_j^2].$$

Proof: The case of $n = \infty$ follows from the case of finite by the MCT. Let

$$S_k = X_1 + \cdots + X_k, \tag{20}$$

$$A_n = \max_{1 \leq k \leq n} S_k, \tag{21}$$

Note that for $n = 1$ we clearly have

$$\mathbb{E}[A_n^2] \leq \mathbb{E}[S_n^2]. \quad (22)$$

Assume (by induction) that (22) is shown for all sums of n random variables. Then, notice that

$$A_n = X_1 + \max(0, X_2, X_2 + X_3, \dots, \sum_2^n X_j).$$

Since first and second terms are independent $\mathbb{E}[X_1] = 0$ we get

$$\mathbb{E}[A_n^2] = \mathbb{E}[X_1^2] + \mathbb{E}[\max(0, X_2, X_2 + X_3, \dots, \sum_2^n X_j)^2] \quad (23)$$

$$\leq \mathbb{E}[X_1^2] + \mathbb{E}[\max(X_2, X_2 + X_3, \dots, \sum_2^n X_j)^2] \quad (24)$$

$$\leq \sum_{j=1}^n \mathbb{E}[X_j^2], \quad (25)$$

where the first inequality follows from $x^2 \leq x^2$ and the second one is by the inductive assumption. Thus (22) holds for all

By symmetry, we also must have

$$\mathbb{E}[B_n^2] \leq \mathbb{E}[S_n^2], \quad B_n = \max_{1 \leq k \leq n} -S_k.$$

Finally, since $M_n^2 = \max(A_n^2, B_n^2)$ and using $\max(a, b) \leq a + b$ we complete the proof. \square

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

THE BASICS OF STOCHASTIC PROCESSES

Contents

- 1. Stochastic processes: spaces \mathbb{R}^∞ and $\mathbb{R}^{[0,\infty)}$
- 2. The Bernoulli process
- 3. The Poisson process

We now turn to the study of some simple classes of stochastic processes. Examples and a more leisurely discussion of this material can be found in the corresponding chapter of [BT].

A discrete-time stochastic is a sequence of random variables $\{X_n\}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In more detail, a stochastic process is a function X of two variables n and ω . For every n , the function $\omega \mapsto X_n(\omega)$ is a random variable (a measurable function). An alternative perspective is provided by fixing some $\omega \in \Omega$ and viewing $X_n(\omega)$ as a function of n (a “time function,” or “sample path,” or “trajectory”).

A continuous-time stochastic process is defined similarly, as a collection of random variables $\{X_t\}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where t varies over non-negative real values \mathbb{R}_+ .

1 SPACES OF TRAJECTORIES: \mathbb{R}^∞ and $\mathbb{R}^{[0,\infty)}$

1.1 σ -algebras on spaces of trajectories

Recall that earlier we defined the Borel σ -algebra \mathcal{B}^n on \mathbb{R}^n as the smallest σ -algebra containing all measurable rectangles, i.e. events of the form

$$B_1 \times \cdots \times B_n = \{\mathbf{x} \in \mathbb{R}^n : x_j \in B_j \quad \forall j \in [n]\}$$

where B_j are (1-dimensional) Borel subsets of \mathbb{R} . A generalization is the following:

Definition 1. Let T be an arbitrary set of indices. The product space \mathbb{R}^T is defined as

$$\mathbb{R}^T \triangleq \prod_{t \in T} \mathbb{R} = \{(x_t, t \in T)\}.$$

A subset $\mathcal{J}_S(B)$ of \mathbb{R}^T is called a cylinder with base B on time indices $S = \{s_1, \dots, s_n\}$ if

$$\mathcal{J}_S(B) = \{(x_t) : (x_{s_1}, \dots, x_{s_n}) \in B\}, \quad B \subset \mathbb{R}^n, \quad (1)$$

with $B \in \mathcal{B}^n$. The product σ -algebra \mathcal{B}^T is the smallest σ -algebra containing all cylinders:

$$\mathcal{B}^T = \sigma\{\mathcal{J}_S(B) : \forall S\text{-finite and } B \in \mathcal{B}^S\}.$$

For the special case $T = \{1, 2, \dots\}$ the notation \mathbb{R}^∞ and \mathcal{B}^∞ will be used.

The following are measurable subsets of \mathbb{R}^∞ :

$$E_0 = \{x \in \mathbb{R}^\infty : x_n\text{-converges}\}$$

The following are measurable subsets of $\mathbb{R}^{[0, \infty)}$:

$$E_1 = \{x \in \mathbb{R}^{[0, \infty)} : x_t = 0 \quad \forall t \in \mathbb{Q}\} \quad (2)$$

$$E_2 = \{x \in \mathbb{R}^{[0, \infty)} : \sup_{t \in \mathbb{Q}} x_t > 0\} \quad (3)$$

The following are not measurable subsets of $\mathbb{R}^{[0, \infty)}$:

$$E'_1 = \{x \in \mathbb{R}^{[0, \infty)} : x_t = 0 \quad \forall t\} \quad (4)$$

$$E'_2 = \{x \in \mathbb{R}^{[0, \infty)} : \sup_t x_t > 0\} \quad (5)$$

$$E_3 = \{x \in \mathbb{R}^{[0, \infty)} : x_t\text{-continuous}\} \quad (6)$$

Non-measurability of E'_1 and E'_2 will follow from the next result. We mention that since $E_1 \cap E_3 = E'_1 \cap E_3$, then by considering a trace of $\mathcal{B}^{[0, +\infty)}$ on E_3 sets E'_1 and E'_2 can be made measurable. This is a typical approach taken in the theory of continuous stochastic processes.

Proposition 1. The following provides information about \mathcal{B}^T :

- (i) For every measurable set $E \in \mathcal{B}^T$ there exists a countable set of time indices $S = \{s_1, \dots\}$ and a subset $B \in \mathcal{B}^\infty$ such that

$$E = \{(x_t) : (x_{s_1}, \dots, x_{s_n}, \dots) \in B\} \quad (7)$$

- (ii) Every measurable set $E \in \mathcal{B}^T$ can be approximated within arbitrary ϵ by a cylinder:

$$\mathbb{P}[E \Delta \mathcal{J}_S(B)] \leq \epsilon,$$

where \mathbb{P} is any probability measure on $(\mathbb{R}^T, \mathcal{B}^T)$.

- (iii) If $\{X_t, t \in T\}$ is a collection of random variables on (Ω, \mathcal{F}) , then the map

$$X : \Omega \rightarrow \mathbb{R}^T, \quad (8)$$

$$\omega \mapsto (X_t(\omega), t \in T) \quad (9)$$

is measurable with respect to \mathcal{B}^T .

Proof: For (i) simply notice that collection of sets of the form (7) contains all cylinders and closed under countable unions/intersections. To see this simply notice that one can without loss of generality assume that every set in, for example, union $F = \bigcup E_n$ correspond to the same set of indices in (7) (otherwise extend the index sets S first).

(ii) follows from the next exercise and the fact that $\{\mathcal{J}_S(B), B \in \mathcal{B}^S\}$ (under fixed finite S) form a σ -algebra. For (iii) note that it is sufficient to check that $X^{-1}(\mathcal{J}_S(B)) \in \mathcal{F}$ (since cylinders generate \mathcal{B}^T). The latter follows at once from the definition of a cylinder (1) and the fact that

$$\{(X_{s_1}, \dots, X_{s_n}) \in B\}$$

are clearly in \mathcal{F} . □

Exercise 1. Let $\mathcal{F}_\alpha, \alpha \in S$ be a collection of σ -algebras and let $\mathcal{F} = \bigvee_{\alpha \in S} \mathcal{F}_\alpha$ be the smallest σ -algebra containing all of them. Call set B finitary if $B \in \bigvee_{\alpha \in S_1} \mathcal{F}_\alpha$, where S_1 is a finite subset of S . Prove that every $E \in \mathcal{F}$ is finitary approximable, i.e. that for every $\epsilon > 0$ there exists a finitary B such that

$$\mathbb{P}[E \Delta B] \leq \epsilon.$$

(Hint: Let $\mathcal{L} = \{E : E\text{-finitary approximable}\}$ and show that \mathcal{L} contains the algebra of finitary sets and closed under monotone limits.)

With these preparations we are ready to give a definition of stochastic process:

Definition 2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A stochastic process with time set T is a measurable map $X : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}^T, \mathcal{B}^T)$. The pushforward $\mathbb{P}_X \triangleq \mathbb{P} \circ X^{-1}$ is called the law of X .

1.2 Probability measures on spaces of trajectories

According to Proposition 1 we may define probability measures on \mathbb{R}^T by simply computing an induced measure along a map (9). An alternative way to define probabilities on \mathbb{R}^T is via the following construction.

Theorem 1 (Kolmogorov). *Suppose that for any finite $S \subset T$ we have a probability measure \mathbb{P}_S on \mathbb{R}^S and that these measures are consistent. Namely, if $S' \subset S$ then*

$$\mathbb{P}_{S'}[B] = \mathbb{P}_S[B \times \mathbb{R}^{S \setminus S'}].$$

Then there exists a unique probability measure \mathbb{P} on \mathbb{R}^T such that

$$\mathbb{P}[\mathcal{J}_S(B)] = \mathbb{P}_S[B]$$

for every cylinder $\mathcal{J}_S(B)$.

Proof (optional): As a simple exercise, reader is encouraged to show that it suffices to consider the case of countable T (cf. Proposition 1.(i)). We thus focus on constructing a measure on \mathbb{R}^∞ . Let $\mathcal{A} = \bigcup_{n \geq 1} \mathcal{F}_n$, where \mathcal{F}_n is the σ -algebra of all cylinders with time indices $\{1, \dots, n\}$. Clearly \mathcal{A} is an algebra. Define a set-function on \mathcal{A} via:

$$\forall E = \{(x_1, \dots, x_n) \in B\} : \quad \mathbb{P}[E] \triangleq \mathbb{P}_{\{1, \dots, n\}}[B].$$

Consistency conditions guarantee that this assignment is well-defined and results in a finitely additive set-function. We need to verify countable additivity. Let

$$E_n \searrow \emptyset \tag{10}$$

By repeating the sets as needed, we may assume $E_n \in \mathcal{F}_n$. If we can show that

$$\mathbb{P}[E_n] \searrow 0 \tag{11}$$

then Caratheodory's extension theorem guarantees that \mathbb{P} extends uniquely to $\sigma(\mathcal{A}) = \mathcal{B}^\infty$.

We will use the following facts about \mathbb{R}^n :

1. Every finite measure μ on $(\mathbb{R}^n, \mathcal{B}^n)$ is *inner regular*, namely for every $E \in \mathcal{B}^n$

$$\mu[E] = \sup_{K \subset E} \mu[K], \tag{12}$$

supremum over all compact subsets of E .

2. Every decreasing sequence of non-empty compact sets has non-empty intersection:

$$K_n \neq \emptyset, K_n \searrow K \Rightarrow K \neq \emptyset \quad (13)$$

3. If $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$ is continuous, then $f(K)$ is compact for every compact K .

Then according to (12) for every E_n and every $\epsilon > 0$ there exists a compact subset $K'_n \subset \mathbb{R}^n$ such that such that

$$\mathbb{P}[E_n \setminus \mathcal{J}_{1,\dots,n}(K'_n)] \leq \epsilon 2^{-n}.$$

Then, define by induction

$$K_n = K'_n \cap (K_{n-1} \times \mathbb{R}).$$

(Note that $K_{n-1} \subset \mathbb{R}^{n-1}$ and the set $K_{n-1} \times \mathbb{R}$ is simply an extension of K_{n-1} into \mathbb{R}^n by allowing arbitrary last coordinates.) Since $E_n \subset E_{n-1}$ we have

$$\mathbb{P}[E_n \setminus \mathcal{J}_{1,\dots,n}(K_n)] \leq \epsilon 2^{-n} + \mathbb{P}[E_{n-1} \setminus \mathcal{J}_{1,\dots,n-1}(K_{n-1})].$$

Thus, continuing by induction we have shown that

$$\mathbb{P}[E_n \setminus \mathcal{J}_{1,\dots,n}(K_n)] \leq \epsilon(2^{-1} + \dots + 2^{-n}) < \epsilon \quad (14)$$

We will show next that $K_n = \emptyset$ for all n large enough. Since by construction

$$E_n \supset \mathcal{J}_{1,\dots,n}(K_n) \quad (15)$$

we then have from (14) and $K_n = \emptyset$ that

$$\limsup_{n \rightarrow \infty} \mathbb{P}[E_n] < \epsilon.$$

By taking ϵ to 0 we have shown (11) and the Theorem.

It thus remains to show that $K_n = \emptyset$ for all large enough n . Suppose otherwise, then by construction we have

$$K_n \subset K_{n-1} \times \mathbb{R} \subset K_{n-2} \times \mathbb{R}^2 \subset \dots \subset K_1 \times \mathbb{R}^{n-1}.$$

Thus by projecting each K_n onto first coordinate we get a decreasing sequence of non-empty compacts, which by (13) has non-empty intersection. Then we can pick a point $x_1 \in \mathbb{R}$ such that

$$x_1 \in \text{Proj}_{n \rightarrow 1}(K_n) \quad \forall n.$$

Repeating the same argument but projecting onto first two coordinates, we can now pick $x_2 \in \mathbb{R}$ such that

$$(x_1, x_2) \in \text{Proj}_{n \rightarrow 2}(K_n) \quad \forall n.$$

By continuing in this fashion we will have constructed the sequence

$$(x_1, x_2, \dots) \in \mathcal{J}_{1, \dots, n}(K_n) \quad \forall n.$$

By (15) then we have

$$(x_1, x_2, \dots) \in \bigcap_{n \geq 1} E_n$$

which contradicts (10). Thus, one of K_n must be empty. \square

1.3 Tail σ -algebra and Kolmogorov's 0/1 law

Definition 3. Consider $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ and let \mathcal{F}_n^∞ be a sub- σ -algebra generated by all cylinders $\mathcal{J}_{s_1, \dots, s_k}(B)$ with $s_j \geq n$. Then the σ -algebra

$$\mathcal{T} \triangleq \bigcap_{n > 0} \mathcal{F}_n^\infty$$

is called a tail σ -algebra on \mathbb{R}^∞ . If $X : \Omega \rightarrow \mathbb{R}^\infty$ is a stochastic process, then σ -algebra $X^{-1}\mathcal{T}$ is called a tail σ -algebra of X .

Examples of tail events:

$$E_1 = \{\text{sequence } X_n \text{ converges}\} \tag{16}$$

$$E_2 = \{\text{series } \sum X_n \text{ converges}\} \tag{17}$$

$$E_3 = \{\limsup_{n \rightarrow \infty} X_n > 0\}, \tag{18}$$

An example of the event which is not a tail event:

$$E_4 = \{\limsup_{n \rightarrow \infty} \sum_{k=1}^n X_k > 0\}$$

Theorem 2 (Kolmogorov's 0/1 law). If $X_j, j = 1, \dots$ are independent then any event in the tail σ -algebra of X has probability 0 or 1.

Proof: Let \mathbb{P}_X be the law of X (so that \mathbb{P}_X is a measure on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$). Take $E \in \mathcal{T}$, then $E \in \mathcal{F}_n^\infty$ for every n . Thus under \mathbb{P}_X event E is independent of every cylinder:

$$\mathbb{P}_X[E \cap \mathcal{J}_{s_1, \dots, s_k}(B)] = \mathbb{P}_X[E]\mathbb{P}_X[\mathcal{J}_{s_1, \dots, s_k}(B)] \quad (19)$$

On the other hand, by Proposition 1 every element of \mathcal{B}^∞ can be arbitrarily well approximated with cylinders. Taking a sequence of such approximations converging to E in (19) we derive that E must be independent of itself:

$$\mathbb{P}_X[E \cap E] = \mathbb{P}_X[E]\mathbb{P}_X[E],$$

implying $\mathbb{P}_X[E] = 0$ or 1 . □

2 THE BERNOULLI PROCESS

In the Bernoulli process, the random variables X_n are i.i.d. Bernoulli, with common parameter $p \in (0, 1)$. The natural sample space in this case is $\Omega = \{0, 1\}^\infty$.

Let $S_n = X_1 + \dots + X_n$ (the number of “successes” or “arrivals” in n steps). The random variable S_n is binomial, with parameters n and p , so that

$$p_{S_n}(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

$$\mathbb{E}[S_n] = np, \quad \text{var}(S_n) = np(1-p).$$

Let T_1 be the time of the first success. Formally, $T_1 = \min\{n \mid X_n = 1\}$. We already know that T_1 is geometric:

$$p_{T_1}(k) = (1-p)^{k-1}p, \quad k = 1, 2, \dots; \quad \mathbb{E}[T_1] = \frac{1}{p}.$$

2.1 Stationarity and memorylessness

The Bernoulli process has a very special structure. The discussion below is meant to capture some of its special properties in an abstract manner.

Consider a Bernoulli process $\{X_n\}$. Fix a particular positive integer m , and let $Y_n = X_{m+n}$. Then, $\{Y_n\}$ is the process seen by an observer who starts watching the process $\{X_n\}$ at time $m+1$, as opposed to time 1. Clearly, the process $\{Y_n\}$ also involves a sequence of i.i.d. Bernoulli trials, with the same parameter p . Hence, it is also a Bernoulli process, and has the same distribution as the process $\{X_n\}$. More precisely, for every k , the distribution of (Y_1, \dots, Y_k)

is the same as the distribution of (X_1, \dots, X_k) . This property is called **stationarity** property.

In fact a stronger property holds. Namely, even if we are given the values of X_1, \dots, X_m , the distribution of the process $\{Y_n\}$ does not change. Formally, for any measurable set $A \subset \mathbb{R}^m$, we have

$$\begin{aligned}\mathbb{P}((X_{n+1}, X_{n+2}, \dots) \in A \mid X_1, \dots, X_n) &= \mathbb{P}((X_{n+1}, X_{n+2}, \dots) \in A) \\ &= \mathbb{P}((X_1, X_2, \dots, \dots) \in A).\end{aligned}$$

We refer to the first equality as a **memorylessness** property. (The second inequality above is just a restatement of the stationarity property.)

2.2 Stopping times

We just discussed a situation where we start “watching” the process at some time $m + 1$, where m is an integer constant. We next consider the case where we start watching the process at some random time $N + 1$. So, let N be a nonnegative integer random variable. Is the process $\{Y_n\}$ defined by $Y_n = X_{N+n}$ a Bernoulli process with the same parameter? In general, this is not the case. For example, if $N = \min\{n \mid X_{n+1} = 1\}$, then $\mathbb{P}(Y_1 = 1) = \mathbb{P}(X_{N+1} = 1) = 1 \neq p$. This inequality is due to the fact that we chose the special time N by “looking into the future” of the process; that was determined by the future value X_{n+1} .

This motivates us to consider random variables N that are determined causally, by looking only into the past and present of the process. Formally, a nonnegative random variable N is called a **stopping time** if, for every n , the occurrence or not of the event $\{N = n\}$ is completely determined by the values of X_1, \dots, X_n . Even more formally, for every n , there exists a function h_n such that

$$I_{\{N=n\}} = h_n(X_1, \dots, X_n).$$

We are now in a position to state a stronger version of the memorylessness property. If N is a stopping time, then for all n , we have

$$\begin{aligned}\mathbb{P}((X_{N+1}, X_{N+2}, \dots) \in A \mid N = n, X_1, \dots, X_n) &= \mathbb{P}((X_{n+1}, X_{n+2}, \dots) \in A) \\ &= \mathbb{P}((X_1, X_2, \dots, \dots) \in A).\end{aligned}$$

In words, the process seen if we start watching right after a stopping time is also Bernoulli with the same parameter p .

2.3 Arrival and interarrival times

For $k \geq 1$, let Y_k be the k **th arrival time**. Formally, $Y_k = \min\{n \mid S_n = k\}$. For convenience, we define $Y_0 = 0$. The k **th interarrival time** is defined as $T_k = Y_k - Y_{k-1}$.

We already mentioned that T_1 is geometric. Note that T_1 is a stopping time, so the process $(X_{T_1+1}, X_{T_1+2}, \dots)$ is also a Bernoulli process. Note that the second interarrival time T_2 , in the original process is the first arrival time in this new process. This shows that T_2 is also geometric. Furthermore, the new process is independent from (X_1, \dots, X_{T_1}) . Thus, T_2 (a function of the new process) is independent from (X_1, \dots, X_{T_1}) . In particular, T_2 is independent from T_1 .

By repeating the above argument, we see that the interarrival times T_k are i.i.d. geometric. As a consequence, Y_k is the sum of k i.i.d. geometric random variables, and its PMF can be found by repeated convolution. In fact, a simpler derivation is possible. We have

$$\begin{aligned}\mathbb{P}(Y_k = t) &= \mathbb{P}(S_{t-1} = k-1 \text{ and } X_t = 1) = \mathbb{P}(S_{t-1} = k-1) \cdot \mathbb{P}(X_t = 1) \\ &= \binom{t-1}{k-1} p^{k-1} (1-p)^{t-k} \cdot p = \binom{t-1}{k-1} p^k (1-p)^{t-k}.\end{aligned}$$

The PMF of Y_k is called a **Pascal PMF**.

2.4 Merging and splitting of Bernoulli processes

Suppose that $\{X_n\}$ and $\{Y_n\}$ are independent Bernoulli processes with parameters p and q , respectively. Consider a “merged” process $\{Z_n\}$ which records an arrival at time n if and only if one or both of the original processes record an arrival. Formally,

$$Z_n = \max\{X_n, Y_n\}.$$

The random variables Z_n are i.i.d. Bernoulli, with parameter

$$\mathbb{P}(Z_n = 1) = 1 - \mathbb{P}(X_n = 0, Y_n = 0) = 1 - (1-p)(1-q) = p + q - pq.$$

In particular, $\{Z_n\}$ is itself a Bernoulli process.

“Splitting” is in some sense the reverse process. If there is an arrival at time n (i.e., $X_n = 1$), we flip an independent coin, with parameter q , and record an arrival of “type I” or “type II”, depending on the coin’s outcome. Let $\{X_n\}$ and $\{Y_n\}$ be the processes of arrivals of the two different types. Formally, let $\{U_n\}$ be a Bernoulli process with parameter q , independent from the original process $\{Z_n\}$. We then let

$$X_n = Z_n \cdot U_n, \quad Y_n = Z_n \cdot (1 - U_n).$$

Note that the random variables X_n are i.i.d. Bernoulli, with parameter pq , so that $\{X_n\}$ is a Bernoulli process with parameter pq . Similarly, $\{Y_n\}$ is a Bernoulli process with parameter $p(1 - q)$. Note however that the two processes are dependent. In particular, $\mathbb{P}(X_n = 1 \mid Y_n = 1) = 0 \neq pq = \mathbb{P}(X_n = 1)$.

3 THE POISSON PROCESS

The Poisson process is best understood intuitively as a continuous-time analog of the Bernoulli process. The process starts at time zero, and involves a sequence of arrivals, at random times. It is described in terms of a collection of random variables $N(t)$, for $t \geq 0$, all defined on the same probability space, where $N(0) = 0$ and $N(t)$, $t > 0$, represents the number of arrivals during the interval $(0, t]$.

If we fix a particular outcome (sample path) ω , we obtain a time function whose value at time t is the realized value of $N(t)$. This time function has discontinuities (unit jumps) whenever an arrival occurs. Furthermore, this time function is right-continuous: formally, $\lim_{\tau \downarrow t} N(\tau) = N(t)$; intuitively, the value of $N(t)$ incorporates the jump due to an arrival (if any) at time t .

We introduce some notation, analogous to the one used for the Bernoulli process:

$$Y_0 = 0, \quad Y_k = \min\{t \mid N(t) = k\}, \quad T_k = Y_k - Y_{k-1}.$$

We also let

$$P(k; t) = \mathbb{P}(N(t) = k).$$

The Poisson process, with parameter $\lambda > 0$, is defined implicitly by the following properties:

- (a) The numbers of arrivals in disjoint intervals are independent. Formally, if $0 < t_1 < t_2 < \dots < t_k$, then the random variables $N(t_1), N(t_2) - N(t_1), \dots, N(t_k) - N(t_{k-1})$ are independent. This is an analog of the independence of trials in the Bernoulli process.
- (b) The distribution of the number of arrivals during an interval is determined by λ and the length of the interval. Formally, if $t_1 < t_2$, then

$$\mathbb{P}(N(t_2) - N(t_1) = k) = \mathbb{P}(N(t_2 - t_1) = k) = P(k; t_2 - t_1).$$

- (c) There exist functions o_1, o_2, o_3 such that

$$\lim_{\delta \downarrow 0} \frac{o_k(\delta)}{\delta} = 0, \quad k = 1, 2, 3,$$

and

$$\begin{aligned} P(0; \delta) &= 1 - \lambda\delta + o_1(\delta) \\ P(1; \delta) &= \lambda\delta + o_2(\delta), \\ \sum_{k=2}^{\infty} P(k; \delta) &= o_3(\delta), \end{aligned}$$

for all $\delta > 0$.

The o_k functions are meant to capture second and higher order terms in a Taylor series approximation.

3.1 The distribution of $N(t)$

Let us fix the parameter λ of the process, as well as some time $t > 0$. We wish to derive a closed form expression for $P(k; t)$. We do this by dividing the time interval $(0, t]$ into small intervals, using the assumption that the probability of two or more arrivals in a small interval is negligible, and then approximate the process by a Bernoulli process.

Having fixed $t > 0$, let us choose a large integer n , and let $\delta = t/n$. We partition the interval $[0, t]$ into n “slots” of length δ . The probability of at least one arrival during a particular slot is

$$p = 1 - P(0; \delta) = \lambda\delta + o(\delta) = \frac{\lambda t}{n} + o(1/n),$$

for some function o that satisfies $o(\delta)/\delta \rightarrow 0$.

We fix k and define the following events:

- A : exactly k arrivals occur in $(0, t]$;
- B : exactly k slots have one or more arrivals;
- C : at least one of the slots has two or more arrivals.

The events A and B coincide unless event C occurs. We have

$$B \subset A \cup C, \quad A \subset B \cup C,$$

and, therefore,

$$\mathbb{P}(B) - \mathbb{P}(C) \leq \mathbb{P}(A) \leq \mathbb{P}(B) + \mathbb{P}(C).$$

Note that

$$\mathbb{P}(C) \leq n \cdot o_3(\delta) = (t/\delta) \cdot o_3(\delta),$$

which converges to zero, as $n \rightarrow \infty$ or, equivalently, $\delta \rightarrow 0$. Thus, $\mathbb{P}(A)$, which is the same as $P(k; t)$ is equal to the limit of $\mathbb{P}(B)$, as we let $n \rightarrow \infty$.

The number of slots that record an arrival is binomial, with parameters n and $p = \lambda t/n + o(1/n)$. Thus, using the binomial probabilities,

$$\mathbb{P}(B) = \binom{n}{k} \left(\frac{\lambda t}{n} + o(1/n) \right)^k \left(1 - \frac{\lambda t}{n} + o(1/n) \right)^{n-k}.$$

When we let $n \rightarrow \infty$, essentially the same calculation as the one carried out in Lecture 6 shows that the right-hand side converges to the Poisson PMF, and

$$P(k; t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

This establishes that $N(t)$ is a Poisson random variable with parameter λt , and $\mathbb{E}[N(t)] = \text{var}(N(t)) = \lambda t$.

3.2 The distribution of T_k

In full analogy with the Bernoulli process, we will now argue that the interarrival times T_k are i.i.d. exponential random variables.

3.2.1 First argument

We have

$$\mathbb{P}(T_1 > t) = \mathbb{P}(N(t) = 0) = P(0; t) = e^{-\lambda t}.$$

We recognize this as an exponential CDF. Thus,

$$f_{T_1}(t) = \lambda e^{-\lambda t}, \quad t > 0.$$

Let us now find the joint PDF of the first two interarrival times. We give a heuristic argument, in which we ignore the probability of two or more arrivals during a small interval and any $o(\delta)$ terms. Let $t_1 > 0$, $t_2 > 0$, and let δ be a small positive number, with $\delta < t_2$. We have

$$\begin{aligned} \mathbb{P}(t_1 \leq T_1 \leq t_1 + \delta, \quad t_2 \leq T_2 \leq t_2 + \delta) \\ \approx P(0; t_1) \cdot P(1; \delta) \cdot P(0; t_2 - t_1 - \delta) \cdot P(1; \delta) \\ = e^{-\lambda t_1} \lambda \delta e^{-\lambda(t_2 - \delta)} \lambda \delta. \end{aligned}$$

We divide both sides by δ^2 , and take the limit as $\delta \downarrow 0$, to obtain

$$f_{T_1, T_2}(t_1, t_2) = \lambda e^{-\lambda t_1} \lambda e^{-\lambda t_2}. \quad t_1, t_2 > 0.$$

This shows that T_2 is independent of T_1 , and has the same exponential distribution. This argument is easily generalized to argue that the random variables T_k are i.i.d. exponential, with common parameter λ .

3.2.2 Second argument

We will first find the joint PDF of Y_1 and Y_2 . Suppose for simplicity that $\lambda = 1$. let us fix some s and t that satisfy $0 < s \leq t$. We have

$$\begin{aligned}\mathbb{P}(Y_1 \leq s, Y_2 \leq t) &= \mathbb{P}(N(s) \geq 1, N(t) \geq 2) \\ &= \mathbb{P}(N(s) = 1)\mathbb{P}(N(t) - N(s) \geq 1) + \mathbb{P}(N(s) \geq 2) \\ &= se^{-s}(1 - e^{-(t-s)}) + (1 - e^{-s} - se^{-s}) \\ &= -se^{-t} + 1 - e^{-s}.\end{aligned}$$

Differentiating, we obtain

$$f_{Y_1, Y_2}(s, t) = \frac{\partial^2}{\partial t \partial s} \mathbb{P}(Y_1 \leq s, Y_2 \leq t) = e^{-t}, \quad 0 \leq s \leq t.$$

We point out an interesting consequence: conditioned on $Y_2 = t$, Y_1 is uniform on $(0, t)$; that is given the time of the second arrival, all possible times of the first arrival are “equally likely.”

We now use the linear relations

$$T_1 = Y_1, \quad T_2 = Y_2 - Y_1.$$

The determinant of the matrix involved in this linear transformation is equal to 1. Thus, the Jacobian formula yields

$$f_{T_1, T_2}(t_1, t_2) = f_{Y_1, Y_2}(t_1, t_1 + t_2) = e^{-t_1}e^{-t_2},$$

confirming our earlier independence conclusion. Once more this approach can be generalized to deal with more than two interarrival times, although the calculations become more complicated

3.2.3 Alternative definition of the Poisson process

The characterization of the interarrival times leads to an alternative, but equivalent, way of describing the Poisson process. Start with a sequence of independent exponential random variables T_1, T_2, \dots , with common parameter λ , and record an arrival at times $T_1, T_1 + T_2, T_1 + T_2 + T_3$, etc. It can be verified that starting with this new definition, we can derive the properties postulated in our original definition. Furthermore, this new definition, being constructive, establishes that a process with the claimed properties does indeed exist.

3.3 The distribution of Y_k

Since Y_k is the sum of k i.i.d. exponential random variables, its PDF can be found by repeating convolution.

A second, somewhat heuristic, derivation proceeds as follows. If we ignore the possibility of two arrivals during a small interval, We have

$$\mathbb{P}(y \leq Y_k \leq y + \delta) = P(k-1; y)P(1; \delta) = \frac{\lambda^{k-1}}{(k-1)!} y^{k-1} e^{-\lambda y} \lambda \delta.$$

We divide by δ , and take the limit as $\delta \downarrow 0$, to obtain

$$f_{Y_k}(y) = \frac{\lambda^{k-1}}{(k-1)!} y^{k-1} e^{-\lambda y} \lambda, \quad y > 0.$$

This is called a **Gamma** or **Erlang** distribution, with k degrees of freedom.

For an alternative derivation that does not rely on approximation arguments, note that for a given $y \geq 0$, the event $\{Y_k \leq y\}$ is the same as the event

$$\{ \text{number of arrivals in the interval } [0, y] \text{ is at least } k \}.$$

Thus, the CDF of Y_k is given by

$$F_{Y_k}(y) = \mathbb{P}(Y_k \leq y) = \sum_{n=k}^{\infty} P(n, y) = 1 - \sum_{n=0}^{k-1} P(n, y) = 1 - \sum_{n=0}^{k-1} \frac{(\lambda y)^n e^{-\lambda y}}{n!}.$$

The PDF of Y_k can be obtained by differentiating the above expression, and moving the differentiation inside the summation (this can be justified). After some straightforward calculation we obtain the Erlang PDF formula

$$f_{Y_k}(y) = \frac{d}{dy} F_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}.$$

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MARKOV CHAINS**Contents**

1. Homogeneous Finite State Markov Chain
2. Stationary Distribution
3. Recurrent and Transient States

1 INTRODUCTION

Recall a model we considered earlier: random walk. We have $X_j \stackrel{d}{=} \text{Ber}(p)$, i.i.d. Then $S_n = \sum_{j=1}^n X_j$ was defined to be a simple random walk. One of its key property is that the distribution S_{n+1} conditioned on the state $S_n = x$ at n is independent from the past history, namely, $m \leq n - 1$. To see this formally note that

$$\begin{aligned} & \mathbb{P}(S_{n+1} = y | S_n = x, S_{n-1} = z_1, \dots, S_1 = z_{n-1}) \\ &= \frac{\mathbb{P}(X_{n+1} = y - x, S_n = x, S_{n-1} = z_1, \dots, S_1 = z_{n-1})}{\mathbb{P}(S_n = x, S_{n-1} = z_1, \dots, S_1 = z_{n-1})} \\ &= \frac{\mathbb{P}(X_{n+1} = y - x) \mathbb{P}(S_n = x, S_{n-1} = z_1, \dots, S_1 = z_{n-1})}{\mathbb{P}(S_n = x, S_{n-1} = z_1, \dots, S_1 = z_{n-1})} \\ &= \mathbb{P}(X_{n+1} = y - x), \end{aligned}$$

where the second equality follows from the independence assumption for the sequence $X_n, n \geq 1$. A similar derivation gives $\mathbb{P}(S_{n+1} = y | S_n = x) = \mathbb{P}(X_{n+1} = y - x)$ and we get the required equality $\mathbb{P}(S_{n+1} = y | S_n = x, S_{n-1} = z_1, \dots, S_1 = z_{n-1}) = \mathbb{P}(S_{n+1} = y | S_n = x)$.

Definition 1. A discrete time stochastic process $(X_n, n \geq 1)$ is defined to be a Markov chain if it takes values in some countable set \mathcal{X} , and for every $x_1, x_2, \dots, x_n \in \mathcal{X}$ it satisfies the property

$$\begin{aligned}\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \dots, X_1 = x_1) \\ = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1})\end{aligned}$$

The elements of \mathcal{X} are called *states*. We say that the Markov chain is in state $s \in \mathcal{X}$ at time n if $X_n = s$. Mostly for now we will consider the case when finite. In this case we call $(X_n, n \geq 1)$ a *finite state Markov chain* and, without the loss of generality, we will assume that $\mathcal{X} = \{1, 2, \dots, N\}$.

Let us establish some properties of Markov chains.

Proposition 1. Given a Markov chain $(X_n, n \geq 1)$,

1. For every collection of states x_1, x_2, \dots, x_{n-1} and every m

$$\begin{aligned}\mathbb{P}(X_{n+m} = s | X_{n-1} = x_{n-1}, \dots, X_1 = x_1) \\ = \mathbb{P}(X_{n+m} = s | X_{n-1} = x_{n-1}).\end{aligned}$$

2. For every collection of states x_1, x_2, \dots, x_n and $k = 1, 2, \dots, n$

$$\begin{aligned}\mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_1 = x_1 | X_k = x_k) \\ = \mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_{k+1} = x_{k+1} | X_k = x_k) \times \\ \times \mathbb{P}(X_{k-1} = x_{k-1}, \dots, X_1 = x_1 | X_k = x_k).\end{aligned}$$

Proof. Exercise. □

2 EXAMPLES

We already have an example of a Markov chain - random walk.

Consider now the following example (Exercise 2, Section 6.1 [1]).- Suppose we roll a die repeatedly and X_n is the number of 6-s we have seen so far. Then X_n is a Markov chain and $\mathbb{P}(X_n = x+1 | X_{n-1} = x) = 1/6$, $\mathbb{P}(X_n = x | X_{n-1} = x) = 5/6$ and $\mathbb{P}(X_n = y | X_{n-1} = x) = 0$ for all $y \neq x, x+1$. Note, that we can think X_n as a random walk, where the transition to the

right occurs with probability $1/6$ and the transition to the same state with the probability $5/6$.

Now let X_n be the largest of the six possible outcomes observed up to time n . Then X_n is again a Markov chain. What are its transition probabilities?

For our next example consider the following model of an inventory process. The inventory can hold finish goods up to capacity $C \in \mathbb{N}$. Every month there is some current inventory level I_n and a certain fixed amount of product $x \in \mathbb{N}$ is produced, as long as limit is not reached, namely $x \leq C$. If $I_n + x > C$, than just enough $C - I_n$ is produced to reach the capacity. Every month there is a random demand $D_n, n \geq 1$, which we assume is i.i.d. If the current inventory level is at least as large as the demand, then the full demand is satisfied. Otherwise as much of the demand is satisfied as possible, bringing the inventory level down to zero.

Let I_n be the inventory level in month n . Then I_n is a Markov chain. Note

$$I_{n+1} = \min((I_n - D_n)^+ + x, C).$$

Specifically, the probability distribution of I_{n+1} given $I_n = i$, is independent from the values $I_m, m \leq n-1$. I_n is a Markov chain taking values in $0, 1, \dots, C$.

3 HOMOGENEOUS FINITE STATE MARKOV CHAINS

We say that the Markov chain X_n is homogeneous if $\mathbb{P}(X_{n+1} = y | X_n = x) = \mathbb{P}(X_2 = y | X_1 = x)$ for all n . Observe that all of our examples are homogeneous Markov chains. For a homogenous Markov chain we can specify transition probabilities $\mathbb{P}(X_{n+1} = y | X_n = x)$ by a sequence of values $p_{x,y} = \mathbb{P}(X_{n+1} = y | X_n = x)$. For the case of finite state Markov chain, say the state space is $\{1, 2, \dots, N\}$. Then the transition probabilities are $p_{i,j}, 1 \leq i, j \leq N$. We call $P = (p_{i,j})$ the transition matrix of X_n . The transition matrix P has the following obvious property $\sum_j p_{i,j} = 1$ for all i . Any non-negative matrix with such property is called stochastic matrix, for obvious reason.

Observe that

$$\begin{aligned}
& \mathbb{P}(X_{n+2} = j | X_n = i) \\
&= \sum_{1 \leq k \leq N} \mathbb{P}(X_{n+2} = j | X_{n+1} = k, X_n = i) \mathbb{P}(X_{n+1} = k | X_n = i) \\
&= \sum_{1 \leq k \leq N} \mathbb{P}(X_{n+2} = j | X_{n+1} = k) \mathbb{P}(X_{n+1} = k | X_n = i) \\
&= \sum_{1 \leq k \leq N} p_{k,j} p_{i,k}.
\end{aligned}$$

This means that the matrix P^2 gives the two-step transition probabilities of the underlying Markov chain. Namely, the (i,j) -th entry of P^2 , which we denote by $p_{i,j}^{(2)}$ is precisely $\mathbb{P}(X_{n+2} = j | X_n = i)$. This observation is not hard to extend to the general case: for every $r \geq 1$, P^r is the transition matrix of r steps of the Markov chain. One of our goals is understanding the long-term dynamics of P^r as $r \rightarrow \infty$. We will see that for a broad class of Markov chains the following property holds: the limit $\lim_{r \rightarrow \infty} p_{i,j}^{(r)}$ exists and depends only on i and j . Namely, the starting state is irrelevant, as far as the limit is concerned. This property is called *mixing* and is a very important property of Markov chains.

Let e_j denote the j -th N -dimensional column vector. Namely, e_j has j -th coordinate equal to one, and all the other coordinates equal to zero. We also let e denote the N -dimensional column vector consisting of ones. Suppose $X_0 = i$, for some state $i \in \{1, \dots, N\}$. Then the probability vector \mathbf{x}_0 can be written as $e_i^T P^n$ in vector form. Suppose at time zero, the state of the chain is random and is given by some probability vector μ . Namely, $\mathbb{P}(X_0 = i) = \mu_i, i = 1, 2, \dots, N$. Then the probability vector \mathbf{x}_n is precisely $\mu^T P^n$ in vector form.

4 STATIONARY DISTRIBUTION

Consider the following simple Markov chain on states $\{1, 2\}$: $p_{1,1} = p_{1,2} = 1/2, p_{2,1} = 1, p_{2,2} = 0$. Suppose we start at random at time zero with the following probability distribution μ : $\mu_1 = \mathbb{P}(X_0 = 1) = 2/3, \mu_2 = \mathbb{P}(X_0 = 2) = 1/3$. What is the probability distribution \mathbf{x}_1 ? We have $\mathbb{P}(X_1 = 1) = (1/2)\mathbb{P}(X_0 = 1) + \mathbb{P}(X_0 = 2) = (1/2)(2/3) + (1/3) = 2/3$. From this we find $\mathbb{P}(X_1 = 2) = 1 - \mathbb{P}(X_1 = 1) = 1/3$. We see that the probability distribution of X_0 and X_1 are identical. The same applies to every

Definition 2. A probability vector $\pi = (\pi_i), 1 \leq i \leq N$ is defined to be a **stationary distribution** if $\mathbb{P}(X_n = i) = \pi_i$ for all times $n \geq 1$ and states $i = 1, \dots, N$, conditioned on $\mathbb{P}(X_0 = i) = \pi_i, 1 \leq i \leq N$. In this case we also say that the Markov chain is **steady-state**.

Repeating the derivation above for the case of general Markov chains, it is not hard to see that the vector is stationary iff it satisfies the following properties: $\pi_i \geq 0, \sum_i \pi_i = 1$ and

$$\pi_i = \sum_{1 \leq k \leq N} p_{k,i} \pi_k, \quad \forall i.$$

In vector form this can be written as

$$\pi^T = \pi^T P, \quad (1)$$

where w^T denotes the (row) transpose of a column vector

One of the fundamental properties of finite state Markov chains is that a stationary distribution always exists.

Theorem 1. Given a finite state Markov chain with transition matrix there exists at least one stationary distribution. Namely the system of equation (1) has at least one solution satisfying $\geq 0, \sum_i \pi_i = 1$.

Proof. There are many proofs of this fundamental results. One possibility is to use Brower's Fixed Point Theorem. Later on we will give another probabilistic proof which provides important intuition about the meaning of For now let us give a quick proof, but one that relies on linear programming (LP). If you are not familiar with linear programming theory, you can simply skip the proof.

Consider the following LP problem in variables π_1, \dots, π_N .

$$\max \sum_{1 \leq i \leq N} \pi_i$$

Subject to:

$$P^T \pi - \pi = 0,$$

$$\pi \geq 0.$$

Note that a stationary vector exists iff this LP has an unbounded optimal solution. Indeed, if it is a stationary vector, then it clearly is a feasible solution to this LP. Note that $\alpha\pi$ is also a solution for every $\alpha > 0$. Since $\alpha \sum_{1 \leq i \leq N} \pi_i = \alpha$, then we can obtain a feasible solution as large as we want. On the other hand, suppose this LP has an unbounded objective value. In particular, there exists a solution x satisfying $\sum_i x_i > 0$. Taking $\pi_i = x_i / \sum_i x_i$ we obtain a stationary distribution.

Now using LP duality theory, this LP has an unbounded solution iff the dual solution is infeasible. The dual solution is

$$\min \sum_{1 \leq i \leq N} 0y_i$$

Subject to:

$$Py - y \geq e.$$

Let us show that indeed this dual LP problem is infeasible. Suppose the contrary is true. Namely, there exists y satisfying $Py - y \geq e$. Take any such y and find $k^*, 1 \leq k^* \leq N$ such that $y_{k^*} = \max_i y_i$. Observe that $\sum_i p_{k^*,i} y_i \leq \sum_i p_{k^*,i} y_{k^*} = y_{k^*} < 1 + y_{k^*}$, since the rows of P sum to one. Thus the constraint $Py - y \geq e$ is violated in the k^* -th row. We conclude that the dual problem is indeed infeasible. Thus the primal LP problem is unbounded and the stationary distribution exists. \square

As we mentioned above, stationary distribution is not necessarily unique, though in many special cases it is. The uniqueness can be verified by checking whether the following system has a unique solution $T = \pi^T P$, $\sum_j \pi_j = 1$, $\pi_j \geq 0$.

Example.[6.6 from [2]] An absent-minded professor has two umbrellas, used when commuting from home to work and back. If it rains and umbrella is available, the professor takes it. If umbrella is not available, the professor gets wet. If it does not rain the professor does not take the umbrella. It rains on a given commute with probability p independently for all days. What is the steady-state probability that the professor will get wet on a given day?

We model the process as a Markov chain with states 0, 1, 2. The state j means the location where the professor is currently in j umbrellas. Then the corresponding transition probabilities are $p_{0,0} = 1$, $p_{2,1} = p$, $p_{1,2} = p$, $p_{1,1} = 1 - p$, $p_{2,0} = 1 - p$. The corresponding equations for π_j , $j = 0, 1, 2$ are then $\pi_0 = \pi_2(1-p)$, $\pi_1 = (1-p)\pi_1 + p\pi_2$, $\pi_2 = \pi_0 + p\pi_1$. From the second equation $\pi_1 = \pi_2$. Combining with the first equation and with the fact $\pi_1 + \pi_2 = 1$,

we obtain $\pi_1 = \pi_2 = \frac{1}{3-p}$, $\pi_0 = \frac{1-p}{3-p}$. The steady-state probability that the professor gets wet is the probability of being in state zero times probability that it rains on this day. Namely $\text{ifP(wet)} = \frac{(1-p)p}{3-p}$.

5 CLASSIFICATION OF STATES. RECURRENT AND TRANSIENT STATES

Given a finite state homogeneous Markov chain with transition matrix P , construct a directed graph as follows: the nodes are $1, 2, \dots, N$. Put edges (i, j) for every pair of states such that $p_{i,j} > 0$. Given two states i, j suppose there is a directed path from i to j . We say that i communicates with j and write $i \rightarrow j$. By allowing paths of lengths zero, we obtain i communicates with itself ($i \rightarrow i$), although it is possible that starting from i , after time $n = 0$ the chain never returns to i . What is the probabilistic interpretation of this? It means there is a positive probability of getting to j starting from i . Formally $\sum_n p_{i,j}^{(n)} > 0$. Suppose, there is a path from j to i , but not from i to j . This means that if the chain starting from j got to i , then it will never return to j again. Since, there is a positive chance of going from j to i , intuitively, this will happen with probability one. Thus with probability one we will never return to i . We would like to formalize this intuition.

Definition 3. A state i is called transient if there exists a state j such that $i \rightarrow j$, but $j \not\rightarrow i$. Otherwise i is called recurrent.

We write $i \leftrightarrow j$ if states i and j communicate with each other. Observe that if $i \leftrightarrow j$ then $j \leftrightarrow i$. Also, if $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$. Finally, observe that if i is recurrent then it must be the case that $i \leftrightarrow i$. Indeed, consider any state j (possibly i itself) such that $p_{i,j} > 0$. If there is a path from j to i , then there is a path from i to j as well and the assertion is established. Otherwise, we find that $i \rightarrow j$, but $j \not\rightarrow i$, and therefore i is not recurrent. We conclude that there is an equivalency relationship on the set of recurrent states, and we can partition all the recurrent states into equivalency classes R_1, R_2, \dots, R_r . The entire states space $\{1, 2, \dots, N\}$ then can be partitioned as $T \cup R_1 \cup \dots \cup R_r$, where T is the (possibly empty) set of transient states.

References

- [1] G. R. Grimmett and D. R. Stirzak *Probability and Random Processes*, Oxford University Press, 3rd edition, 2001.
- [2] D. P. Bertsekas and J. N. Tsitsik *Introduction to probability*, Athena Scientific, 2002.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MARKOV CHAINS II**Contents**

1. Markov chains with a single recurrence class
2. Uniqueness of Stationary Distributions

1 Markov chains with a single recurrence class

Recall the relations $\rightarrow, \leftrightarrow$ introduced in the previous lecture for the class of finite state Markov chains. Recall that we defined a state i to be recurrent if whenever $i \rightarrow j$ we also have $j \rightarrow i$, namely $i \leftrightarrow j$. We have observed that \leftrightarrow is an equivalency relation, so that set of recurrent states is partitioned into equivalency classes R_1, \dots, R_r . The remaining states \mathcal{T} are transient.

Lemma 1. For every $l = 1, \dots, r$ and every $i \in R_l, j \notin R_l$ we must have $p_{i,j} = 0$.

This means that once the chain is in some recurrent class R it stays there forever.

Proof. The proof is simple: $p_{i,j} > 0$ implies $i \rightarrow j$. Since i is recurrent then also $j \rightarrow i$ implying $j \in R$ - contradiction. \square

Introduce the following basic random quantities. Given states i, j let

$$T_i = \min\{n \geq 1 : X_n = i | X_0 = i\}.$$

In case no such n exists, we set $T_i = \infty$. Thus the range of T_i is $\mathcal{N} \cup \{\infty\}$. The quantity is called the *the first passage time*. Let $\mu_i = \mathbb{E}[T_i]$, possibly with $\mu_i = \infty$. This is called mean recurrence time of the state i .

Lemma 2. For every state $i \in \mathcal{T}$, $\mathbb{P}(X_n = i, \text{i.o.}) = 0$. Namely, almost surely, after some finite time n_0 , the chain will never return to i . In addition $\mathbb{E}[T_i] = \infty$.

Proof. By definition there exists a state j such that $i \rightarrow j$, but $j \not\rightarrow i$. It then follows that $\mathbb{P}(T_i = \infty) > 0$ implying $\mathbb{E}[T_i] = \infty$. Now, let us establish the first part.

Let $I_{i,m}$ be the indicator of the event that the M.c. returned to state i at least m times. Notice that $\mathbb{P}(I_{i,1}) = \mathbb{P}(T_i < \infty) < 1$. Also by M.c. property we have $\mathbb{P}(I_{i,m}|I_{i,m-1}) = \mathbb{P}(T_i < \infty)$, as conditioning that at some point the M.c. returned to state i $m-1$ times does not impact its likelihood to return to this state again. Also notice $I_{i,m} \subset I_{i,m-1}$. Thus $\mathbb{P}(I_{i,m}) = \mathbb{P}(I_{i,m}|I_{i,m-1})\mathbb{P}(I_{i,m-1}) = \mathbb{P}(T_i < \infty)\mathbb{P}(I_{i,m-1}) = \dots = \mathbb{P}^m(T_i < \infty)$. Since $\mathbb{P}(T_i < \infty) < 1$, then by continuity of probability property we obtain $\mathbb{P}(\cap_m I_{i,m}) = \lim_{m \rightarrow \infty} \mathbb{P}(I_{i,m}) = \lim_{m \rightarrow \infty} \mathbb{P}^m(T_i < \infty) = 0$. Notice that the event $\cap_m I_{i,m}$ is precisely the event $X_n = i$, i.o. \square

Exercise 1. Show that $\mathcal{T} \neq \mathcal{X}$. Namely, in every finite state M.c. there exists at least one recurrent state.

Exercise 2. Let $i \in \mathcal{T}$ and let π be an arbitrary stationary distribution. Establish that $\pi_i = 0$.

Exercise 3. Suppose M.c. has one recurrent class R . Show that for every $i \in R$ $\mathbb{P}(X_n = i, \text{i.o.}) = 1$. Moreover, show that there exists $0 < q < 1$ and $C > 0$ such that $\mathbb{P}(T_i > t) \leq Cq^t$ for all $t \geq 0$. As a result, show that $\mathbb{E}[T_i] < \infty$.

We now focus on the family of Markov chains with only one recurrent class. Namely $\mathcal{X} = \mathcal{T} \cup R$. If in addition $\mathcal{T} = \emptyset$, then such a M.c. is called *irreducible*.

2 Uniqueness of the stationary distribution

We now establish a fundamental result on M.c. with a single recurrence class.

Theorem 1. A finite state M.c. with a single recurrence class has a unique stationary distribution π , which is given as $\pi_i = \frac{1}{\mu_i}$ for all states i . Specifically, $\pi_i > 0$ iff the state i is recurrent.

Proof. Let P be the transition matrix of the chain. We let the state space be $\mathcal{X} = \{1, \dots, N\}$. We fix an arbitrary recurrent state k . We know that one exists by Exercise 1. Assume $X_0 = k$. Let N_i be the number of visits to state i between two successive visits to state k . In case $i = k$, the last visit is counted but the initial is not. Namely, in the special case $i = k$ the number of visits is 1 with probability one. Let $\rho_i(k) = \mathbb{E}[N_i]$. Consider the event $\{X_n = i, T_k \geq n\}$ and consider the indicator function $\sum_{n \geq 1} I_{X_n=i, T_k \geq n} = \sum_{1 \leq n \leq T_k} I_{X_n=i}$. Notice that this sum is precisely N_i . Namely,

$$\rho_i(k) = \sum_{n \geq 1} \mathbb{P}(X_n = i, T_k \geq n | X_0 = k). \quad (1)$$

Then using the formula $\mathbb{E}[Z] = \sum_{n \geq 1} \mathbb{P}(Z \geq n)$ for integer valued r.v., we obtain

$$\sum_i \rho_i(k) = \sum_{n \geq 1} \mathbb{P}(T_k \geq n | X_0 = k) = \mathbb{E}[T_k] = \mu_k. \quad (2)$$

Since k is recurrent, then by Exercise 3, $\mu_k < \infty$ implying $\rho_i(k) < \infty$. We let $\rho(k)$ denote the vector with components $\rho_i(k)$.

Lemma 3. $\rho(k)$ satisfies $\rho^T(k) = \rho^T(k)P$. In particular, for every recurrent state k , $\pi_i = \frac{\rho_i(k)}{\mu_k}, 1 \leq i \leq N$ defines a stationary distribution.

Proof. The second part follows from (2) and the fact that $\mu_k < \infty$. Now we prove the first part. We have for every $n \geq 2$

$$\begin{aligned} \mathbb{P}(X_n = i, T_k \geq n | X_0 = k) &= \sum_{j \neq k} \mathbb{P}(X_n = i, X_{n-1} = j, T_k \geq n | X_0 = k) \\ &\quad (3) \end{aligned}$$

$$= \sum_{j \neq k} \mathbb{P}(X_{n-1} = j, T_k \geq n - 1 | X_0 = k) p_{j,i} \quad (4)$$

Observe that $\mathbb{P}(X_1 = i, T_k \geq 1 | X_0 = k) = p_{k,i}$. We now sum the (3) over n and apply it to (1) to obtain

$$\rho_i(k) = p_{k,i} + \sum_{j \neq k} \sum_{n \geq 2} \mathbb{P}(X_{n-1} = j, T_k \geq n - 1 | X_0 = k) p_{j,i}$$

We recognize $\sum_{n \geq 2} \mathbb{P}(X_{n-1} = j, T_k \geq n-1 | X_0 = k)$ as $\rho_j(k)$. Using $\rho_k(k) = 1$ we obtain

$$\rho_i(k) = \rho_k(k)p_{k,i} + \sum_{j \neq k} \rho_j(k)p_{j,i} = \sum_j \rho_j(k)p_{j,i}$$

which is in vector form precisely $\rho^T(k) = \rho^T(k)P$. \square

We now return to the proof of the theorem. Let π denote an *arbitrary* stationary distribution of our M.c. We know one exists by Lemma 3 and, independently by our linear programming based proof. By Exercise 2 we already know that $\pi_i = 1/\mu_i = 0$ for every transient state i .

We now show that in must be that $\pi_k = 1/\mu_k$ for every recurrent state k . In particular, the stationary distribution is unique. Assume that at time zero we start with distribution π . Namely $\mathbb{P}(X_0 = i) = \pi_i$ for all i . Of course this implies that $\mathbb{P}(X_n = i)$ is also π_i for all n . On the other hand, fix any recurrent state k and consider

$$\begin{aligned} \mu_k \pi_k &= \mathbb{E}[T_k | X_0 = k] \mathbb{P}(X_0 = k) \\ &= \sum_{n \geq 1} \mathbb{P}(T_k \geq n | X_0 = k) \mathbb{P}(X_0 = k) \\ &= \sum_{n \geq 1} \mathbb{P}(T_k \geq n, X_0 = k). \end{aligned}$$

On the other hand $\mathbb{P}(T_k \geq 1, X_0 = k) = \mathbb{P}(X_0 = k)$ and for $n \geq 2$

$$\begin{aligned} \mathbb{P}(T_k \geq n, X_0 = k) &= \mathbb{P}(X_0 = k, X_j \neq k, 1 \leq j \leq n-1) \\ &= \mathbb{P}(X_j \neq k, 1 \leq j \leq n-1) - \mathbb{P}(X_j \neq k, 0 \leq j \leq n-1) \\ &\stackrel{(*)}{=} \mathbb{P}(X_j \neq k, 0 \leq j \leq n-2) - \mathbb{P}(X_j \neq k, 0 \leq j \leq n-1) \\ &= a_{n-2} - a_{n-1}, \end{aligned}$$

where $a_n = \mathbb{P}(X_j \neq k, 0 \leq j \leq n)$ and $(*)$ follows from stationarity of π . Now $a_0 = \mathbb{P}(X_0 \neq k)$. Putting together, we obtain

$$\begin{aligned} \mu_k \pi_k &= \mathbb{P}(X_0 = k) + \sum_{n \geq 2} (a_{n-2} - a_{n-1}) \\ &= \mathbb{P}(X_0 = k) + \mathbb{P}(X_0 \neq k) - \lim_n a_n \\ &= 1 - \lim_n a_n \end{aligned}$$

But by continuity of probabilities $\lim_n a_n = \mathbb{P}(X_n \neq k, \forall n)$. By Exercise 3, the state k , being recurrent is visited infinitely often with probability one. We conclude that $\lim_n a_n = 0$, which gives $\mu_k \pi_k = 1$, implying that π_k is uniquely defined as $1/\mu_k$. \square

3 Ergodic theorem

Let $N_i(t)$ denote the number of times the state i is visited during the times $0, 1, \dots, t$. What can be said about the behavior of $N_i(t)/t$ when t is large? The answer turns out to be very simple: it is π_i . These type of results are called *ergodic* properties, as they show how the time average of the system, namely $N_i(t)/t$ relates to the spatial average, namely π_i .

Theorem 2. For arbitrary starting state $X_0 = k$ and for every state i ,

$$\lim_{t \rightarrow \infty} \frac{N_i(t)}{t} = \pi_i$$

almost surely. Also

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}[N_i(t)]}{t} = \pi_i.$$

Proof. Suppose $X_0 = k$. If i is a transient state, then, as we have established, almost surely after some finite time, the chain will never enter i , meaning $\lim_t N_i(t)/t = 0$ almost surely. Since also $\pi_i = 0$, then we have established the required equality for the case when i is a transient state.

Suppose now i is a recurrent state. Let T_1, T_2, T_3, \dots denote the time of successive visits to i . Then the sequence $T_n, n \geq 2$ is i.i.d. Also T_1 is independent from the rest of the sequence, although its distribution is different from the one of $T_m, m \geq 2$ since we have started the chain from k which is in general different from i . By the definition of $N_i(t)$ we have

$$\sum_{1 \leq m \leq N_i(t)} T_m \leq t < \sum_{1 \leq m \leq N_i(t)+1} T_m$$

from which we obtain

$$\frac{\sum_{1 \leq m \leq N_i(t)} T_m}{N_i(t)} \leq \frac{t}{N_i(t)} < \frac{\sum_{1 \leq m \leq N_i(t)+1} T_m}{N_i(t)+1} \frac{N_i(t)+1}{N_i(t)}. \quad (5)$$

We know from Exercise 3 that $\mathbb{E}[T_m] < \infty, m \geq 2$. Using a similar approach it can be shown that $\mathbb{E}[T_1] < \infty$, in particular $T_1 < \infty$ a.s. Applying SLLN we have that almost surely

$$\lim_{n \rightarrow \infty} \frac{\sum_{2 \leq m \leq n} T_m}{n} = \lim_{n \rightarrow \infty} \frac{\sum_{2 \leq m \leq n} T_m}{n-1} \frac{n-1}{n} = \mathbb{E}[T_2]$$

which further implies

$$\lim_{n \rightarrow \infty} \frac{\sum_{1 \leq m \leq n} T_m}{n} = \lim_{n \rightarrow \infty} \frac{\sum_{2 \leq m \leq n} T_m}{n} + \lim_{n \rightarrow \infty} \frac{T_1}{n} = \mathbb{E}[T_2]$$

almost surely. \square

Since i is a recurrent state then by Exercise 3, $N_i(t) \rightarrow \infty$ almost surely as $t \rightarrow \infty$. Combining the preceding identity with (5) we obtain

$$\lim_{t \rightarrow \infty} \frac{t}{N_i(t)} = \mathbb{E}[T_2] = \mu_i,$$

from which we obtain $\lim_t N_i(t)/t = \mu_i^{-1} = \pi_i$ almost surely.

To establish the convergence in expectation, notice that $N_i(t) \leq t$ almost surely, implying $N_i(t)/t \leq 1$. Applying bounded convergence theorem, we obtain that $\lim_t \mathbb{E}[N_i(t)]/t = \pi_i$, and the proof is complete.

4 Markov chains with multiple recurrence classes

How does the theory extend to the case when the M.c. has several recurrence classes R_1, \dots, R_r ? The summary of the theory is as follows (the proofs are very similar to the case of single recurrent class case and is omitted). It turns out that such a M.c. chain possesses r stationary distributions $\pi^i = (\pi_1^i, \dots, \pi_N^i), 1 \leq i \leq r$, each "concentrating" on the class R_i . Namely for each i and each state $k \notin R_i$ we have $\pi_k^i = 0$. The i -th stationary distribution is described by $\pi_k^i = 1/\mu_k$ for all $k \in R_i$ and where μ_k is the mean return time from state $k \in R_j$ into itself. Intuitively, the stationary distribution π^i corresponds to the case when the M.c. "lives" entirely in the class R_i . One can prove that the family of all of the stationary distributions of such a M.c. can be obtained by taking all possible convex combinations of $\pi^i, 1 \leq i \leq r$, but we omit the proof. (Exercise: show that a convex combination of stationary distributions is a stationary distribution).

References

- [1] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, Oxford University Press, 3rd edition, 2001.
- [2] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to probability*, Athena Scientific, 2002.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MARKOV CHAINS III**Contents**

1. Periodicity
2. Mixing
3. Absorbtion

1 PERIODICITY

Previously we showed that when a finite state M.c. has only one recurrent class and π is the corresponding stationary distribution, then $[N_i(t)|X_0 = k]/t \rightarrow \pi_i$ as $t \rightarrow \infty$, irrespective of the starting state. Since $N_i(t) = \sum_{n=1}^t I_{\{X_n=i\}}$ is the number of times state i is visited up till time t , we have shown that $\frac{1}{t} \sum_{n=1}^t \mathbb{P}(X_n = i|X_0 = k) \rightarrow \pi_i$ for every state k , i.e., $p_{ki}^{(n)}$ converges to π_i in the Cesaro sense. However, $p_{ki}^{(n)}$, which from now on we call *transient probability distribution* of a Markov chain, need not converge, as the following example shows. Consider a 2 state Markov Chain with states $\{1, 2\}$ and $p_{12} = 1 = p_{21}$. Then $p_{12}^{(n)} = 1$ when n is odd and 0 when n is even.

Let x be a recurrent state and consider all the times when accessible from itself, i.e., the times in the I_x set $= \{n \geq 1 : p_{xx}^{(n)} > 0\}$ (note that this set is non-empty since x is a recurrent state). One property I_x will make use of is that it is closed under addition, i.e., if $n \in I_x$, then $m+n \in I_x$. This is easily seen by observing that $p_{xx}^{(m+n)} \geq p_{xx}^{(m)} p_{xx}^{(n)} > 0$. Let d_x be the greatest common divisor of the numbers in I_x . We call d_x the *period* of x . We now show that all states in the same recurrent class has the same period.

Lemma 1. *If x and y are in the same recurrent class, then $d_x = d_y$.*

Proof. Let m and n be such that $p_{xy}^{(m)}, p_{yx}^{(n)} > 0$. Then $p_{yy}^{(m+n)} \geq p_{xy}^{(m)} p_{yx}^{(n)} > 0$. So d_y divides $m+n$. Let l be such that $p_{xx}^{(l)} > 0$, then $p_{yy}^{(m+n+l)} \geq p_{yx}^{(n)} p_{xx}^{(l)} p_{xy}^{(m)} > 0$. Therefore d_y divides $m+n+l$, hence it divides d_x . This implies that d_y divides d_x . A similar argument shows that d_x divides d_y , so $d_x = d_y$. \square

A recurrent class is said to be *periodic* if the period d is greater than 1 and *aperiodic* if $d = 1$. The 2 state Markov Chain in the example above has a period of 2 since $p_{11}(n) > 0$ iff n is even. A recurrent class with period d can be divided into d subsets, so that all transitions from one subset lead to the next subset.

Why is periodicity of interest to us? It is because periodicity is exactly what prevents the convergence of $p_{xy}^{(n)}$ to π_y . Suppose y is a recurrent state with period $d > 1$. Then $p_{yy}^{(n)} = 0$ unless n is a multiple of d , but $\pi_y > 0$. However, if $d = 1$, we have positive probability of returning to y for all time steps sufficiently large.

Lemma 2. *If $d_y = 1$, then there exists some $N \geq 1$ such that $p_{yy}^{(n)} > 0$ for all $n \geq N$.*

Proof. We first show that $I_y = \{n \geq 1 : p_{yy}^{(n)} > 0\}$ contains two consecutive integers. Let n and $n+k$ be elements of I_y . If $k=1$, then we are done. If not, then since $d_y = 1$, we can find $n_1 \in I_y$ such that k is not a divisor of n_1 . Let $n_1 = mk + r$ where $0 < r < k$. Consider $(m+1)(n+k)$ and $(m+1)n + n_1$, which are both in I_y since I_y is closed under addition. We have

$$(m+1)(n+k) - ((m+1)n + n_1) = k - r < k.$$

So by repeating the above argument at most m times, we eventually obtain a pair of consecutive integers $n, m+1 \in I_y$. If $N = m^2$, then for all $n \geq N$, we have $n - N = km + r$, where $0 \leq r < m$. Then $n = m^2 + km + r = r(1+m) + (m-r+k)m \in I_y$. \square

2 COUPLING TECHNIQUE AND MIXING

We now establish that when a Markov chain has one recurrent class (irreducible) and aperiodic, the transient distribution approaches the (unique) steady state distribution as time goes to infinity. Namely, for every two states we have $p_{xy}^{(n)} \rightarrow \pi_y$ as $n \rightarrow \infty$. This is commonly called mixing property of a Markov chain.

Theorem 1. Consider an irreducible, aperiodic Markov chain. Then for all states x, y , $\lim_{n \rightarrow \infty} p_{xy}^{(n)} = \pi_y$.

For the case of periodic chains, there is a similar statement regarding convergence of $p_{xy}^{(n)}$, but now the convergence holds only for certain subsequences of the time index. See [1] for further details.

There are at least two generic ways to prove this theorem. One is based on the Perron-Frobenius Theorem which characterizes eigenvalues and eigenvectors of non-negative matrices. Specifically the largest eigenvalue λ_1 of P is equal to unity and all other eigenvalues are strictly smaller than unity in absolute value. The P-F Theorem is especially useful in the special case of so-called *reversible* M.c.. These are irreducible M.c. for which the unique stationary distribution satisfies $\pi_x p_{xy} = \pi_y p_{yx}$ for all states x, y . The subject of reversible M.c. is a rich subject on its own and is outside of the scope of this lecture. In the special case of reversible M.c. the following important refinement of Theorem 1 is known.

Theorem 2. Consider an irreducible aperiodic Markov chain which is reversible. Then there exists a constant C such that for all states x, y , $|p_{xy}^{(n)} - \pi_y| \leq C|\lambda_2|^n$, where λ_2 is the second largest (in absolute value) eigenvalue of P .

Since by P-F Theorem $|\lambda_2| < 1$, this theorem is indeed a refinement of Theorem 1 as it gives a concrete rate of convergence to the steady-state.

We adopt a different approach which does not rely on the reversibility assumption. The main technique underlying our approach is the method of coupling, which we now discuss. The method of coupling allows combining two Markov chains into one by building them on the same probability space. Intuitively, two Markov chains X_n and Y_n are coupled if we construct out of them a single Markov chain $Z_n = (X_n, Y_n)$ such that each "marginal" Markov chain X_n and Y_n behaves as an individual Markov chain before coupling, but the evolution of X_n and Y_n is in general dependent. Formally

Definition 1. Given two Markov chains X_n and Y_n on state spaces $\{1, \dots, N\}$ and $\{1, \dots, M\}$, respectively, and transition probability matrices $P = (p_{xy}, 1 \leq x, y \leq N)$ and $Q = (q_{xy}, 1 \leq x, y \leq M)$, a coupling of X_n and Y_n is any Markov chain Z_n with a state space $\{1, \dots, N\} \times \{1, \dots, M\}$ and a transition matrix

$$R = (r_{(x_1, x_2), (y_1, y_2)}, 1 \leq x_1, y_1 \leq N, 1 \leq x_2, y_2 \leq M),$$

which satisfies the following properties: for every $x_1, y_1 \leq N$ and $1 \leq x_2 \leq M$,

$$\sum_{y_2=1}^M r_{(x_1, x_2), (y_1, y_2)} = p_{x_1, y_1},$$

and for every $1 \leq x_1 \leq N$ and $1 \leq x_2, y_2 \leq M$,

$$\sum_{y_1=1}^N r_{(x_1, x_2), (y_1, y_2)} = q_{x_2, y_2},$$

In words, coupling $Z_n = (X_n, Y_n)$ of two Markov chains X_n and Y_n means that the Markov chain X_n transitions from state x_1 to state y_1 with probability p_{x_1, y_1} , regardless of the state of Y_n , and vice versa.

How do we know that the definition of the coupling is not vacuous and at least one coupling exists? This is easy: simply consider running chains X_n and Y_n independently and set $Z_n = (X_n, Y_n)$. Formally, set $r_{(x_1, x_2), (y_1, y_2)} = p_{x_1, y_1} q_{x_2, y_2}$ for all x_1, x_2, y_1, y_2 . It is easy to check that this leads to a valid coupling. However, this is also the least useful coupling. We now consider a different coupling idea applied in the special case where $N = M$ and $Q = P$. Namely, we will couple a Markov chain with itself. For convenience, we again use notation X_n and Y_n , though now Y_n has the same state space $\{1, \dots, N\}$ and transition matrix $Q = P$ as X_n . We now defined the coupled chain $Z_n = (X_n, Y_n)$ according to the following rules

$$r_{(x_1, x_2), (y_1, y_2)} = \begin{cases} p_{x_1, y_1} p_{x_2, y_2}, & \text{when } x_1 \neq x_2; \\ p_{x_1, y_1}, & \text{when } x_1 = x_2, y_1 = y_2; \\ 0, & \text{when } x_1 = x_2, y_1 \neq y_2; \end{cases}$$

In words, the Markov chain X_n and Y_n run independently until they collide for the first time in the same state $X_n = Y_n = x$. Once this happen, they transition

to a new state which is the same for X_n and Y_n with probability p_{xy} - the transition probability of the original Markov chain. Again, it is easy to check that $R = (r_{(x_1, x_2), (y_1, y_2)})$ defines a valid coupling of the Markov chain with itself. We define T to be the first (random) time when two copies of M.c. collide for the first time. Namely, $T = \min\{n \geq 0 : X_n = Y_n\}$. Then, $X_n = Y_n$ for all $n \geq T$. We define T to be infinite if the states never collide. We are now ready to prove the "mixing" theorem.

Proof of Theorem 1. Fix an arbitrary two states x_0 and y_0 . We need to show that $\lim_{n \rightarrow \infty} p_{x_0, y_0}^{(n)} = \pi_{y_0}$. We assume for simplicity that all transition probabilities are positive: $p_{xy} > 0, \forall x, y \in \{1, \dots, N\}$. The general case is the subject of an exercise. Fix any $\delta > 0$ such that $p_{xy} \geq \delta$ for all states x, y . Consider a coupling $Z_n = (X_n, Y_n)$ of X_n with itself described above. To completely describe the probabilistic evolution of Z_n we need to specify the initial distribution. We will be judicious about this. Specifically, let $X_0 = x_0$ with probability one and let Y_n be distributed according to π . Formally, $\mathbb{P}(Z_0 = (x_0, x)) = \pi_x$, and $\mathbb{P}(Z_0 = (x', x)) = 0$ for all $x' \neq x$. This in particular, means that $\mathbb{P}(Y_n = x) = \pi_x$ for all n and $\mathbb{P}(X_n = x) = p_{x_0, x}^{(n)}$, though notice that we explicitly write down the joint probability of X_n and Y_n in terms of P and π , since X_n and Y_n run independently.

Let $T \geq 0$ be defined as above - the first time when $X_n = Y_n$. Observe that, if the Markov chains X_n and Y_n did not collide by time t , they will collide at time $n + 1$ with probability at least δ , since every state is reachable with probability δ . Therefore, $\mathbb{P}(T \geq t) \leq \delta^t$. In particular, by continuity of probabilities, $\mathbb{P}(T = \infty) = \lim_t \mathbb{P}(T \geq t) = 0$. Now we have

$$\begin{aligned} \mathbb{P}(X_n = y_0) &= \mathbb{P}(X_n = y_0, T \leq n) + \mathbb{P}(X_n = y_0, T > n) \\ &\leq \mathbb{P}(X_n = y_0, T \leq n) + \mathbb{P}(T > n) \\ &= \mathbb{P}(Y_n = y_0, T \leq n) + \mathbb{P}(T > n) \\ &\leq \mathbb{P}(Y_n = y_0) + \mathbb{P}(T > n) \\ &= \pi_{y_0} + \mathbb{P}(T > n). \end{aligned}$$

Here the second equality is valid since on the event $T \leq n$ we have $X_n = Y_n$ (the collision took place at time n or earlier). Taking the limit of both sides we obtain

$$\limsup_n \mathbb{P}(X_n = y_0) \leq \pi_{y_0}.$$

But recall that $\mathbb{P}(X_n = y_0) = p_{x_0, y_0}^{(n)}$. Similarly, we have

$$\begin{aligned}\mathbb{P}(X_n = y_0) &\geq \mathbb{P}(X_n = y_0, T \leq n) \\ &= \mathbb{P}(Y_n = y_0, T \leq n) \\ &= \mathbb{P}(Y_n = y_0) - \mathbb{P}(Y_n = y_0, T > n) \\ &\geq \mathbb{P}(Y_n = y_0) - \mathbb{P}(T > n) \\ &\geq \pi_{y_0} - \mathbb{P}(T > n).\end{aligned}$$

Again, by taking limits, we obtain $\liminf_n \mathbb{P}(X_n = y_0) \geq \pi_{y_0}$. Combining, we obtain $\lim_n \mathbb{P}(X_n = y_0) = \lim_n p_{x_0, y_0}^{(n)} = \pi_{y_0}$ and the proof is complete. \square

3 ABSORPTION PROBABILITY AND EXPECTED TIME TILL ABSORPTION

We have considered the long-term behavior of Markov chains. Now, we study the short-term behavior. In such considerations, we are concerned with the behavior of the chain starting in a transient state, till it enters one of the recurrent state. For simplicity, we can therefore assume that every recurrent ~~is a~~-absorbing, i.e., $p_{ii} = 1$. The Markov chain that we will work with in this section has only transient and absorbing states.

If there is only one absorbing state then $\pi_i = 1$, and i is reached with probability 1. If there are multiple absorbing states, the state that is entered is random, and we are interested in the absorbing probability

$$a_{ki} = \mathbb{P}(X_n \text{ eventually equals } i \mid X_0 = k),$$

i.e., the probability that state i is eventually reached, starting from state k . Note that $a_{ii} = 1$ and $a_{ji} = 0$ for all absorbing $j \neq i$. When k is a transient state, we have

$$\begin{aligned}a_{ki} &= \mathbb{P}(\exists n : X_n = i \mid X_0 = k) \\ &= \sum_{j=1}^N \mathbb{P}(\exists n : X_n = i \mid X_1 = j) p_{kj} \\ &= \sum_{j=1}^N a_{ji} p_{kj}.\end{aligned}$$

So we can find the absorption probabilities by solving the above system of linear equations.

Example: Gambler's Ruin A gambler wins 1 dollar at each round, with probability p , and loses a dollar with probability p . Different rounds are independent. The gambler plays continuously until he either accumulates a target amount m or loses all his money. What is the probability of losing his fortune?

We construct a Markov chain with state space $\{0, 1, \dots, m\}$, where the state i is the amount of money the gambler has. So 0 corresponds to losing his entire fortune, and m corresponds to accumulating the target amount. The states 0 and m are absorbing states. We have the transition probabilities $p_{i,i+1} = p$, $p_{i,i-1} = 1 - p$ for $i = 1, 2, \dots, m - 1$, and $p_{00} = p_{mm} = 1$. To find the absorbing probabilities for the state 0 , we have

$$\begin{aligned} a_{00} &= 1, \\ a_{m0} &= 0, \\ a_{i0} &= (1 - p)a_{i-1,0} + pa_{i+1,0}, \quad \text{for } i = 1, \dots, m - 1. \end{aligned}$$

Let $b_i = a_{i0} - a_{i+1,0}$, $\rho = (1 - p)/p$, then the above equation gives us

$$(1 - p)(a_{i-1,0} - a_{i0}) = p(a_{i0} - a_{i+1,0}).$$

Namely, $b_i = \rho b_{i-1}$. So we obtain $b_0 = \rho^i b_0$. Note that $b_0 + b_1 + \dots + b_{m-1} = a_{00} - a_{m0} = 1$, hence $(1 + \rho + \dots + \rho^{m-1})b_0 = 1$, which gives us

$$b_i = \begin{cases} \frac{\rho^i(1-\rho)}{1-\rho^m}, & \text{if } \rho \neq 1, \\ \frac{1}{m}, & \text{if } \rho = 1. \end{cases}$$

Finally, $a_{i,0}$ can be calculated. For $\rho \neq 1$, we have for $i = 1, \dots, m - 1$,

$$\begin{aligned} a_{i0} &= a_{00} - b_{i-1} - \dots - b_0 \\ &= 1 - (\rho^{i-1} + \dots + \rho + 1)b_0 \\ &= 1 - \frac{1 - \rho^i}{1 - \rho} \frac{1 - \rho}{1 - \rho^m} \\ &= \frac{\rho^i - \rho^m}{1 - \rho^m} \end{aligned}$$

and for $\rho = 1$,

$$a_{i0} = \frac{m - i}{m}.$$

This shows that for any fixed i if $\rho > 1$, i.e., $p < 1/2$, the probability of losing goes to 1 as $i \rightarrow \infty$. Hence, it suggests that if the gambler aims for a

large target while under unfavorable odds, financial ruin is almost certain.

The expected time of absorption μ_k when starting in a transient state can be defined as $\mu_k = \mathbb{E}[\min\{n \geq 1 : X_n \text{ is recurrent}\} | X_0 = k]$. A similar analysis by conditioning on the first step of the Markov chain shows that the expected time to absorption can be found by solving

$$\begin{aligned}\mu_k &= 0 \quad \text{for all recurrent states } k, \\ \mu_k &= 1 + \sum_{j=1}^N p_{kj} \mu_j \quad \text{for all transient states } k.\end{aligned}$$

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

INFINITE MARKOV CHAINS. CONTINUOUS TIME MARKOV CHAINS.**Contents**

1. Recurrence and transience
2. Stationary distributions
3. Generators

1 INTRODUCTION

In this lecture we discuss infinite state Markov chains. Then we consider finite and infinite state M.c. where the transition between the states occurs during some random time interval, as opposed to unit time steps. Most of the times we state the results without proofs. Our treatment of this material is also very brief. A more in depth analysis of these concepts is devoted by the course 6.262 - Discrete Stochastic Processes.

2 INFINITE STATE MARKOV CHAINS

Suppose we have a (homogeneous) Markov chain whose state space is countably infinite $\mathcal{X} = \{0, 1, 2, \dots\}$. In this case the theory is similar in some respects to the finite state counterpart, but different in other respects. We denote again by $p_{i,j}$ the probability of transition from state i to state j . Thus we will consider only homogeneous Markov chains, without explicitly saying this. We introduce the notion of i communicates with j , written as $i \rightarrow j$, in the same manner as before. Thus again we may decompose the state space into transient states i , namely states such that for some j , $i \rightarrow j$ but $j \not\rightarrow i$; and the remaining states which are recurrent. However, in the case of infinite M.c. a new complication appears. To discuss it, let us again define a probability distribution π on \mathcal{X} to be

stationary if it is time invariant. The necessary and sufficient condition for this is $\pi_i \geq 0$, $\sum_{i \in \mathcal{X}} \pi_i = 1$ and for every state i

$$\pi_i = \sum_j \pi_j p_{j,i}.$$

As a result, if the M.c. X_n has the property $X_0 \stackrel{d}{=} \pi$, then $X_n \stackrel{d}{=} \pi$ for all n .

Now let us consider the following M.c. on \mathbb{Z}_+ . A parameter p is fixed. For every integer $i > 0$, $p_{i,i+1} = p$, $p_{i,i-1} = 1 - p$ and $p_{0,1} = p$, $p_{0,0} = 1 - p$. This M.c. is called *random walk with reflection at zero*. Let us try to find a stationary distribution π of this M.c. It must satisfy

$$\begin{aligned}\pi_i &= \pi_{i-1} p_{i-1,i} + \pi_{i+1} p_{i+1,i} = \pi_{i-1} p + \pi_{i+1} (1-p), \quad i \geq 1. \\ \pi_0 &= \pi_0 (1-p) + \pi_1 (1-p).\end{aligned}$$

From this we obtain $\pi_1 = \frac{p}{1-p} \pi_0$ and iterating

$$\pi_{i+1} = \frac{p}{1-p} \pi_i. \tag{1}$$

This gives $\pi_i = (p/(1-p))^i \pi_0$. Now if $p > 1/2$ then $\pi_i \rightarrow \infty$ and we cannot possibly have that $\sum_i \pi_i = 1$. Thus no stationary distribution exists. Note, that however all pairs of states i, j communicate, as we can get from i to $j > i$ in $j-i$ steps with probability $p^{j-i} > 0$, and from j to i in $j-i$ steps with probability $(1-p)^{j-i}$.

We conclude that an infinite state M.c. does not necessarily have a stationary distribution. Recall that in the case of finite state M.c. if i is a recurrent state, then its recurrence time T_i has finite expected value (as it has geometrically decreasing tails). For the case of infinite M.c. the difficulty is the fact that while every state i communicates with every other state j , it is possible that the chain starting from i wanders off to "infinity" for ever without ever returning to i . Furthermore, it is possible that even if the chain returns to i infinitely often with probability one, the *expected* return time from i to i is infinite. Recall, that the return time is defined to be $T_i = \min\{n \geq 1 : X_n = i\}$, when the M.c. starts at i at time 0, when such n exists, and defined to be $T_i = \infty$ when the chain never returns to i .

Definition 1. Given an infinite M.c. $X_n, n \geq 1$, the state i is defined to be transient if the probability of never returning to i is positive. Namely,

$$\mathbb{P}(X_n \neq i, \forall n \geq 1 | X_0 = i) > 0.$$

Otherwise the state is defined to be recurrent. It is defined to be positive recurrent if $\mathbb{E}[T_i] < \infty$ and null-recurrent if $\mathbb{E}[T_i] = \infty$.

Thus, unlike the finite state case, the state is transient if there is a positive probability of no return, as opposed to existence of a state from which the return to starting state has probability zero. It is an exercise to see that the definition above when applied to the finite state case is consistent with the earlier definition. Namely, it is an *implication* of how we defined the transient and recurrent states, rather than the definition. Also, observe that there are no null-recurrent states in the finite state case.

The following theorem holds, the proof of which we skip.

Theorem 1. Given an infinite M.c. $X_n, n \geq 1$ suppose all the states communicate. Then there exists a stationary distribution π iff there exists at least one positive recurrent state i . In this case in fact all the states are positive recurrent and the stationary distribution π is unique. It is given as $\pi_j = 1/\mathbb{E}[T_j] > 0$ for every state j .

We see that in the case when all the states communicate, all states have the same status: positive recurrent, null recurrent or transient. In this case we will say the M.c. itself is positive recurrent, null recurrent, or transient. There is an extension of this theorem to the cases when not all states communicate, but we skip the discussion of those. Similarly, if there are several communicating classes, then there exists at least one stationary distribution per class which contains at least one positive recurrent state (and as a result all states in the class are positive recurrent).

Theorem 2. A random walk with reflection X_n on \mathbb{Z}_+ is positive recurrent if $p < 1/2$, null-recurrent if $p = 1/2$ and transient if $p > 1/2$.

Proof. The case $p < 1/2$ will be resolved by exhibiting explicitly at least one

steady state distribution π . Since all the states communicate, then by Theorem 1 we know that the stationary distribution is unique and $\mathbb{E}[T_i] = 1/\pi_i < \infty$ for all i . Thus the chain is positive recurrent. In fact we can find explicitly the stationary distribution. Consider again at the recurrence (1), which suggests $\pi_i = (p/(1-p))^i \pi_0$. From this we obtain

$$\pi_0(1 + \sum_{i>0} (p/(1-p))^i) = 1$$

implying $\pi_0 = 1 - p/(1-p) = (1-2p)/(1-p)$ and

$$\pi_i = \frac{1-2p}{1-p} \left(\frac{p}{1-p}\right)^i, \quad i \geq 0.$$

This gives us a probability vector π with $\sum_i \pi_i = 1$ and completes the proof for the case $p < 1/2$.

The case $p \geq 1/2$ will be analyzed using our earlier result on random walk on \mathbb{Z} . Recall that for such a r.w. the probability of return to zero is = 1 iff $p = 1/2$. In the case $p = 1/2$ we have also established that the expected return time to zero is infinite. Thus suppose $p = 1/2$. A r.w. without reflection makes the first step into 1 or -1 with probability $1/2$ each. Conditioning on $X_1 = 1$ and conditioning on $X_1 = -1$, we have that the expected return time to zero is again infinite. If the first transition is into 1, then the behavior of this r.w. till the first return to zero is the same as of our r.w. with reflection at zero. In particular, the return to zero happens with probability one and the expected return time is infinite. We conclude that the state 0 is null-recurrent.

Finally, suppose $p > 1/2$. We already saw that the M.c. cannot have a stationary distribution. Thus by Theorem 1, since all the states communicate we have that all states are null-recurrent or transient. We just need to refine this result to show that in fact all states are transient.

For the unreflected r.w. we have that with positive probability the walk never returns to zero. Let, T_0 denote return time to 0 - the time it takes to come back to zero for unreflected random walk, when it at zero. We claim that $\mathbb{P}(T_0 = \infty | X_1 = 1) > 0, \mathbb{P}(T_0 = \infty | X_1 = -1) = 0$. Namely, the "no return to zero" happens iff the first step is to the right. First let us see why just the first inequality, namely $\mathbb{P}(T_0 = \infty | X_1 = 1) > 0$ implies our result. Conditioned on the event that the first step is to the right, the r.w. behaves as r.w. with reflection at zero until the first return to zero. The assumption $\mathbb{P}(T_0 = \infty | X_1 = 1) > 0$ means there is a positive probability of no return for random walk without reflection when the first step is to the right. Then there is a positive probability of no return for the reflected r.w. conditioned on $X_1 = 1$. Since the transition from

zero to 1 occurs with positive probability p , then there is a positive probability of no return to zero starting from zero for the random walk with reflection and thus state 0 is transient. Since all states communicate, this means that all states of the random walk with reflection are transient.

Now we establish that claim. We have $\mathbb{P}(T_0 = \infty) = p\mathbb{P}(T_0 = \infty|X_1 = 1) + (1 - p)\mathbb{P}(T_0 = \infty|X_1 = -1)$. We also have that $\mathbb{P}(T_0 = \infty) > 0$. We now establish that $\mathbb{P}(T_0 = \infty|X_1 = -1) = 0$. This immediately implies $\mathbb{P}(T_0 = \infty|X_1 = 1) > 0$, which we need. Now assume $X_1 = -1$. Consider $Y_n = -X_n$. Observe that, until the first return to zero, Y_n is a reflected r.w. with parameter $q = 1 - p$. Since $q < 1/2$, then, as we established at the beginning of the proof, the process Y_n returns to zero with probability one (moreover the return time has finite expected value). We conclude that X_n returns to zero with probability one, namely $\mathbb{P}(T_0 = \infty|X_1 = -1) = 0$. This completes the proof. \square

3 CONTINUOUS TIME MARKOV CHAINS

We consider a stochastic process $X(t)$ which is a function of a real argument t instead of integer n . Let \mathcal{X} be the state space of this process, which is assumed to be finite or countably infinite.

Definition 2. *$X(t)$ is defined to be a continuous time Markov chain if for every $j, i_1, \dots, i_{n-1} \in \mathcal{X}$ and every sequence of times $t_1 < t_2 < \dots < t_n$,*

$$\mathbb{P}(X(t_n) = j|X(t_{n-1}) = i_{n-1}, \dots, X(t_1) = i_1) \quad (2)$$

$$= \mathbb{P}(X(t_n) = j|X(t_{n-1}) = i_{n-1}). \quad (3)$$

The process is defined to be homogeneous if $\mathbb{P}(X(t) = j|X(s) = i) = \mathbb{P}(X(t-s) = j|X(0) = i)$ for every i, j and $s < t$.

From now on we assume without explicitly saying that our M.c. is homogeneous. We write $p_{i,j}^{(t)}$ for $\mathbb{P}(X(t) = j|X(0) = i)$. The continuous time Markov chain is a special case of a Markov process, the definition of which we skip. Loosely speaking, a stochastic process is a Markov process if its future trajectory is completely determined by its current state, independently from the past. We already know an example of a continuous time M.c. - Poisson process. It is given as $\mathbb{P}(X(t) = i+k|X(s) = i) \stackrel{d}{=} \text{Pois}(\lambda(t-s)), k \geq 0$ and $\mathbb{P}(X(t) = i+k|X(s) = i) = 0$ for $k < 0$.

Given a state i and time t_0 introduce ‘‘holding time’’ $U(i, t_0)$ as $\inf\{s > 0 : X(t_0 + s) \neq i\}$, when $X(t_0) = i$. Namely, it is the time that the chain spends in state i after time t_0 , assuming that it is in i at time t_0 . It might turn out in special cases that $U(i, t_0) = 0$ with positive probability. But in many special cases this will not happen. For now we assume that $U(i, t_0) > 0$ a.s. In special cases we can establish this directly.

Proposition 1. *For every state i and time t_0 , $U(i, t_0) \stackrel{d}{=} \text{Exp}(\mu_i)$ for some parameter μ_i which depends only on the state.*

Since, per proposition above, the distribution of holding time is exponential, and therefore memoryless, we see that the time till the next transition occurs is independent from the past history of the chain and only depends on the current state i . The parameter μ_i is usually called transition rate out of state i . This is a very fundamental (and useful) property of continuous time Markov chains.

Proof sketch. Consider

$$\mathbb{P}(U(i, t_0) > x + y | U(i, t_0) > x, X(t_0) = i).$$

The event $U(i, t_0) > x, X(t_0) = i$ implies in particular $X(t_0 + x) = i$. Since we have a M.c. the trajectory of $X(t)$ for $t \geq t_0 + x$ depends only on the state at time $t_0 + x$ which is i in our case. Namely

$$\mathbb{P}(U(i, t_0) > x + y | U(i, t_0) > x, X(t_0) = i) = \mathbb{P}(U(i, t_0 + x) > y | X(t_0 + x) = i).$$

But the latter expression by homogeneity is $\mathbb{P}(U(i, t_0) > y | X(t_0) = i)$, as it is the probability of the holding time being larger than y when the current state is i . We conclude that

$$\mathbb{P}(U(i, t_0) > x + y | U(i, t_0) > x, X(t_0) = i) = \mathbb{P}(U(i, t_0) > y | X(t_0) = i),$$

namely

$$\mathbb{P}(U(i, t_0) > x + y | X(t_0) = i) = \mathbb{P}(U(i, t_0) > y | X(t_0) = i) \mathbb{P}(U(i, t_0) > x | X(t_0) = i).$$

Since the exponential function is the only one satisfying this property, then $U(i, t_0)$ must be exponentially distributed. \square

There is an omitted subtlety in the proof. We assumed that for every $t, z > 0$ and state i , $\mathbb{P}(X(t+s) = i, \forall s \in [0, z] | X(t) = i, \mathfrak{I}_t) = \mathbb{P}(X(t+s) = i, \forall s \in$

$[0, z] | X(t) = i$ where \mathfrak{S}_t denotes the history of the process up to time t . We deduced this based on the assumption (2). This requires a technical proof, which we ignored above.

Thus the evolution of a continuous M.c. $X(t)$ can be described as follows. It stays in a given state i during some exponentially distributed time U_i , with parameter μ_i which only depends on the state. After this time it makes a transition to the next state j . If we consider the process only at the random times of transitions, denoted say by $t_1 < t_2 < \dots$, then we obtain an *embedded* discrete time process $Y_n = X(t_n)$. It is an exercise to show that Y_n is in fact a homogeneous Markov chain. Denote the transition rates of this Markov chain by $p_{i,j}$. The value $q_{i,j} = \mu_i p_{i,j}$ is called “transition rate” from state i to state j . Note, that the values $p_{i,j}$ were introduced only for $j \neq i$, as they were derived from M.c. changing its state. Define $q_{i,i} = -\sum_{j \neq i} q_{i,j}$. The matrix $G = (q_{i,j})$, $i, j \in \mathcal{X}$ is defined to be the **generator** of the M.c. $X(t)$ and plays an important role, specifically for the discussion of a stationary distribution.

A stationary distribution π of a continuous M.c. is defined in the same way as for the discrete time case: it is the distribution which is time invariant. The following fact can be established.

Proposition 2. A vector $(\pi_i), i \in \mathcal{X}$ is a stationary distribution iff $\pi_i \geq 0, \sum_i \pi_i = 1$ and $\sum_j \pi_j q_{j,i} = 0$ for every state i . In vector form $\pi^T G = 0$.

As for the discrete time case, the theory of continuous time M.c. has a lot of special structure when the state space is finite. We now summarize without proofs some of the basic results. First there always exists a stationary distribution. The condition for uniqueness of the stationary distribution is the same - single recurrence class, with communications between the states defined similarly. A nice “advantage” of continuous M.c. is the lack of periodicity. There is no notion of a period of a state. Moreover, and most importantly, suppose the chain has a unique recurrence class. Then, letting π denote the corresponding unique stationary distribution, the mixing property

$$\lim_{t \rightarrow \infty} p_{i,j}^{(t)} = \pi_j$$

holds for all states i, j . For the modeling purposes, it is useful sometimes to consider a continuous as opposed to a discrete M.c.

There is an alternative way to describe a continuous M.c. and the embedded discrete time M.c. Assume that to each pair of states i, j we associate an exponential “clock” - exponentially distributed r.v. $U_{i,j}$ with parameter $\mu_{i,j}$. Each

time the process jumps into i all of the clocks turned on simultaneously. Then at time $U_i \triangleq \min U_{i,j}$ the process jumps into state $j = \arg \min_j U_{i,j}$. It is not hard to establish the following: the resulting process is a continuous time finite state M.c. The embedded discrete time M.c. has then transition probabilities $\mathbb{P}(X(t_{n+1}) = j | X(t_n) = i) = \frac{\mu_{i,j}}{\sum_k \mu_{i,k}}$, as the probability that $U_{i,j} = \min_k U_{i,k}$ is given by this expression, when the distribution of $U_{i,j}$ is exponential with parameters $\mu_{i,j}$. The holding time has then the distribution $\text{Exp}(\mu_i)$ where $\mu_i = \sum_k \mu_{i,k}$. Thus we obtain an alternative description of a M.c. The transition rates of this M.c. are $q_{i,j} = \mu_i p_{i,j} = \mu_{i,j}$. In other words, we described the M.c. via the rates $q_{i,j}$ as given.

This description extends to the infinite M.c., when the notion of holding times is well defined (see the comments above).

References

- [1] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, Oxford University Press, 3rd edition, 2001.
- [2] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to probability*, Athena Scientific, 2002.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436/15.085 Lecture 25

Lecturer: Yury Polyanskiy

Scribe: MIT Class Participants

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They are posted to serve class purposes.*

MARTINGALES I

Content.

0. Background
1. Martingales: definition, examples
2. Azuma-Hoeffding Inequality
3. Optional stopping theorem

0 BACKGROUND

- Recall we defined conditional expectation $V = \mathbb{E}[A|X]$ as follows:

$$\forall B = f(X), \mathbb{E}[AB] = \mathbb{E}[VB]$$

We also learned that one computes conditional expectations, usually, by integrating

$$\mathbb{E}[A|X = x] = \int_{\mathbb{R}} a P_{A|X}(da|x)$$

- We can also define conditional expectation with respect to a sigma-algebra \mathcal{F} :

$$V = \mathbb{E}[A|\mathcal{F}]$$

Namely, random variable V is a conditional expectation of A given \mathcal{F} if
a) $V \in \mathcal{F}$ and b) $\forall B \in \mathcal{F}, \mathbb{E}[AB] = \mathbb{E}[VB]$. Here we used common abuse of notation $V \in \mathcal{F}$ meaning “ V is \mathcal{F} -measurable” (which, recall, means $\{V \leq v\} \in \mathcal{F}$ for every $v \in \mathbb{R}$).

- Recall $\sigma(X_0, X_1, \dots, X_k) = \mathcal{F}_k$ where \mathcal{F}_k is the smallest σ -algebra containing all events $\{X_i \leq a\}$. Recall also that

$$A \in \mathcal{F}_{\parallel} \iff \exists f : A = f(X_0, \dots, X_k) \quad (1)$$

Given a stochastic process X_0, X_1, \dots

$$\mathcal{F}_k = \sigma(X_0, X_1, \dots, X_k), \mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots \subset \mathcal{F}_{\infty}$$

where $\mathcal{F}_{\infty} = \sigma(X_i, i \in \mathbb{Z}_+)$. The \mathcal{F}_k we have defined here is known as the **standard filtration** generated by the stochastic process. We can think about each \mathcal{F}_k as the valid questions you can ask (and answer) if you only know realization of the stochastic process up to time k .

- Before we were talking about a stochastic process in isolation. Now we will talk about stochastic process being adapted to some filtration \mathcal{F}_k . For simplicity, you can always think of a standard filtration generated by a mother (complicated) random process $\{X_k\}$. We say that Y_k is a stochastic process adapted to filtration \mathcal{F}_k if $Y_k \in \mathcal{F}_k$ holds $\forall k \geq 0$.

As an example, we can look at a simple process $Y_k = \text{sign}(X_0 + \dots + X_k)$. Note that $Y_k \in \mathcal{F}_k$, i.e. Y_k is \mathcal{F}_k -measurable, because it is a function of X_0, \dots, X_k . However, knowledge of Y_0, \dots, Y_k is insufficient to reconstruct trajectory X_0, \dots, X_k . So while Y_k is adapted to \mathcal{F}_k , the filtration \mathcal{F}_k is much richer. This is a common situation in applications (since we are interested in functions of the mother process), and that's why we need the concept of filtration.

1 MARTINGALES

1.1 Definition

We introduce our main definition of a Martingale:

Definition 1. A process $(M_t, t = 0, \dots, \infty)$ is a **martingale** with respect to filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots$ if:

1. $M_t \in \mathcal{F}_t \forall t \geq 0$
2. $\mathbb{E}|M_t| < \infty$, i.e. M_t is integrable
3. $\mathbb{E}[M_t | \mathcal{F}_{t-1}] = M_{t-1}$

In the special case if $\mathcal{F}_t = \sigma(M_0, \dots, M_t)$ we simply say “ M_t is a martingale” (without mentioning filtration). In this case, the property 1) is automatic and being a martingale becomes essentially just the requirement that $\mathbb{E}[M_t | M_0, \dots, M_{t-1}] = M_{t-1}$, for all $t \geq 1$.

1.2 History of Martingales

The word *martingale* comes from gambling. It describes a strategy in which a gambler makes a series of bets. For each bet, he wins if a coin lands on heads and loses if the coin lands on tails. For each successive loss, he doubles his bet, starting with \$1 on the first flip. At the time of winning (i.e. first time the coin lands on heads), the gambler will receive a net gain of $\$1 \cdot 2^{t+1} - (\$1 + \$2 + \dots + \$2^t) = \$1$; however the expected loss at the time of winning is ∞ .

1.3 Examples

For the following examples of martingales, we introduce the notation

$$\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot | \mathcal{F}_t]$$

Example 1. $S_n = X_0 + \dots + X_n$ for X_i independent and $\mathbb{E}[X_i] = 0$. $\mathcal{F}_k = \sigma(X_0, \dots, X_k)$. Then, we have $\mathbb{E}_{n-1}S_n = S_{n-1}$.

Example 2. $Y_n = X_0 \cdot X_1 \cdot \dots \cdot X_n$ for X_i independent and $\mathbb{E}[X_i] = 1$. Then, we have $\mathbb{E}_{n-1}Y_n = Y_{n-1}$.

Example 3 (Doob Martingale). Let Z be any random variable with finite expectation ($\mathbb{E}|Z| < \infty$) and \mathcal{F}_t be any filtration. We define a Doob martingale:

$$M_t = \mathbb{E}[Z | \mathcal{F}_t] \tag{2}$$

This make look like a rather special case, but it will turn out that many martingales we work with will turn out to be of this type. Think of it as if we have a “secret” Z and we are observing its average given the known information at time n . Over time, we learn more and more about this mother random variable, and approach knowing Z itself. A Doob martingale has the martingale property with respect to the given filtration:

$$\begin{aligned} \mathbb{E}_{t-1}M_t &= \mathbb{E}[M_t | X_0, \dots, X_{t-1}] \\ &= \mathbb{E}[\mathbb{E}[Z | X_0, \dots, X_t] | X_0, \dots, X_{t-1}] \\ &= \mathbb{E}[Z | X_0, \dots, X_{t-1}] \\ &= M_{t-1} \end{aligned}$$

where in the second line, we use the tower property of conditional expectation.

As we will see below, Examples 1 and 2 are not Doob martingales unless they converge. We can, however, modify examples 1 and 2 so that they are Doob martingales. Suppose we restrict the martingales to within a certain window, for instance S_n for n such that $-100 \leq S_n \leq 100$, and freeze the process once it exceeds the boundary. Then, the martingales are Doob martingales (since they are bounded!).

2 AZUMA'S INEQUALITY

By performing a simple computation and induction, we can see that $\mathbb{E}M_t = \mathbb{E}M_0$:

$$\begin{aligned}\mathbb{E}[M_t] &= \mathbb{E}[\mathbb{E}_{t-1}[M_t]] \\ &= \mathbb{E}[M_{t-1}]\end{aligned}$$

To compute the variance of M_t , assume without loss of generality that $\mathbb{E}[M] = 0$.

$$\text{var}(M_t) = \mathbb{E}M_t^2 \tag{3}$$

$$= \mathbb{E}[M_t - M_{t-1} + M_{t-1}]^2 \tag{4}$$

$$= \mathbb{E}[\mathbb{E}_{t-1}[(M_t - M_{t-1})^2 + M_{t-1}^2 + 2M_{t-1}(M_t - M_{t-1})]] \tag{5}$$

$$= \mathbb{E}[M_{t-1}^2] + \mathbb{E}[M_t - M_{t-1}]^2 \tag{6}$$

$$= \sum_{s=1}^t \mathbb{E}[M_s - M_{s-1}]^2 + \text{var}(M_0) \tag{7}$$

We obtain (5) by using the tower property of conditional expectation and we use the following simplification to obtain (6):

$$\mathbb{E}_{t-1}[M_{t-1}(M_t - M_{t-1})] = M_{t-1}(\mathbb{E}[M_t - M_{t-1}]) = 0$$

Finally, we obtain (7) by induction.

From this derivation, we can see that $|M_s - M_{s-1}| \leq c \Rightarrow \text{var}M_t \leq c^2 t$. Martingales with bounded increments (within a constant c) grow with speed $\sim \sqrt{t}$. This leads us to the Azuma-Hoeffding inequality.

Theorem 1 (Azuma-Hoeffding Inequality). *If M_t is a martingale with $|M_t - M_{t-1}| \leq c_t$ a.s. $\forall t$, then*

$$\mathbb{P}(M_t - \mathbb{E}[M_t] > h) \leq \exp\left(\frac{-h^2}{2 \sum_{s=1}^t c_s^2}\right)$$

Proof. From Chernoff bound, we have $\mathbb{P}[\cdot] \leq e^{-\lambda h + \psi_t(\lambda)}$ $\forall \lambda > 0$ where $\psi_t(\lambda) = \ln \mathbb{E}[e^{\lambda M_t}]$ is the log MGF. Without loss of generality, we are assuming $\mathbb{E}M = 0$.

It is sufficient to prove that $\psi_t(\lambda) \leq \psi_{t-1}(\lambda) + \frac{\lambda^2 c_t^2}{2}$.

We can rewrite the following expression:

$$\mathbb{E}_{t-1} e^{\lambda M_t} = \mathbb{E}_{t-1} e^{\lambda(M_t - M_{t-1})} e^{\lambda M_{t-1}}$$

$\forall |x| \leq c_t$:

$$e^{\lambda x} \leq e^{-\lambda c_t} + (x + c_t) \frac{(e^{\lambda c_t} - e^{-\lambda c_t})}{2c_t}$$

Plugging in $x = M_t - M_{t-1}$.

$$\mathbb{E}_{t-1} e^{\lambda(M_t - M_{t-1})} \leq \frac{e^{-\lambda c_t} + e^{\lambda c_t}}{2}$$

because $\mathbb{E}_{t-1}(M_t - M_{t-1}) = 0$. Finally, using the fact that

$$\frac{e^{-p} + e^p}{2} \leq e^{\frac{p^2}{2}},$$

which can be checked using Matlab/Python, and substituting $p = \lambda c_t$, we get

$$\mathbb{E}_{t-1} e^{\lambda(M_t - M_{t-1})} \leq \frac{e^{-\lambda c_t} + e^{\lambda c_t}}{2} \leq e^{\frac{\lambda^2 c_t^2}{2}}$$

□

Example 4. Suppose we throw M balls into n bins. Let V be the number of occupied bins. Let us define a process:

$$M_t \triangleq \mathbb{E}[V | X_1, \dots, X_t] \tag{8}$$

where X_i is the bin selected by the i^{th} ball. We can see that this process is a Doob martingale because we are conditioning on increasing σ -algebras. Intuitively, we can see that at any step of the process, the conditional expectation

will not change by more than 1: $|M_t - M_{t-1}| \leq 1$. Thus, we have by Azuma-Hoeffding:

$$\mathbb{P}[|V - \mathbb{E}V| > h] \leq e^{-\frac{h^2}{2M}}$$

This example demonstrates that, even if the exact mean is unknown, we can already guarantee that the distribution of V concentrates sharply around its mean. So all complexity of understanding V boils down to computing its expectation (which, in turn, can be done by sampling a few realizations, thanks to the concentration phenomenon).

3 OPTIONAL STOPPING THEOREM

Recall the following definition of the stopping time of a filtration:

Definition 2. $\tau : \Omega \rightarrow \mathbb{Z}_+ \cup \{+\infty\}$ is a **stopping time of filtration** $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots$ if

$$\{\tau \leq n\} \in \mathcal{F}_n \quad \forall n \iff \{\tau = n\} \in \mathcal{F}_n \quad \forall n$$

Now, using the notation $a \wedge b = \min(a, b)$, let us define $Y_t = M_{\tau \wedge t}$. Think of this as a process that has values M_t until time τ and then has constant value M_τ . We are essentially defining a new process that follows the trajectory of M_t but then freezes once it reaches τ . For instance, we could define $\tau = \inf\{t : |M_t| \geq 100\}$.

Theorem 2. $Y_t = M_{\tau \wedge t}$ is a martingale for any martingale M_t and stopping time τ .

Proof.

$$Y_t = \sum_{r=0}^{t-1} M_r \mathbf{1}\{\tau = r\} + M_t \mathbf{1}\{\tau \geq t\}$$

Note that $\{\tau \geq t\} = \{\tau \leq t-1\}^c \in \mathcal{F}_{t-1}$, so $\mathbb{E}_{t-1} \sum_{r=0}^{t-1} M_r \mathbf{1}\{\tau = r\} = \sum_{r=0}^{t-1} M_r \mathbf{1}\{\tau = r\}$. Substituting this, we get

$$\begin{aligned}
\mathbb{E}_{t-1} Y_t &= \sum_{r=0}^{t-1} M_t \mathbb{1}\{\tau = r\} + (\mathbb{E}_{t-1} M_t) \mathbb{1}\{\tau \geq t\} \\
&= \sum_{r=0}^{t-1} M_t \mathbb{1}\{\tau = r\} + M_{t-1} \mathbb{1}\{\tau \geq t\} \\
&= M_{(t-1) \wedge \tau}
\end{aligned}$$

where the equation from the first to the second line follows from $\mathbb{E}_{t-1} M_t = M_{t-1}$. We are done with the proof. \square

This relates to the efficient market hypothesis: the price of a stock should be a martingale (with respect to filtration generated by all public information). Indeed, in this case defining a smart stopping time one is unable to improve the average price still.

Now we return to the idea of uniform integrability and introduce the crucial concept of a **uniformly integrable martingale (uim)**. First we consider a very useful and simple criterion for getting a wealth of uims.

Proposition 1. *Let M_t be a martingale, τ be a stopping time such that $\mathbb{E}\tau < \infty$, and $\mathbb{E}_{t-1}|M_t - M_{t-1}| \leq c$ a.s., then $Y_t \triangleq M_{t \wedge \tau}$ is a **uniformly integrable Martingale**.*

Proof.

$$\begin{aligned}
|Y_t - Y_0| &\leq \sum_{s=1}^t |Y_s - Y_{s-1}| \\
&= \sum_{s=1}^t |M_s - M_{s-1}| \mathbb{1}\{\tau \geq s\} \\
&\leq \sum_{s=1}^{\infty} |M_s - M_{s-1}| \mathbb{1}\{\tau \geq s\} =: W
\end{aligned}$$

The conditions imply

$$\mathbb{E}W < \infty \Rightarrow |Y_t| \leq W + |Y_0| \quad \forall t$$

\square

We will see in the next lecture that every u.i.m. \iff Doob Martingale. In particular, every bounded martingale is Doob. For now we state the crown jewel of martingale theory:

Theorem 3 (Optional stopping theorem). *Let M_t be a uniformly integrable Martingale and let τ be a stopping time such that $\mathbb{P}[\tau < \infty] = 1$. Then:*

$$\mathbb{E}M_\tau = \mathbb{E}M_0$$

Proof. We first prove a special case: Suppose $\tau \leq L$ a.s. where L is some constant. Then:

$$M_\tau = \sum_{t=0}^L M_t \mathbb{1}\{\tau = t\} \quad (9)$$

$$= \sum_{t=0}^L (\mathbb{E}_t M_L) \mathbb{1}\{\tau = t\} \quad (10)$$

$$= \sum_{t=0}^L \mathbb{E}_t M_L \mathbb{1}\{\tau = t\} \quad (11)$$

The key insight to obtain (10) was to use the property of martingales from part 3 of the definition. Now, we can take the expected value of both sides of (11):

$$\mathbb{E}M_\tau = \sum_t \mathbb{E}M_L \mathbb{1}\{\tau = t\} \quad (12)$$

$$= \mathbb{E}M_L \quad (13)$$

$$= \mathbb{E}M_0 \quad (14)$$

Note that (12) forms a partition because $\sum_t \mathbb{1}\{\tau = t\} = 1$ a.s.

Note that a similar argument shows

$$\mathbb{E}|M_\tau| \leq \mathbb{E}|M_L|$$

Indeed, one only needs to notice that $|\mathbb{E}_t M_L| \leq \mathbb{E}_t |M_L|$.

The general case follows in two steps. First define $\tau_L = \tau \wedge L$, then by the previous argument we have

$$\mathbb{E}|M_{\tau_L}| \leq \sup_L \mathbb{E}|M_L| < \infty,$$

where the last inequality follows from uniform integrability (which implies uniform boundedness). So since $M_{\tau_L} \rightarrow M_\tau$ as $L \rightarrow \infty$ almost surely, we have via Fatou's lemma

$$\mathbb{E}|M_\tau| < \infty.$$

Finally,

$$M_\tau = M_{\tau_L} + (M_\tau - M_{\tau_L})1\{\tau \geq L\}$$

By the first part of the proof $\mathbb{E}[M_{\tau_L}] = \mathbb{E}[M_0]$. So we only need to show that as $L \rightarrow \infty$ the expectation of the second term vanishes.

Note that as $L \rightarrow \infty$ we have $\mathbb{P}[\tau \geq L] \rightarrow 0$. Thus $\mathbb{E}[|M_\tau|1\{\tau \geq L\}] \rightarrow 0$. Similarly, from uniform integrability of $\{M_L\}$ we have $\mathbb{E}[|M_{\tau_L}|1\{\tau \geq L\}] = \mathbb{E}[|M_L|1\{\tau \geq L\}] \rightarrow 0$. This completes the proof. \square

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436/15.085J Lecture 26

Lecturer: Yury Polyanskiy

Scribe: MIT Class Participants

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They are posted to serve class purposes.*

Martingales II

Content.

1. Review
2. Some applications of Optional Stopping Theorem
3. Martingale Convergence Theorem

1 Review

Definition 1 (Martingale). $\{M_t\}$ is a martingale with respect to $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots \subset \mathcal{F}$ if it satisfies:

1. $M_t \in \mathcal{F}_t, \quad t \geq 0$
2. $\mathbb{E}|M_t| < \infty, \quad t \geq 0$
3. $\mathbb{E}[M_t | \mathcal{F}_{t-1}] = M_{t-1}, \quad t \geq 1$

In other words, $\{M_t\}$ is a martingale w.r.t. $\{X_t\}$ if:

1. $M_t = f(X_0, \dots, X_t), \quad t \geq 0$
2. $\mathbb{E}|M_t| < \infty, \quad t \geq 0$
3. $\mathbb{E}[M_t | X_0, \dots, X_{t-1}] = M_{t-1}, \quad t \geq 1.$

For convenience, we denote it by

$$\mathbb{E}_s = \mathbb{E}[\cdot | \mathcal{F}_s] = \mathbb{E}[\cdot | X_0, \dots, X_s],$$

and the third condition can be written as: $\mathbb{E}_{t-1} M_t = M_{t-1}$ for any $t \geq 1$.

When we say $\{M_t\}$ is a martingale without specifying the filtration, we mean that it is a martingale with respect to its natural filtration, i.e. $\mathcal{F}_t = \sigma(M_0, \dots, M_t)$. We consider it as a special case of the definition.

Now if $\{M_t\}$ is a martingale with respect to the filtration $\mathcal{F}_t = \sigma(X_0, \dots, X_t)$, it is also a martingale with respect to its filtration $\sigma(M_0, \dots, M_t)$. In fact, by the tower property, we have:

$$\mathbb{E}[M_t | M_0, \dots, M_{t-1}] = \mathbb{E}[\mathbb{E}[M_t | \mathcal{F}_t] | M_0, \dots, M_{t-1}] = \mathbb{E}[M_{t-1} | M_0, \dots, M_{t-1}] = M_{t-1}.$$

Properties of a martingale $\{M_t\}$

- $\mathbb{E}_s M_t = M_{t \wedge s}$
- $\mathbb{E} M_t = \mathbb{E} M_0, \forall t \geq 0$
- If M_t is a martingale, and τ is a stopping time, then $Y_t = M_{t \wedge \tau}$ is a martingale.
- Side note: \mathbb{E}_n also works like a martingale: $\mathbb{E}_n \mathbb{E}_m = \mathbb{E}_{n \wedge m}$, $\mathbb{E} \mathbb{E}_n = \mathbb{E}_n \mathbb{E} = \mathbb{E}$. In fact, you can also define \mathbb{E}_τ and even have $\mathbb{E}_\tau M_t = M_{\tau \wedge t}$. But we won't do it in this class.

Let A_t be the gambler's ruin Markov chain starting from k . Now let's consider the simple random walk S_t starting from $S_0 = k$, and $S_t = S_{t-1} + X_t$ with $\mathbb{P}(X_t = \pm 1) = \frac{1}{2}$. Let,

$$\tau = \inf\{t : S_t = 0 \text{ or } S_t = n\}.$$

Notice that $A_t = S_{t \wedge \tau}$. Since S_t is a martingale, it follows that its stopped martingale A_t is a martingale as well. This implies a good property of gambler's ruin Markov chain, which is

$$\mathbb{E} A_t = A_0 = k.$$

We will use the definitions and notations of S_t and A_t for several times in this lecture.

Theorem 1 (Optional Stopping Theorem). *If $\{M_t\}$ is a martingale and τ is a stopping time such that $\{M_t\}$ is uniformly integrable and $\mathbb{P}(\tau < \infty) = 1$, then*

$$\mathbb{E}M_\tau = \mathbb{E}M_0.$$

Proposition 1 (Uniformly integrable martingales). *The following propositions about uniformly integrable martingales (u.i.M.) hold:*

1. $M_t = \mathbb{E}[Z|\mathcal{F}_t]$ for any Z such that $\mathbb{E}|Z| < \infty$ is always u.i.M.
2. If there exists $G(t)$ such that $G(t)/t \rightarrow \infty$ as $t \rightarrow \infty$. If $\sup_t \mathbb{E}[G(M_t)] < \infty$, then M_t is u.i.M.
3. If $|M_t - M_{t-1}| \leq c < \infty$ and $\mathbb{E}\tau < \infty$, then $Y_t = M_{t \wedge \tau}$ is u.i.M.

$\{S_n\}$ is not uniformly integrable. Indeed, the magnitude of $|S_n|$ is approximately $O(\sqrt{n})$. For any b , one can always find some $N \approx b^2$ such that $\mathbb{E}[|S_N|1\{|S_N| \geq b\}] \geq c$, so $\sup_n \mathbb{E}[|S_n|1\{|S_n| \geq b\}] \not\rightarrow 0$ as $b \rightarrow \infty$. Hence, S_n is not uniformly integrable.

2 Some applications of O.S.T.

2.1 Gambler's Ruin

For the **gambler's ruin** problem, we start with $A_0 = k$, and we want to find $\mathbb{P}[win] = \mathbb{P}[A_\infty = n]$.

Note that $A_t = S_{t \wedge \tau}$ is a u.i.M (since A_t is bounded), therefore

$$n\mathbb{P}[win] = \mathbb{E}A_\tau = \mathbb{E}A_0 = k,$$

as

$$A_\tau = \begin{cases} 0, & \text{"ruined"} \\ n, & \text{"won"} \end{cases}.$$

Therefore, $\mathbb{P}[win] = \frac{k}{n}$.

Now let $M_t = S_t^2 - t = (S_{t-1} + X_t)^2 - t = S_{t-1}^2 + 2X_t S_{t-1} + X_t^2 - t = S_{t-1}^2 - (t-1) + 2X_t S_{t-1} = M_{t-1} + 2X_t S_{t-1}$. Therefore,

$$\mathbb{E}_{t-1} M_t = M_{t-1}.$$

$M_{t \wedge \tau}$ is uniformly integrable since the increment $|M_t - M_{t-1}| = 2|X_t S_{t-1}|$ is bounded.

Therefore, by OST, we have

$$\frac{k}{n} \cdot n^2 - \mathbb{E}\tau = \mathbb{E}[M_\tau] = M_0 = k^2,$$

and thus, $\mathbb{E}\tau = k(n - k)$.

2.2 Null recurrence of S_t

We start with $S_0 = k$. Let $\tau_1 = \inf\{t : S_t = 0\}$, and $B_t = S_{t \wedge \tau_1}$. One can think of B_t as a Markov chain with 0 the absorbing state.

We know from recurrence of S_t that $\tau_1 < \infty$ a.s.. We also know that $\mathbb{E}S_t = \mathbb{E}B_t = \mathbb{E}S_0 = k$.

If B_t were a u.i.M, then OST applies, we will have $\mathbb{E}B_{\tau_1} = k$. However, by definition, $B_{\tau_1} = 0$ a.s., so $\mathbb{E}B_{\tau_1} = 0 \neq k$. By Proposition 2(3), the only thing that prevents B_t from being a u.i.M is $\mathbb{E}\tau = \infty$. Therefore, S_t is null recurrent.

2.3 Gambler's Ruin in the asymmetric case

For the asymmetric case, i.e. $S_t = S_{t-1} + X_t$ with $\mathbb{P}(X_t = 1) = p$ and $\mathbb{P}(X_t = -1) = 1 - p$, one can use the following two martingales to compute $\mathbb{P}[\text{win}]$ and $\mathbb{E}\tau$:

1. $M_t = S_t - (2p - 1)t$
2. $N_t = e^{\lambda S_t - t\psi_X(\lambda)}$, where $\psi_{X_1}(\lambda) = \ln M_{X_1}(\lambda)$

From OST, we have

$$\mathbb{E}S_\tau - (2p - 1)\mathbb{E}\tau = \mathbb{E}M_\tau = M_0 = k$$

and

$$e^{\lambda n} \mathbb{P}[\text{win}] + \mathbb{P}[\text{ruined}] = \mathbb{E}N_\tau = N_0 = e^{\lambda k},$$

with some $\lambda (= \ln \frac{p}{1-p})$ such that $\psi_{X_1}(\lambda) = 0$.

3 Martingale Convergence Theorem

Think of M_t as the price of stock. At time $t - 1$, you decide to move your possession of stock to F_t shares, where $F_t \in \mathcal{F}_{t-1}$ is determined by all the

observed information at time $t - 1$. Then the value of your portfolio at time t is

$$V_t = F_0 M_0 + F_1(M_1 - M_0) + \dots + F_t(M_t - M_{t-1}) \stackrel{\Delta}{=} \int_0^t F dM.$$

Proposition 2. *If M_t is a martingale, then V_t is a martingale. In particular, $\mathbb{E}V_t = \mathbb{E}V_0$.*

The important consequence is that if you start with F_0 shares priced at M_0 then no trading strategy (and no finite cash-out time) can yield an expectation different from what you had $\mathbb{E}[F_0 M_0]$ in the beginning. Assuming the market price is a martingale with respect to the same filtration \mathcal{F}_t that determines the available information you have to execute the trading decisions.

Definition 2. Starting $S_0 = 0$, define $T_k = \inf\{t \geq S_{k-1} : M_t \leq a\}$, $S_k = \inf\{t \geq T_k : M_t \geq b\}$. Define $U_n(a, b) = \# \text{ of upcrossings of } (a, b) \text{ in } 0 \leq t \leq n$, i.e.

$$U_n(a, b) = \sup\{k : S_k \leq n\}.$$

Lemma 1 (Upcrossing Lemma).

$$\mathbb{E}[U_n(a, b)] \leq \frac{\mathbb{E}(M_n - a)_-}{b - a}.$$

Proof. Starting with $F_0 = 0$ and do trading: buy 1 share when $M_t \leq a$ and sell it when $M_t \geq b$. Since $V_0 = 0$, we have

$$V_n \geq (b - a)U_n + (M_n - a) \wedge 0 = (b - a)U_n - (M_n - a)_-.$$

Since V_n is a martingale, it follows from Optional Stopping Theorem that

$$\mathbb{E}U_n \leq \frac{\mathbb{E}(M_n - a)_-}{b - a}.$$

□

Theorem 2. If M_n is a martingale such that $\mu = \sup_n \mathbb{E}|M_n| < \infty$, then there exists an integrable random variable M_∞ such that

$$M_n \xrightarrow{\text{a.s.}} M_\infty, \quad \text{and} \quad \mathbb{E}[|M_\infty|] \leq \mu < \infty.$$

If M_t is u.i.M, then $M_t \xrightarrow{L^1} M_\infty$ and

$$M_t = \mathbb{E}[M_\infty | \mathcal{F}_t].$$

Remark: Note that if M_t is u.i.M. then $\mu < \infty$ automatically. Thus, the second part of the theorem shows that every u.i.M. is in fact a Doob martingale.

Proof. **Proof of part 1:** Fix $b > a$,

$$U(a, b) = \lim_{n \rightarrow \infty} U_n(a, b).$$

By the **upcrossing lemma**, we have

$$\mathbb{E}U_n(a, b) \leq \frac{\mathbb{E}(M_n - a)_-}{b - a} \leq \sup_n \frac{\mathbb{E}|M_n| + |a|}{b - a} < \infty.$$

Therefore, by **Monotone Convergence Theorem**, we have

$$\mathbb{E}U(a, b) = \lim_{n \rightarrow \infty} \mathbb{E}U_n(a, b) \leq \sup_n \frac{\mathbb{E}|M_n| + |a|}{b - a} < \infty.$$

This imples that,

$$\mathbb{P}(U(a, b) = \infty \text{ for any } b > a, a, b \in \mathbb{Q}) = 0.$$

So with probability 1 the trajectory M_n intersects any arbitrary small interval only finitely many times. Thus there must exist a (possibly extended real-valued) random variable M_∞ such that $M_n \xrightarrow{\text{a.s.}} M_\infty$.

To show that M_∞ is in fact integrable (and hence real-valued) we use Fatou's lemma:

$$\mathbb{E}[|M_\infty|] = \mathbb{E}[\liminf_{n \rightarrow \infty} |M_n|] \leq \liminf_{n \rightarrow \infty} \mathbb{E}[|M_n|] \leq \mu < \infty.$$

Proof of part 2: To show $M_t = \mathbb{E}[M_\infty | \mathcal{F}_t]$, it suffices to show that for any $B \in \mathcal{F}_t$, we have

$$\mathbb{E}M_\infty 1_B = \mathbb{E}M_t 1_B.$$

For any $m \geq t$, we have

$$\mathbb{E}M_m 1_B = \mathbb{E}[\mathbb{E}_t[M_m 1_B]] = \mathbb{E}[1_B M_t].$$

Since $M_m 1_B \xrightarrow{a.s.} M_\infty 1_B$ and $\{M_m 1_B\}$ is uniformly integrable, it follows that $M_m 1_B \xrightarrow{L_1} M_\infty 1_B$. Therefore,

$$\mathbb{E}[M_t 1_B] = \lim_{m \rightarrow \infty} \mathbb{E}[M_m 1_B] = \mathbb{E}[M_\infty 1_B].$$

□

Corollary 1. If $M_n \geq 0$, M_n is a martingale, then it converges almost surely to integrable M_∞ .

Proof. Since for any n ,

$$\mathbb{E}|M_n| = \mathbb{E}M_n = \mathbb{E}M_0,$$

it follows that

$$\sup_n \mathbb{E}|M_n| < \infty.$$

□

In particular, $M_n = X_1 \dots X_n$ such that $X_n \geq 0$, $\mathbb{E}X_n = 1$. Then, M_n converges almost surely.

4 Further topics

Martingale and stopping time theory is rich subject. The key omissions are:

- A lot of results about martingales are also available for submartingales (i.e. when $\mathbb{E}_{t-1}[M_t] \geq M_{t-1}$) and supermartingales (i.e. when $\mathbb{E}_{t-1}[M_t] \leq M_{t-1}$).
- Maximal inequalities for martingales/submartingales/supermartingales). These establish results similar to Kolmogorov's maximal inequalities (for sums of independent r.v.s) but for general martingales. To get a flavor of such results, if $M_0 = 0$ then

$$\mathbb{P}\left[\max_{0 \leq t \leq n} M_t > b\right] = \mathbb{P}[U_n(0, b) \geq 1] \leq \frac{1}{b} \mathbb{E}[|M_n|],$$

where in the last step we applied the upcrossing Lemma and Markov's inequality. So in particular, in the setting of convergence theorem we see that life-time maximum of M_t is of the order of μ . Other maximal inequalities bound p -th norm of the maximum in terms of the p -th norm of M_n etc.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Recitation 1

Fall 2018

BACKGROUND MATERIAL ON SETS AND REAL ANALYSIS

1 SETS

A **set** is a collection of objects, which are the **elements** of the set. If A is a set and x is an element of A , we write $x \in A$. If x is not an element of A , we write $x \notin A$. A set can have no elements, in which case it is called the **empty set**, denoted by \emptyset .

Sets can be specified in a variety of ways. If A contains a finite number of elements, say x_1, x_2, \dots, x_n , we write it as a list of the elements, in braces:

$$A = \{x_1, x_2, \dots, x_n\}.$$

For example, the set of possible outcomes of a die roll is $\{1, 2, 3, 4, 5, 6\}$, and the set of possible outcomes of a coin toss is $\{H, T\}$, where H stands for “heads” and T stands for “tails.”

More generally, we can consider the set of all x that have a certain property P , and denote it by

$$\{x \mid x \text{ satisfies } P\}.$$

(The symbol “ \mid ” is to be read as “such that.”) For example, the set of even integers can be written as $\{k \mid k/2 \text{ is integral}\}$. Similarly, the set of all real numbers x in the interval $[0, 1]$ can be written as $\{x \mid 0 \leq x \leq 1\}$.

If A contains infinitely many elements x_1, x_2, \dots , that can be enumerated in a list (so that the elements are in a one-to-one correspondence with the positive integers), we write

$$A = \{x_1, x_2, \dots\},$$

and we say that A is **countably infinite**. For example, the set of even integers can be written as $\{0, 2, -2, 4, -4, \dots\}$, and is countably infinite. The term **countable** is sometimes used to refer to a set which is either finite or countably infinite. A set which is not countable is said to be **uncountable**.

If every element of a set A is also an element of a set B , we say that A is a **subset** of B , and we write $A \subset B$ or $B \supset A$. If $A \subset B$ and $A \subset B$, the two

sets are **equal**, and we write $A = B$.¹ It is sometimes expedient to introduce a **universal set**, denoted by Ω , which contains all objects that could conceivably be of interest in a particular context. Having specified a context in terms of a universal set Ω , one then only considers sets A that are subsets of Ω .

2 SET OPERATIONS

The **complement** of a set A , with respect to a universal set Ω , is the set $\{x \in \Omega \mid x \notin A\}$ of all elements of Ω that do not belong to A , and is denoted by A^c . Note that $\Omega^c = \emptyset$.

The **union** of two sets A and B is the set of all elements that belong to A or B (or both), and is denoted by $A \cup B$. The **intersection** of two sets A and B is the set of all elements that belong to both A and B , and is denoted by $A \cap B$. Thus,

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\},$$

and

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

We also define

$$A \setminus B = A \cap B^c = \{x \mid x \in A \text{ and } x \notin B\},$$

which is the set of all elements that belong to A but not in B .

We will often deal with the union or the intersection of several, even infinitely many sets, defined in the obvious way. In particular, if I is a (possibly infinite) index set, and for each $i \in I$ we have a set A_i , the union of these sets is defined as

$$\bigcup_{i \in I} A_i = \{x \mid x \in A_i \text{ for some } i \in I\},$$

and their intersection is defined as

$$\bigcap_{i \in I} A_i = \{x \mid x \in A_i \text{ for all } i \in I\}.$$

In case we are dealing with the union or intersection of countably many sets A_i , the notation $\cup_{i=1}^{\infty} A_i$ and $\cap_{i=1}^{\infty} A_i$, respectively, is used.

Two sets are said to be disjoint if their intersection is empty. More generally, several sets are said to be **disjoint** if no two of them have a common element.

¹Some texts use the notation $A \subseteq B$ to indicate that A is a subset of B , and reserve the notation $A \subset B$ for the case where A is a **proper** subset of B , i.e., a subset of B which is not equal to B .

Disjoint sets are also said to be **mutually exclusive**. A collection of sets is said to be a **partition** of a set A if the sets in the collection are disjoint and their union is A .

2.1 The Algebra of Sets

Set operations have several properties, which are elementary consequences of the definitions. Some examples are:

$$\begin{array}{ll} A \cup B = B \cup A, & A \cap B = B \cap A, \\ A \cup (B \cup C) = (A \cup B) \cup C, & A \cap (B \cap C) = (A \cap B) \cap C, \\ A \cap (B \cup C) = (A \cap B) \cup (A \cap C), & A \cup (B \cap C) = (A \cup B) \cap (A \cup C), \\ (A^c)^c = A, & A \cap A^c = \emptyset, \\ A \cup \Omega = \Omega, & A \cap \Omega = A. \end{array}$$

Two particularly useful properties are given by **De Morgan's laws** which state that

$$\left(\bigcup_{i \in I} A_i \right)^c = \bigcap_{i \in I} A_i^c, \quad \left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} A_i^c.$$

To establish the first law, suppose that $x \in (\bigcup_{i \in I} A_i)^c$. Then, $x \notin \bigcup_{i \in I} A_i$, which implies that for every $i \in I$, we have $x \notin A_i$. Thus, x belongs to the complement of every A_i , and $x \in \bigcap_{i \in I} A_i^c$. This shows that $(\bigcup_{i \in I} A_i)^c \subset \bigcap_{i \in I} A_i^c$. The reverse inclusion is established by reversing the above argument, and the first law follows. The argument for the second law is similar.

3 NOTATION: SOME COMMON SETS

We now introduce the notation that will be used to refer to some common sets:

- (a) \mathbb{R} denotes the set of all **real numbers**;
- (b) $\overline{\mathbb{R}}$ denotes $\mathbb{R} \cup \{-\infty, \infty\}$, the set of **extended real numbers**.
- (c) \mathbb{Z} denotes the set of all **integers**;
- (d) \mathbb{N} denotes the set of **natural numbers** (the positive integers).

Also, for any $a, b \in \overline{\mathbb{R}}$, we use the following notation:

- (a) $[a, b]$ denotes the set $\{x \in \overline{\mathbb{R}} \mid a \leq x \leq b\}$;
- (b) (a, b) denotes the set $\{x \in \overline{\mathbb{R}} \mid a < x < b\}$;

- (c) $[a, b)$ denotes the set $\{x \in \overline{\mathbb{R}} \mid a \leq x < b\}$;
- (d) $(a, b]$ denotes the set $\{x \in \overline{\mathbb{R}} \mid a < x \leq b\}$.

We finally introduce some definitions related to products of sets.

- (a) The **Cartesian product** of n sets A_1, \dots, A_n , denoted by $A_1 \times A_2 \times \dots \times A_n$, or $\prod_{i=1}^n A_i$ for short, is the set of all n -tuples that can be formed by picking one element from each set, that is,

$$\prod_{i=1}^n A_i = \{(a_1, \dots, a_n) \mid a_i \in A_i, \forall i\}.$$

The set $A \times A$ is also denoted by A^2 . The notation A^n is defined similarly.

- (b) The Cartesian product $\prod_{i=1}^{\infty} A_i$ of an infinite sequence of sets A_i is defined as the set of all sequences (a_1, a_2, \dots) where $a_i \in A_i$ for each i . The simpler notation A^∞ is used if $A_i = A$ for all i .
- (c) The set of all subsets of a set A is denoted by 2^A .
- (d) Given two sets A and B , A^B stands for the set of functions from B to A .

As defined above, a sequence (a_1, a_2, \dots) of elements of a set A belongs to A^∞ . However, such a sequence can also be viewed as a function from \mathbb{N} into A , which belongs to $A^{\mathbb{N}}$. Thus, there is a one-to-one correspondence between A^∞ and $A^{\mathbb{N}}$.

In the special case where $A = \{0, 1\}$, a sequence (a_1, a_2, \dots) can be identified with a subset of \mathbb{N} , namely the set $\{n \in \mathbb{N} \mid a_n = 1\}$. We conclude that there is a one-to-one correspondence between $\{0, 1\}^\infty$, $\{0, 1\}^{\mathbb{N}}$, and $2^{\mathbb{N}}$.

4 REMARKS ON THE CARDINALITY OF SETS

For any finite set, its cardinality is intuitively defined as the number of elements in it. For infinite sets, one might be tempted to define their cardinality as infinity. However, not all infinite sets are created equal, so we need to find a way to differentiate them. We begin by defining a way to compare the cardinality of any two arbitrary sets A and B .

Definition 1.

- (a) We say that A and B have the same cardinality, denoted by $|A| = |B|$, if and only if there exists a bijective function $f : A \rightarrow B$.
- (b) We say that A has cardinality smaller than or equal to B , denoted by $|A| \leq |B|$, if and only if there exists an injective function $g : A \rightarrow B$.
- (c) We say that A has cardinality bigger than or equal to B , denoted by $|A| \geq |B|$, if and only if there exists an surjective function $g : A \rightarrow B$.

Using this definition, we define what it means for a set to be countable, or uncountable.

Definition 2.

- (a) We say that a set is countable, if its cardinality is smaller than or equal to the set of natural numbers \mathbb{N} .
- (b) We say that a set is uncountable, if its cardinality is strictly bigger than the cardinality of the natural numbers \mathbb{N} .

Since the *continuum hypothesis* (a set theoretic axiom) states that there is no set with cardinality between the cardinality of the natural numbers and the cardinality of the real numbers, uncountable sets can also be defined as the ones that have cardinality bigger than or equal to the cardinality of the real numbers.

We collect here some facts that are useful in distinguishing countable and uncountable sets.

Theorem 1.

- (a) *The union of countably many countable sets is a countable set.*
- (b) *If A is finite, of cardinality n , then 2^A has cardinality 2^n .*
- (c) *The Cartesian product of finitely many countable sets is countable.*
- (d) *The set of rational numbers is countable.*
- (e) *The set $\{0, 1\}^\infty$ is uncountable.*
- (f) *The Cartesian product of infinitely many sets (with at least two elements each) is uncountable.*

Proof.

- (a) Left as an exercise.
- (b) When choosing a subset of A , there are two choices for each element of A : whether to include it in the subset or not. Since there are n elements, with two choices for each, the total number of choices is 2^n .
- (c) Suppose that A and B are countable sets, and that $A = \{a_1, a_2, \dots\}$, $B = \{b_1, b_2, \dots\}$. We observe that

$$A \times B = \bigcup_{i=1}^{\infty} (\{a_i\} \times B).$$

For any i , there is a one-to-one correspondence between elements of B and elements of $\{a_i\} \times B$. Therefore $\{a_i\} \times B$ is countable. Using part (a) of the theorem, it follows that $A \times B$ is countable.

We continue by induction. We fix some $k \geq 2$ and use the induction hypothesis that the Cartesian product of k or fewer countable sets is countable. Suppose that the sets A_1, \dots, A_{k+1} are countable. We observe that the set $A_1 \times \dots \times A_{k+1}$ is essentially the same as the set $(A_1 \times \dots \times A_k) \times A_{k+1}$, which is a Cartesian product of two sets. The first is countable, by the induction hypothesis; the second is countable by assumption. The result follows.

- (d) Left as an exercise.
- (e) Suppose, in order to derive a contradiction, that the elements of $\{0, 1\}^\infty$ (each of which is a binary sequences) can be arranged in a sequence s_1, s_2, \dots . Consider the binary sequence s whose k th entry is chosen to be different

from the k th entry of the sequence s_k . This sequence s is certainly an element of $\{0, 1\}^\infty$, but is different from each of the sequences s_k , by construction. This means that the sequence s_1, s_2, \dots cannot exhaust all of the elements of $\{0, 1\}^\infty$ and therefore the latter set is uncountable.

- (f) Follows from (e) since 2^A has at least as many elements as $2^{\mathbb{N}}$, which can be identified with $\{0, 1\}^\infty$. \square

5 SEQUENCES AND LIMITS

Formally, a sequence of elements of a set A is a mapping $f : \mathbb{N} \rightarrow A$. Let $a_i = f(i)$. The corresponding sequence is often written as (a_1, a_2, \dots) or $\{a_k\}$ for short.

Given a sequence $\{a_k\}$ and an increasing sequence of natural numbers $\{k_i\}$, we can construct a new sequence whose i th element is a_{k_i} . This new sequence is called a **subsequence** of $\{a_k\}$. Informally, a subsequence of $\{a_k\}$ is obtained by skipping some of the elements of the original sequence.

Definition 3.

- (a) A sequence $\{x_k\}$ of real numbers (also called a “real sequence”) is said to **converge** to a real number x if for every $\epsilon > 0$ there exists some (positive integer) K such that $|x_k - x| < \epsilon$ for every $k \geq K$.
- (b) A real sequence $\{x_k\}$ is said to converge to ∞ (respectively, $-\infty$) if for every real number c there exists some K such that $x_k \geq c$ (respectively, $x_k \leq c$) for all $k \geq K$.
- (c) If a real sequence converges to some x (possibly infinite), we say that x is the **limit** of x_k ; symbolically, $\lim_{k \rightarrow \infty} x_k = x$.
- (d) A real sequence $\{x_k\}$ is called a **Cauchy sequence** if for every $\epsilon > 0$, there exists some K such that $|x_k - x_m| < \epsilon$ for all $k \geq K$ and $m \geq K$.
- (e) A real sequence $\{x_k\}$ is said to be **bounded above** (respectively, **below**) if there exists some real number c such that $x_k \leq c$ (respectively, $x_k \geq c$) for all k .
- (f) A real sequence $\{x_k\}$ is called **bounded** if the sequence $\{|x_k|\}$ is bounded above.
- (g) A real sequence is said to be **nonincreasing** (respectively, **nondecreasing**) if $x_{k+1} \leq x_k$ (respectively, $x_{k+1} \geq x_k$) for all k . A sequence that is either nonincreasing or nondecreasing is called **monotonic**.

The following result is a fundamental property of the real-number system, and is presented without proof.

Theorem 2. Every monotonic real sequence converges to an extended real number. If the sequence is also bounded, then it converges to a real number.

We continue with the definition of some key quantities associated with sets or sequences of real numbers.

Definition 4.

- (a) *The supremum (or least upper bound) of a set A of real numbers, denoted by $\sup A$, is defined as the smallest extended real number x such that $x \geq y$ for all $y \in A$.*
- (b) *The infimum (or greatest lower bound) of a set A of real numbers, denoted by $\inf A$, is defined as the largest extended real number x such that $x \leq y$ for all $y \in A$.*
- (c) *Given a sequence $\{x_k\}$ of real numbers, the supremum of the sequence, denoted by $\sup_k x_k$, is defined as $\sup\{x_k \mid k = 1, 2, \dots\}$. The infimum of a sequence is similarly defined.*
- (d) *The upper limit of a real sequence $\{x_k\}$, denoted by $\limsup_{k \rightarrow \infty} x_k$, is defined to be equal to $\lim_{m \rightarrow \infty} \sup\{x_k \mid k \geq m\}$.*
- (e) *The lower limit of a real sequence $\{x_k\}$, denoted by $\liminf_{k \rightarrow \infty} x_k$, is defined to be equal to $\lim_{m \rightarrow \infty} \inf\{x_k \mid k \geq m\}$.*

Remarks:

- (a) It turns out that the supremum and infimum of a set of real numbers is guaranteed to exist. This is a direct consequence of the way the real-number system is constructed (see, e.g., [R]). It can also be proved by building on Theorem 2.
- (b) The infimum or supremum of a set need not be an element of a set. For example, if $A = \{1/k \mid k \in \mathbb{N}\}$, then $\inf A = 0$, but $0 \notin A$.
- (c) If $\sup A$ happens to also be an element of A , then $\sup A$ is the maximum (i.e., the largest element) of A , and in that case, it is also denoted as $\max A$. Similarly, if $\inf A$ is an element of A , it is the minimum of A , and is denoted as $\min A$.
- (d) If a set or a sequence of real numbers has arbitrarily large elements (that is, no finite upper bound), then the supremum is equal to ∞ . Similarly, if it has arbitrarily small elements (that is, no finite lower bound), then the infimum is equal to $-\infty$.
- (e) A careful application of the definitions shows that $\sup \emptyset = -\infty$ and $\inf \emptyset = \infty$. On the other hand, if a set is nonempty, then $\inf A \leq \sup A$.
- (f) A sequence need not have a limit (e.g., consider the sequence $x_n = (-1)^n$). On the other hand, the upper and lower limits of a real sequence are al-

ways defined. To see this, let $y_m = \sup\{x_k \mid k \geq m\}$. The sequence $\{y_m\}$ is nonincreasing and therefore has a (possibly infinite) limit. We have $\limsup_{m \rightarrow \infty} x_k = \lim_{m \rightarrow \infty} y_m$, and the latter limit is guaranteed to exist, by Theorem 2. A similar argument applies to the lower limit.

Theorem 3. *Let $\{x_k\}$ be a real sequence.*

(a) *There holds*

$$\inf_k x_k \leq \liminf_{k \rightarrow \infty} x_k \leq \limsup_{k \rightarrow \infty} x_k \leq \sup_k x_k.$$

(b) *The sequence $\{x_k\}$ converges (to an extended real number) if and only if $\liminf_{k \rightarrow \infty} x_k = \limsup_{k \rightarrow \infty} x_k$, and in that case, both of these quantities are equal to the limit of x_k .*

The next definition refers to convergence of finite-dimensional real vectors.

Definition 5.

- (a) *A sequence $\{x_k\}$ of vectors in \mathbb{R}^n is said to converge to some $x \in \mathbb{R}^n$ if the i th coordinate of x_k converges to the i th coordinate of x , for every i . The notation $\lim_{k \rightarrow \infty} x_k = x$ is used again.*
- (b) *A sequence of vectors is called a **Cauchy sequence** (respectively, **bounded**) if each coordinate sequence is a Cauchy sequence (respectively, bounded).*
- (c) *We say that some $x \in \mathbb{R}^n$ is a **limit point** of a sequence $\{x_k\}$ in \mathbb{R}^n if there exists a subsequence of $\{x_k\}$ that converges to x .*
- (d) *Let A be a subset of \mathbb{R}^n . We say that $x \in \mathbb{R}^n$ is an **limit point** of A if there exists a sequence $\{x_k\}$, consisting of elements of A , different from x , that converges to x .*

We summarize some key facts about convergence of vector-valued sequences, see, e.g., [R].

Theorem 4.

- (a) A bounded sequence in \mathbb{R}^n has at least one limit point.
- (b) A bounded sequence in \mathbb{R}^n converges if and only if it has a unique limit point (in which case, the limit point is also the limit of the sequence).
- (c) A sequence in \mathbb{R}^n converges to an element of \mathbb{R}^n if and only if it is a Cauchy sequence.
- (d) Let $\{x_k\}$ be a real sequence. If $\limsup_{k \rightarrow \infty} x_k$ (respectively, $\liminf_{k \rightarrow \infty} x_k$) is finite, then it is the largest (respectively, smallest) limit point of the sequence $\{x_k\}$.

6 LIMITS OF SETS

Consider a sequence $\{A_n\}$ of sets. There are several ways of defining what it means for the sequence to converge to some limiting set. The definitions that will be most useful for our purposes are given below.

Definition 6.

- (a) We define $\limsup_{n \rightarrow \infty} A_n$ as the set of all elements ω that belong to infinitely many of the sets A_n . Formally,

$$\limsup_{n \rightarrow \infty} A_n = \bigcap_{k=1}^{\infty} \left(\bigcup_{n=k}^{\infty} A_n \right).$$

The notation $\{A_n\text{ i.o.}\} = \limsup_{n \rightarrow \infty} A_n$ is also used.

- (b) We define $\liminf_{n \rightarrow \infty} A_n$ as the set of all ω that belong to all but finitely many of the sets A_n . Formally,

$$\liminf_{n \rightarrow \infty} A_n = \bigcup_{k=1}^{\infty} \left(\bigcap_{n=k}^{\infty} A_n \right).$$

- (c) We say that A is the **limit** of the sequence A_n (symbolically, $A_n \rightarrow A$, or $\lim_{n \rightarrow \infty} A_n = A$) if $A = \liminf_{n \rightarrow \infty} A_n = \limsup_{n \rightarrow \infty} A_n$.

Note that a sequence of sets A_n need not have a limit, but $\limsup_{n \rightarrow \infty} A_n$ and $\liminf_{n \rightarrow \infty} A_n$ are always well defined.

In order to parse the above definitions, note that $\omega \in \cup_{n=k}^{\infty} A_n$ if and only if there exists some $n \geq k$ such that $\omega \in A_n$. We then see that ω belongs to the intersection $\cap_{k=1}^{\infty} \cup_{n=k}^{\infty} A_n$ if and only if for every k , there exists some $n \geq k$ such that $\omega \in A_n$; this is equivalent to requiring that ω belong to infinitely many of the sets A_n .

Similarly, $x \in \cup_{k=1}^{\infty} \cap_{n=k}^{\infty} A_n$ if and only if for some k , x belongs to $\cap_{n=k}^{\infty} A_n$. Equivalently, for some k , x belongs to all of the sets A_k, A_{k+1}, \dots , i.e., x belongs to all but finitely many of the sets A_n .

When, the sequence of sets $\{A_n\}$ is monotonic, the limits turn out to behave as expected.

Theorem 5.

- (a) If A_n is an increasing sequence of sets ($A_n \subset A_{n+1}$, for all n), then $\lim_{n \rightarrow \infty} A_n$ exists and is equal to $\cup_{n=1}^{\infty} A_n$.
- (b) If A_n is a decreasing sequence of sets ($A_n \supset A_{n+1}$, for all n), then $\lim_{n \rightarrow \infty} A_n$ exists and is equal to $\cap_{n=1}^{\infty} A_n$.

Reasoning about a sequence of functions is often easier than reasoning about the convergence of a sequence of sets. A link between the two notions of convergence is provided by the following.

Definition 7. The indicator function $I_A : \Omega \rightarrow \{0, 1\}$ of a set A is defined by

$$I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

We then have the following result.

Theorem 6. We have $\lim_{n \rightarrow \infty} A_n = A$ if and only $\lim_{n \rightarrow \infty} I_{A_n}(\omega) = I_A(\omega)$ for all ω .

Proof. To prove one direction of the result, we assume that $\lim_{n \rightarrow \infty} A_n = A$. Consider the two following cases:

- (i) Suppose that $\omega \in A$. Since $\liminf_{n \rightarrow \infty} A_n = A$, ω belongs to all but

finitely many of the sets A_n , which implies that $I_{A_n}(\omega) = 1$ for all but finitely many n . This establishes that $\lim_{n \rightarrow \infty} I_{A_n}(\omega) = 1 = I_A(\omega)$.

- (ii) Suppose now that $\omega \notin A$. Since $\limsup_{n \rightarrow \infty} A_n = A$, ω belongs to at most finitely many of the sets A_n , which implies that $I_{A_n}(\omega) = 0$ for all but finitely many n . This establishes that $\lim_{n \rightarrow \infty} I_{A_n}(\omega) = 0 = I_A(\omega)$, and one direction of the desired result has been proved.

To prove the reverse direction, we consider two cases.

- (i) The limit $\lim_{n \rightarrow \infty} A_n$ exists, and is a set B different than A . Then either $B \setminus A$ or $A \setminus B$ is nonempty.

Suppose $B \setminus A$ is nonempty. Let $\omega \in B \setminus A$. Then, by the part of the result that has already been established, $\lim_{n \rightarrow \infty} I_{A_n}(\omega) = I_B(\omega) = 1$. However, $\omega \notin A$ and $A = \limsup_{n \rightarrow \infty} A_n$, which means ω belongs to finitely many of the sets A_n . Thus, $\lim_{n \rightarrow \infty} I_{A_n}(\omega) = 0$, which is a contradiction.

Similarly, suppose $A \setminus B$ is nonempty. Let $\omega \in A \setminus B$. Then, by the part of the result that has already been established, $\lim_{n \rightarrow \infty} I_{A_n}(\omega) = I_B(\omega) = 0$. However, $\omega \in A$ and $A = \liminf_{n \rightarrow \infty} A_n$, which means ω belongs to all but finitely many of the sets A_n . Thus, $\lim_{n \rightarrow \infty} I_{A_n}(\omega) = 1$, which is a contradiction.

- (ii) The limit $\lim_{n \rightarrow \infty} A_n$ does not exist. In that case, we have $\liminf_{n \rightarrow \infty} A_n < \limsup_{n \rightarrow \infty} A_n$. This implies that there exists some ω that belongs to infinitely many of the sets A_n , but also does not belong to infinitely many of those sets. In that case, $I_{A_n}(\omega) = 0$, for infinitely many choices of n , and also $I_{A_n}(\omega) = 1$, for infinitely many choices of n . This implies that the sequence $I_{A_n}(\omega)$ does not converge, and therefore the condition $\lim_{n \rightarrow \infty} I_{A_n}(\omega) = I_A(\omega)$, for every ω , cannot hold.

□

7 BOREL SETS

We define the Borel σ -algebra in $I = [0, 1]$ as the σ -algebra generated by the intervals of the form $[a, b] = \{x \in I \mid a \leq x \leq b\}$, where $0 \leq a \leq b \leq 1$.

Theorem 1.

- (a) $\{x\}$ is a Borel set for any $x \in I$.
- (b) The set of rational numbers in $[0, 1]$ is a Borel set.
- (c) All sets of the form $(a, b] = \{x \in I \mid a < x \leq b\}$ or $(a, b) = \{x \in I \mid a < x < b\}$, where $0 \leq a < b \leq 1$, are Borel sets.
- (d) If S is an open set contained in I , then S is a Borel set.

Proof. (a) This is because $\{x\}$ is a set of the form $[a, b]$ with $a = b = x$.

Another interesting approach could also be used: for all $x \in [0, 1)$,

$$\{x\} = \bigcap_{n=1}^{\infty} [x, x + (1/n)].$$

For $x = 1$, we write

$$\{1\} = \bigcap_{n=1}^{\infty} [1 - (1/n), 1].$$

- (b) The set $\mathbb{Q} \cap I$ of rational numbers in $[0, 1]$ is countable, i.e., of the form $\{q_1, q_2, \dots\}$, where each q_i is a different rational number. This set can therefore be written as the union of countably many sets of the form $\{q_i\}$:

$$\mathbb{Q} \cap I = \bigcup_{i=1}^{\infty} \{q_i\}.$$

But a countable union of elements of a σ -algebra (in this instance, Borel sets) belongs to that σ -algebra.

- (c) Suppose that $0 \leq a < b \leq 1$. Let k be a positive integer such that $a + (1/k) < b - (1/k)$. Then,

$$(a, b] = \bigcup_{n=k}^{\infty} [a + (1/n), b].$$

Similarly,

$$(a, b) = \bigcup_{n=k}^{\infty} [a + (1/n), b - (1/n)].$$

Note that we can also write (a, b) using complements of Borel sets:

$$(a, b) = \left([0, a] \cup [b, 1] \right)^c.$$

From this and part (a), we get

$$(a, b] = (a, b) \cup \{b\}.$$

- (d) A subset S of I is said to be **open** if for every $x \in S$, there exists an open interval (a, b) which is contained in S and which contains x . By assumption on S , every $x \in S$ is contained in some interval (a, b) which is contained in S . Using the fact that rational numbers are dense in the reals, we can pick rational numbers q_x and r_x such that $a < q_x < x < r_x < b$. We see that any $x \in S$ is contained in one of the above constructed intervals with rational endpoints. Therefore we can write S as the following union of (possibly uncountably many) open intervals:

$$S = \bigcup_{x \in S} (q_x, r_x).$$

Since there are countably many rationals, the number of such intervals is countable. We conclude that S is a union of countably many intervals (which are Borel sets), and is therefore a Borel set.

□

References

- [R] W. Rudin, *Principles of Mathematical Analysis*, McGraw Hill, 1976.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

Complements are independent too

Problem 0.1. Let $\{A_i\}_{i \in T}$ be a (possibly infinite, possibly uncountable) set of independent events. Prove that $\{A_i^c\}_{i \in T}$ is also independent.

Recall that independence means: for every finite $I \subseteq T$,

$$\mathbb{P}\left[\bigcap_{i \in I} A_i\right] = \prod_{i \in I} \mathbb{P}[A_i]$$

This can be done in other ways, e.g. by induction (though it's a little bit of a pain) - the proof we'll use involves the Inclusion-Exclusion formula.

Proof. What we want to prove is that

$$\mathbb{P}\left[\bigcap_i A_i^c\right] = \prod_i \mathbb{P}[A_i^c]$$

given that the $\{A_i\}$ are independent. We start by rewriting

$$\begin{aligned} \prod_i \mathbb{P}[A_i^c] &= \prod_i (1 - \mathbb{P}[A_i]) \\ &= 1 - \sum_{\text{all } i} \mathbb{P}[A_i] + \sum_{\text{all } (i,j)} \mathbb{P}[A_i]\mathbb{P}[A_j] - \sum_{\text{all } (i,j,k)} \mathbb{P}[A_i]\mathbb{P}[A_j]\mathbb{P}[A_k] \dots \end{aligned}$$

(where “all (i, j, k, \dots) ” refers only to *unordered* subsets of $[n]$). By independence of $\{A_i\}$ these products are just the probabilities of intersections, so (grouping the sum terms together) the above is

$$1 - \left(\sum_{\text{all } i} \mathbb{P}[A_i] - \sum_{\text{all } (i,j)} \mathbb{P}[A_i \cap A_j] + \sum_{\text{all } (i,j,k)} \mathbb{P}[A_i \cap A_j \cap A_k] \dots \right)$$

But the thing inside the big parens is just the inclusion-exclusion formula! So we get

$$= 1 - \mathbb{P}\left[\bigcup_i A_i\right] = \mathbb{P}\left[\left(\bigcup_i A_i\right)^c\right] = \mathbb{P}\left[\bigcap_i A_i^c\right]$$

and we are done. \square

Remark: This technique can also be used to show that changing any subset of the A_i to their complements also preserves independence.

Measuring probability of converging to an average density of x heads

Problem 0.2. Consider the infinite-coin-toss model ($\Omega = \{0, 1\}^\infty$, and σ -algebra \mathcal{F} developed in Lecture 2). Fix some $x \in [0, 1]$. Is the set of all sequences whose proportion of 1's converges to x measurable in \mathcal{F} ?

Proof. First, we need to define our event. We call it

$$A_x := \left\{ \omega : \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \omega_i = x \right\}$$

To make this easier to work with, we note that $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \omega_i = x$ just means “for all $m \geq 1$, there exists some $N > 0$ (both m, N are integers) such that

$$\text{for all } n \geq N, \quad \frac{1}{n} \sum_{i=1}^n \omega_i - x \leq \frac{1}{m}$$

We use this to define a collection of sets

$$S_{m,N} := \left\{ \omega : \text{for all } n \geq N, \quad \frac{1}{n} \sum_{i=1}^n \omega_i - x \leq \frac{1}{m} \right\}$$

Replacing “there exists” and “for all” with their equivalent set operations (\cup and \cap respectively) we get

$$A_x = \bigcap_{m=1}^{\infty} \bigcup_{N=1}^{\infty} S_{m,N}$$

So if we can show that $S_{m,N} \in \mathcal{F}$ for all m, N , we are done. To do so, let’s fix m, N and define for all $k \geq 0$,

$$S_{m,N,k} := \left\{ \omega : \text{for } n \in \{N, N+1, \dots, N+k\}, \quad \frac{1}{n} \sum_{i=1}^n \omega_i - x \leq \frac{1}{m} \right\}$$

Then we note two facts:

- $S_{m,N} = \bigcap_{k=0}^{\infty} S_{m,N,k}$;
- $S_{m,N,k} \in \mathcal{F}_{N+k} \subset \mathcal{F}_0$ (the algebra from which the σ -algebra \mathcal{F} is built)

These facts together show that $S_{m,N} \in \sigma(\mathcal{F}_0) = \mathcal{F}$, and therefore $A_x \in \mathcal{F}$ as well. \square

Of monkeys and typewriters: applying Borel-Cantelli

If you've ever heard the common statement that "a monkey at a typewriter will eventually write the entire works of Shakespeare (infinitely many times, no less)", this is what it really means.

Problem 0.3. Suppose we have an infinite sequence of random coin flips - so $\Omega = \{0, 1\}^\infty$ - in which each coin flip is independent and has probability of producing 1 ("heads") with probability $p \in (0, 1)$. Let $b \in \{0, 1\}^\ell$ be any finite pattern (so ℓ is any positive integer). Prove that, almost surely, the pattern b occurs infinitely many times in the sequence.

To help prove this, we have the Borel-Cantelli lemma:

Proposition 0.1 (Borel-Cantelli (part 2)). Given a sequence A_n of events such that (i) $\sum_n \mathbb{P}[A_n] = \infty$ and (ii) the events $\{A_n\}$ are independent, and defining $A := \{A_n \text{ i.o.}\}$ (note: see lecture 3 notes for the definition of this), then $\mathbb{P}[A] = 1$.

Proof. The intuition is that we break up our outcome ω into disjoint ℓ -length blocks (running from bit $(n-1)\ell + 1$ to bit $n\ell$ so the first block goes from 1 to ℓ); letting b have j zeroes and k ones ($j + k = \ell$), and fixing a particular block $\omega_{((n-1)\ell+1):(n\ell)}$, let A_n be the event that this block is actually equal to b , i.e.

$$A_n = \{\omega : \omega_{((n-1)\ell+1):(n\ell)} = b\}$$

Then, we have

$$\mathbb{P}[A_n] = (1-p)^j p^k > 0 \text{ (because } p \neq 0, 1\text{)}$$

Therefore, $\sum_n \mathbb{P}[A_n] = \infty$; furthermore, the events $\{A_n\}$ are independent because the blocks don't overlap. So, almost surely, infinitely many of the A_n come true – and if this happens the sequence b occurs infinitely many times, as we wanted. \square

Remark: In reality, I have a hard time believing that a monkey in front of a typewriter will produce a sequence of independent letters, but for the sake of the metaphor we'll pretend that it does.

Lebesgue measure on \mathbb{R}

See lecture notes (lecture 2).

EXTRA: Pairwise independence is not independence!

Not covered in recitation, and probably most people have already seen this, but something you should definitely know:

Problem 0.4. *If a collection of events $\{A_i\}$ are pairwise independent under a probability distribution (i.e. for any $i \neq j$, $\mathbb{P}[A_i \cap A_j] = \mathbb{P}[A_i]\mathbb{P}[A_j]$) are they necessarily independent as a collection?*

No, they aren't.

Proof. We'll construct a simple counterexample in the two-fair-coins model ($\Omega = \{0, 1\}^2$, $\mathcal{F} = 2^\Omega$, \mathbb{P} uniform). Let “ \oplus ” be the XOR operation, and define:

- $A_1 := \{\omega : \omega_1 = 1\};$
- $A_2 := \{\omega : \omega_2 = 1\};$
- $A_\oplus := \{\omega_1 \oplus \omega_2 = 1\}.$

It is easy to check that each event has two elements, and so $\mathbb{P}[A_1] = \mathbb{P}[A_2] = \mathbb{P}[A_\oplus] = 1/2$; it's also easy to check that every pair of events is only satisfied by one elementary outcome (probability = 1/4), and so they are pairwise independent.

However, for them to be independent we would need $\mathbb{P}[A_1]\mathbb{P}[A_2]\mathbb{P}[A_\oplus] = \mathbb{P}[A_1 \cap A_2 \cap A_\oplus]$ as well – but the left-hand side is 1/8 whereas the right-hand side is actually 0 because no event is in all three at once. \square

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Recitation 3

Fall 2018

1 CONDITIONAL EXPECTATIONS

We have already defined the notion of a conditional PMF, $p_{X|Y}(\cdot | y)$, given the value of a random variable Y . Similarly, given an event A , we can define a conditional PMF $p_{X|A}$, by letting $p_{X|A}(x) = \mathbb{P}(X = x | A)$. In either case, the conditional PMF, as a function of x , is a bona fide PMF (a nonnegative function that sums to one). As such, it is natural to associate a (conditional) expectation to the (conditional) PMF.

Definition 1. Given an event A , such that $\mathbb{P}(A) > 0$, and a discrete random variable X , the **conditional expectation** of X given A is defined as

$$\mathbb{E}[X | A] = \sum_x x p_{X|A}(x),$$

provided that the sum is well-defined.

Note that the preceding also provides a definition for a conditional expectation of the form $\mathbb{E}[X | Y = y]$, for any y such that $p_Y(y) > 0$: just let A be the event $\{Y = y\}$, which yields

$$\mathbb{E}[X | Y = y] = \sum_x x p_{X|Y}(x | y).$$

We note that the conditional expectation is always well defined when either the random variable X is nonnegative, or when the random variable X is integrable. In particular, whenever $\mathbb{E}[|X|] < \infty$, we also have $\mathbb{E}[|X| | Y = y] < \infty$, for every y such that $p_Y(y) > 0$. To verify the latter assertion, note that for every y such that $p_Y(y) > 0$, we have

$$\sum_x |x| p_{X|Y}(x | y) = \sum_x |x| \frac{p_{X,Y}(x, y)}{p_Y(y)} \leq \frac{1}{p_Y(y)} \sum_x |x| p_X(x) = \frac{\mathbb{E}[|X|]}{p_Y(y)}.$$

The converse, however, is not true: it is possible that $\mathbb{E}[|X| | Y = y]$ is finite for every y that has positive probability, while $\mathbb{E}[|X|] = \infty$. This is left as an exercise.

The conditional expectation is essentially the same as an ordinary expectation, except that the original PMF is replaced by the conditional PMF. As such, the conditional expectation inherits all the properties of ordinary expectations (cf. Proposition 4 in the notes for Lecture 6).

1.1 The total expectation theorem

A simple calculation yields

$$\begin{aligned}\sum_y \mathbb{E}[X | Y = y] p_Y(y) &= \sum_y \sum_x x p_{X|Y}(x | y) p_Y(y) \\ &= \sum_y \sum_x x p_{X,Y}(x, y) \\ &= \mathbb{E}[X].\end{aligned}$$

Note that this calculation is rigorous if X is nonnegative or integrable.

Suppose now that $\{A_i\}$ is a countable family of disjoint events that forms a partition of the probability space Ω . Define a random variable Y by letting $Y = i$ if and only if A_i occurs. Then, $p_Y(i) = \mathbb{P}(A_i)$, and $\mathbb{E}[X | Y = i] = \mathbb{E}[X | A_i]$, which yields

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X | A_i] \mathbb{P}(A_i).$$

Example. (The mean of the geometric.) Let X be a random variable with parameter p , so that $p_X(k) = (1-p)^{k-1}p$, for $p \in \mathbb{N}$. We first observe that the geometric distribution is memoryless: for $k \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{P}(X - 1 = k | X > 1) &= \frac{\mathbb{P}(X = k + 1, X > 1)}{\mathbb{P}(X > 1)} \\ &= \frac{\mathbb{P}(X = k + 1)}{\mathbb{P}(X > 1)} \\ &= \frac{(1-p)^k p}{1-p} = (1-p)^{k-1}p \\ &= \mathbb{P}(X = k).\end{aligned}$$

In words, in a sequence of repeated i.i.d., trials, given that the first trial was a failure, the distribution of the remaining trials, $X - 1$, until the first success is the same as the unconditional distribution of the number of trials, X , until the first success. In particular, $\mathbb{E}[X - 1 | X > 1] = \mathbb{E}[X]$.

Using the total expectation theorem, we can write

$$\mathbb{E}[X] = \mathbb{E}[X | X > 1] \mathbb{P}(X > 1) + \mathbb{E}[X | X = 1] \mathbb{P}(X = 1) = (1 + \mathbb{E}[X])(1 - p) + 1 \cdot p.$$

We solve for $\mathbb{E}[X]$, and find that $\mathbb{E}[X] = 1/p$.

Similarly,

$$\mathbb{E}[X^2] = \mathbb{E}[X^2 | X > 1]\mathbb{P}(X > 1) + \mathbb{E}[X^2 | X = 1]\mathbb{P}(X = 1).$$

Note that

$$\mathbb{E}[X^2 | X > 1] = \mathbb{E}[(X-1)^2 | X > 1] + \mathbb{E}[2(X-1)+1 | X > 1] = \mathbb{E}[X^2] + (2/p) + 1.$$

Thus,

$$\mathbb{E}[X^2] = (1-p)(\mathbb{E}[X^2] + (2/p) + 1) + p,$$

which yields

$$\mathbb{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}.$$

We conclude that

$$\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

Example. Suppose we flip a biased coin N times, independently, where N is a Poisson random variable with parameter λ . The probability of heads at each flip is p . Let X be the number of heads, and let Y be the number of tails. Then,

$$\mathbb{E}[X | N = n] = \sum_{m=0}^{\infty} m\mathbb{P}(X = m | N = n) = \sum_{m=0}^n m \binom{n}{m} p^m (1-p)^{n-m}.$$

But X is just the expected number of heads in n trials, so that $\mathbb{E}[X | N = n] = np$.

Let us now calculate $\mathbb{E}[N | X = m]$. We have

$$\begin{aligned} \mathbb{E}[N | X = m] &= \sum_{n=0}^{\infty} n\mathbb{P}(N = n | X = m) \\ &= \sum_{n=m}^{\infty} n \frac{\mathbb{P}(N = n, X = m)}{\mathbb{P}(X = m)} \\ &= \sum_{n=m}^{\infty} n \frac{\mathbb{P}(X = m | N = n)\mathbb{P}(N = n)}{\mathbb{P}(X = m)} \\ &= \sum_{n=m}^{\infty} n \frac{\binom{n}{m} p^m (1-p)^{n-m} (\lambda^n / n!) e^{-\lambda}}{\mathbb{P}(X = m)}. \end{aligned}$$

Recall that $X \stackrel{d}{=} \text{Pois}(\lambda p)$, so that $\mathbb{P}(X = m) = e^{-\lambda p} (\lambda p)^m / m!$. Thus, after some

cancellations, we obtain

$$\begin{aligned}
\mathbb{E}[N \mid X = m] &= \sum_{n=m}^{\infty} n \frac{(1-p)^{n-m} \lambda^{n-m} e^{-\lambda(1-p)}}{(n-m)!} \\
&= \sum_{n=m}^{\infty} (n-m) \frac{(1-p)^{n-m} \lambda^{n-m} e^{-\lambda(1-p)}}{(n-m)!} \\
&\quad + m \sum_{n=m}^{\infty} \frac{(1-p)^{n-m} \lambda^{n-m} e^{-\lambda(1-p)}}{(n-m)!} \\
&= \lambda(1-p) + m.
\end{aligned}$$

A faster way of obtaining this result is as follows. From Theorem 3 in the notes for Lecture 6, we have that X and Y are independent, and that Y is Poisson with parameter $\lambda(1-p)$. Therefore,

$$\mathbb{E}[N \mid X = m] = \mathbb{E}[X \mid X = m] + \mathbb{E}[Y \mid X = m] = m + \mathbb{E}[Y] = m + \lambda(1-p).$$

Exercise. (Simpson's "paradox") Let S be an event and X, Y discrete random variables, all defined on a common probability space. Show that

$$\mathbb{P}[S \mid X = 0, Y = y] > \mathbb{P}[S \mid X = 1, Y = y] \quad \forall y$$

does not imply

$$\mathbb{P}[S \mid X = 0] \geq \mathbb{P}[S \mid X = 1].$$

Thus in a clinical trial comparing two treatments (indexed by X) a drug can be more successful on each group of patients (indexed by Y) yet be less successful overall.

1.2 The conditional expectation as a random variable

Let X and Y be two discrete random variables. For any fixed value of y , the expression $\mathbb{E}[X \mid Y = y]$ is a real number, which however depends on y , and can be used to define a function $\phi : \mathbb{R} \rightarrow \mathbb{R}$, by letting $\phi(y) = \mathbb{E}[X \mid Y = y]$. Consider now the random variable $\phi(Y)$; this random variable takes the value $\mathbb{E}[X \mid Y = y]$ whenever Y takes the value y , which happens with probability $\mathbb{P}(Y = y)$. This random variable will be denoted as $\mathbb{E}[X \mid Y]$. (Strictly speaking, one needs to verify that this is a measurable function, which is left as an exercise.)

Example. Let us return to the last example and find $\mathbb{E}[X \mid N]$ and $\mathbb{E}[N \mid X]$. We found that $\mathbb{E}[X \mid N = n] = np$. Thus $\mathbb{E}[X \mid N] = Np$, i.e., it is a random variable that takes the value np with probability $\mathbb{P}(N = n) = (\lambda^n / n!) e^{-\lambda}$. We found that $\mathbb{E}[N \mid X = m] = \lambda(1-p) + m$. Thus $\mathbb{E}[N \mid X] = \lambda(1-p) + X$.

Note further that

$$\mathbb{E}[\mathbb{E}[X | N]] = \mathbb{E}[Np] = \lambda p = \mathbb{E}[X],$$

and

$$\mathbb{E}[\mathbb{E}[N | X]] = \lambda(1 - p) + \mathbb{E}[X] = \lambda(1 - p) + \lambda p = \lambda = \mathbb{E}[N].$$

This is not a coincidence; the equality $\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]$ is always true, as we shall now see. In fact, this is just the total expectation theorem, written in more abstract notation.

Theorem 1. *Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a measurable function such that $Xg(Y)$ is either nonnegative or integrable. Then,*

$$\mathbb{E}[\mathbb{E}[X | Y]g(Y)] = \mathbb{E}[Xg(Y)].$$

In particular, by letting $g(y) = 1$ for all y , we obtain $\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]$.

Proof: We have

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X | Y]g(Y)] &= \sum_y \mathbb{E}[X | Y = y]g(y)p_Y(y) \\ &= \sum_y \sum_x x p_{X|Y}(x | y)g(y)p_Y(y) \\ &= \sum_{x,y} x g(y)p_{X,Y}(x, y) = \mathbb{E}[Xg(Y)].\end{aligned}$$

□

The formula in Theorem 1 can be rewritten in the form

$$\mathbb{E}[(\mathbb{E}[X | Y] - X)g(Y)] = 0. \quad (1)$$

Here is an interpretation. We can think of $\mathbb{E}[X | Y]$ as an estimate of X , on the basis of Y , and $\mathbb{E}[X | Y] - X$ as an estimation error. The above formula says that the estimation error is uncorrelated with every function of the original data.

Equation (1) can be used as the basis for an abstract definition of conditional expectations. Namely, we define the conditional expectation as a random variable of the form $\phi(Y)$, where ϕ is a measurable function, that has the property

$$\mathbb{E}[(\phi(Y) - X)g(Y)] = 0,$$

for every measurable function g . The merits of this definition is that it can be used for all kinds of random variables (discrete, continuous, mixed, etc.). However, for this definition to be sound, there are two facts that need to be verified:

- (a) Existence: It turns out that as long as X is integrable, a function ϕ with the above properties is guaranteed to exist. We already know that this is the case for discrete random variables: the conditional expectation as defined in the beginning of this section does have the desired properties. For general random variables, this is a nontrivial and deep result. It will be revisited later in this course.
- (b) Uniqueness: It turns out that there is essentially only one function ϕ with the above properties. More precisely, any two functions with the above properties are equal with probability 1.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

Practical push-forward measure

The push-forward measure is one of those things which sounds horrible and complex when done in the abstract, but is pretty simple, intuitive, and really useful when actually used on things.

Problem 0.1 (Flipping coins with uniform r.v.). Suppose you want to flip a fair coin, but you don't have a coin – you only have access to $X \sim \text{unif}[0, 1]$. How would you do it? What if you wanted the coin to not be fair but instead have probability p to land heads?

We can do this by using push-forward probability measures. The random variable we *have* operates on $([0, 1], \mathcal{B}, \lambda)$ (so $\Omega_1 = [0, 1]$); the random variable we *want* operates on $\Omega_2 = \{\text{H, T}\}$.

We now use a measurable function $f : \Omega_1 \rightarrow \Omega_2$ which you probably already figured out:

$$f(\omega) = \begin{cases} \text{H} & \text{if } \omega \leq 1/2 \\ \text{T} & \text{otherwise} \end{cases}$$

That this is measurable is obvious. We then note that $\mathbb{P}[\text{H}] = \lambda(f^{-1}(\text{H})) = 1/2$ and we're done.

To do this with a p probability of heads, just replace $1/2$ with p in the above.

Problem 0.2. What if you want $X \sim \text{unif}[0, 1]$, but you only have a fair coin? (You can flip it more than once and, as MIT students all do, you have infinite time).

Now we have the probability space $(\{0, 1\}^\infty, \mathcal{F}, \mathbb{P})$ of infinite coin-flips (we're using 0 and 1 now because it's easier to write the push-forward function) and we want to move it to $[0, 1]$. Note that an infinite bit-string looks suspiciously like a real number written in binary. Thus, the proper function f to push this forward into $[0, 1]$ is

$$f(\omega) = \sum_{i=1}^{\infty} \omega_i 2^{-i}$$

I'm not going to prove this works with all the rigorous bells and whistles, but the way to do it is to show that for any arbitrary interval $I = [a, b]$, the probability of getting an $X \in [a, b]$ is $b - a$ (which then shows that the probability matches on *every* Borel-measurable subset) – and the way to do *that* is to first consider only a, b of the form $m/2^n$ (so that we can decide the inclusion with only n bits, except for silly edge cases involving having all 0's after the n th bit) and then show that any $[a, b]$ can be approximated in this way.

Some MCT Example Problems

This is taken from last year's Week 7 recitation.

Problem 0.3. Let X_1, X_2, \dots, X, Y be random variables on the same probability space. Then, we want to know:

- Suppose $0 \geq X_n \searrow X$. Does it necessarily hold that $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$?
- Suppose $Y \leq X_n \nearrow X$, and $\mathbb{E}[|Y|] < \infty$. Does it necessarily hold that $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$?
- Suppose $0 \leq X_n \searrow X$. Does it necessarily hold that $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$?
- Suppose X_n is a continuous r.v. on \mathbb{R} for all n , and has density function f_{X_n} ; and suppose that $f_{X_n} \rightarrow f$ (pointwise). Then, is it necessarily true that:

$$1. \text{ Is } \int_{\mathbb{R}} f d\lambda \leq 1? \quad 2. \text{ Is } \int_{\mathbb{R}} f d\lambda = 1?$$

(Note that the last one is a *little* unfair because we haven't really discussed *probability density functions* yet - last year, pdf's were introduced before abstract integration - but probably many people already have some familiarity with them. If not, don't worry about that one.).

Solution: Our main tools here are the MCT and linearity of expectations.

- Yes it does. We note that $0 \leq -X_n \nearrow -X$. Then the MCT implies $\mathbb{E}[-X_n] \rightarrow \mathbb{E}[-X]$; and by linearity of expectation we can pull out the “-” and conclude $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$.
- Yes it does. We write $Z_n := X_n - Y$ and $Z = X - Y$. Then $0 \leq Z_n \nearrow Z$; furthermore, because the integral of Y is absolutely convergent, $\mathbb{E}[Z_n]$ and $\mathbb{E}[Z]$ remain well-defined (exercise for those who want: construct a counterexample when this condition is removed). Therefore, by the MCT, $\mathbb{E}[Z_n] \rightarrow \mathbb{E}[Z]$.
- No it doesn't. We can construct a counterexample: suppose we have the probability space $([0, 1], \mathcal{B}, \lambda)$ (Lebesgue) and $X_n = \frac{1}{n}(\frac{1}{\omega})$. Then note that $X_n \searrow 0$ at all ω ; but $\mathbb{E}[X_n] = \infty$ for all n (see note at bottom for why) but $\mathbb{E}[X] = 0$.
- (1) Yes. Here we use *Fatou's Lemma*. First, since f_{X_n} is a density function, $\int_{\mathbb{R}} f_{X_n} d\lambda = 1$. Second, by definition if $\lim_{n \rightarrow \infty} f_{X_n} = f$, then $\liminf_{n \rightarrow \infty} f_{X_n} = f$. But then by Fatou,

$$1 = \liminf_{n \rightarrow \infty} \int_{\mathbb{R}} f_{X_n} d\lambda \geq \int_{\mathbb{R}} (\liminf_{n \rightarrow \infty} f_{X_n}) d\lambda = \int_{\mathbb{R}} f d\lambda$$

- (2) No. Consider $X_n \sim \text{unif}[n, n+1]$ (so $f_{X_n} = \mathbf{1}_{[n, n+1]}$). Then $f_{X_n}(x) \rightarrow 0$ at all x , so f is just 0. In which case, of course, $\int_{\mathbb{R}} f d\lambda = 0 \neq 1$.

Note: Why is $\int_0^1 t^{-1} dt = \infty$? The integral of t^{-1} is $\log(t)$; and $\log(t)$ is unbounded below as $t \rightarrow 0+$ – so the definite integral ends up being $\log(1) - \log(0) = \infty$.

Fubini Failures: When does swapping order of summation change the result?

Here's an interesting thing to think about: suppose we have some ping-pong balls labeled $1, 2, \dots$, and a (really big) bucket. We make the following procedure: at each iteration n , we put in balls $10n - 9$ through $10n$ (ten balls at a time, starting with 1 through 10) and then remove ball n . Taking this to the limit - infinite iterations - how many balls are left in the bucket?

This is actually a well-known paradox. There are two perspectives:

- *Iterations view:* At each iteration, we put in 10 balls and remove 1, so we added 9 balls in total. So the number of balls grows toward ∞ and we have ∞ balls at the end.
- *Balls view:* Each ball m was put in at step $\lceil m/10 \rceil$ and then removed later at step m . So at the end, 0 balls can in the bucket.

This cute little story shows how Fubini's Theorem would break down if you failed to meet the absolute convergence condition. To make it more "mathematical", we can explicitly construct a double-sum whose value changes when the order of summation changes. The above helps to see how to do this - we'll make rows refer to iterations and columns to balls. The entry $f(m, n)$ will refer to what happened to ball n at iteration m : 1 if it was put in, -1 if it was taken out, and 0 (the vast majority of the time) when nothing happens. Formally,

$$f(m, n) = \begin{cases} 1 & \text{if } n = \lceil m/10 \rceil \\ -1 & \text{if } n = m \\ 0 & \text{otherwise} \end{cases}$$

(technically there's one exception case for ball 1 which is put in and taken out at the same iteration, so we just use $f(1, 1) = 0$ because that's really what happened).

So then, summing by rows: each row has ten 1's in it and one -1 (except the first row which just has nine 1's) and so each sums to 9; so summing over all of them gives

$$\sum_m \sum_n f(m, n) = \infty$$

But summing by columns: each column has one 1 and one -1 (except the first row which is all 0's) and so each sums to 0; so summing over all of them gives

$$\sum_n \sum_m f(m, n) = 0$$

Hence the "paradox".

Why doesn't Fubini prevent this from happening? Because f is not absolutely convergent, nor is it nonnegative – it contains infinite 1's and infinite -1 's.

Remark: A very similar example is in the lecture notes.

A weird integration counterexample

In general, Lebesgue integration is more general than Riemann integration – for example, $\int_{[0,1]} \mathbf{1}_{\mathbb{Q}}(x)dx$ is well-defined as a Lebesgue integral, but not as a Riemann integral. However, when dealing with *improper* Riemann integrals - which have an infinite domain or are unbounded around certain points - sometimes we get a function for which the Riemann integral converges but the Lebesgue integral is undefined.

Note: For this class, it is not required for you to understand this section; but it is something I thought we should mention.

Problem 0.4. Let $f(x) = \frac{\sin(x)}{x}$ (where $f(0) = 1$, which keeps it continuous). Consider:

$$\int_0^\infty f(x)dx := \lim_{t \rightarrow \infty} \int_0^t f(x)dx$$

We want to show that this is Riemann-integrable but that the integral $\int_{\mathbb{R}_+} f(x)dx$ is not Lebesgue-integrable.

We're not going to formally prove this, but we'll give a sketch. We need the following:

Fact 0.1 (Facts about the Harmonic Series).

- The Harmonic Series $h_n := \sum_{i=1}^n \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \dots$ diverges, i.e. $\lim_{n \rightarrow \infty} h_n = \infty$.
- The Alternating Harmonic Series $a_n := \sum_{i=1}^n \frac{(-1)^{n-1}}{i} = 1 - \frac{1}{2} + \frac{1}{3} - \dots$ converges to a finite result (which happens to be $\log(2)$ but honestly it doesn't matter for this).

Note the the first fact means that the Alternating Harmonic Series is *not* absolutely convergent.

Note that $f(x) = \frac{\sin(x)}{x}$ oscillates about 0, crossing at $k\pi$ for all $k > 0$. Let s_k be the area between $f(x)$ and 0 over the interval $[(k-1)\pi, k\pi]$. Note that $\sin(x)/x \leq 1$ and so $s_1 \leq \pi$.

As $k \rightarrow \infty$, these oscillations get closer and closer to being scaled-down versions of $\sin(x)$; and therefore there is a constant c such that $s_k \approx c/k$ for large values of k (there is a rigorous def'n of this but we're not going to worry about it, this is for intuition only).

Then, because the Riemann integral is a limit as $t \rightarrow \infty$, it is explicitly being summed in order: $s_1 - s_2 + s_3 - \dots$ which converges to a finite value; on the other hand, the Lebesgue integral splits off the positive from the negative parts and therefore has the order

$$(s_1 + s_3 + s_5 + \dots) - (s_2 + s_4 + s_6 + \dots) = \infty - \infty, \text{ which results in an undefined integral}$$

(note that the two sums are both more than half the Harmonic Series and therefore diverge).

EXTRA: The probabilistic method for max-cuts

The basic idea of the *probabilistic method* is as follows: Suppose we want to show that some structure exists, or that some structure of a particular size exists – say, an independent set of size k in some graph. But building the thing and showing the build works is hard. So instead we build it randomly and show that the expected size is $\geq k$; then there must be some outcome which actually achieves that size (or greater), thus showing existence.

Remark: Note that this doesn't show, at all, how to build it. However, there is a way to (sometimes) convert this kind of proof into a (deterministic) algorithm for actually constructing whatever it is. This is called the *method of conditional expectations*, and interested students are encouraged to look it up. I promise it's really cool.

Problem 0.5. Let $G = (V, E)$ be an undirected graph where $E \neq \emptyset$ (at least one edge exists). Show that V can be partitioned into $(S, V - S)$ in such a way so that strictly more than half the edges are between S and $V - S$ (as opposed to being internal in S or internal in $V - S$).

Solution: As with all probabilistic method proofs, we put vertices into S at random – specifically, $v \in S$ with probability $1/2$, independent of all other vertices. Then every edge (u, v) has a $1/2$ probability of being across S to $V - S$, because whichever set u ends up in, v has a $1/2$ chance of being in the other one. Denote the set of edges crossing $S, V - S$ to be $E(S, V - S)$, and denote S^* to be the subset maximizing $|E(S^*, V - S^*)|$.

Therefore, if $|E| = m > 0$, then $\mathbb{E}[|E(S, V - S)|] = m/2$, and so

$$E(S^*, V - S^*) \geq \mathbb{E}[|E(S, V - S)|] = m/2$$

But this isn't exactly what we wanted - we wanted a *strict* inequality. Luckily, we have the following fact (easy to verify):

Fact 0.2. If X is a random variable, then $\max_{\omega} X(\omega) = \mathbb{E}[X]$ only if $X = \mathbb{E}[X]$ almost everywhere. Furthermore, if we're in a discrete probability space (like the graph-cutting example here) and every ω has a positive probability, then

$$\max_{\omega} X(\omega) = \mathbb{E}[X] \text{ only if } X \text{ is constant.}$$

Specifically, if we find some ω for which $X(\omega) < \mathbb{E}[X]$ then we've shown that $\max_{\omega} X(\omega) > \mathbb{E}[X]$. Here we use the event that $S = \emptyset$, for which

$$|E(\emptyset, V)| = 0 < m/2 = \mathbb{E}[|E(S, V - S)|]$$

This completes the proof.

Exercise in precision: If we remove the assumption that $E \neq \emptyset$, is the theorem still true? What part of the proof breaks down?

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

1 Sum of independent random variables

Lemma 1. *If X and Y are independent random variables, then*

$$\mathbb{P}(X + Y \leq z) = \mathbb{E}[F_X(z - Y)] = \mathbb{E}[F_Y(z - X)].$$

Proof. We have

$$\begin{aligned} \mathbb{P}(X + Y \leq z) &= \mathbb{E}[1_{\{X+Y \leq z\}}] \\ &= \int_{\mathbb{R}^2} 1_{\{x+y \leq z\}} d(\mathbb{P}_X \times \mathbb{P}_Y)(x, y) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} 1_{\{x+y \leq z\}} d\mathbb{P}_X(x) \right) d\mathbb{P}_Y(y) \\ &= \int_{\mathbb{R}} F_X(z - y) d\mathbb{P}_Y(y) \\ &= \mathbb{E}[F_X(z - Y)], \end{aligned}$$

where in the third inequality we used Fubini's Theorem. \square

If X and Y are continuous, $X + Y$ is also continuous, and its density can be derived by differentiating the above expression, and using Exercise 7 of HW 5 to bring the differentiation inside the integral.

2 Gaussian, Gamma, and Exponential distributions

Theorem 1.

- (a) If $N_1 \sim N(\mu_1, \sigma_1^2)$ and $N_2 \sim N(\mu_2, \sigma_2^2)$, then $N_1 + N_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
- (b) If $G_1 \sim \text{Gamma}(k_1, \theta)$ and $G_2 \sim \text{Gamma}(k_2, \theta)$, then $G_1 + G_2 \sim \text{Gamma}(k_1 + k_2, \theta)$.
- (c) If $N \sim N(0, 1)$, then $N^2 \sim \text{Gamma}(1/2, 2)$
- (d) If $X, Y \sim N(0, 1)$, then $X^2 + Y^2 \sim \text{Exp}(2)$.
- (e) If $X, Y \sim N(0, 1)$, then $\sqrt{X^2 + Y^2}$ and $\arcsin(Y/\sqrt{X^2 + Y^2})$ are independent. Furthermore, $\arcsin(Y/\sqrt{X^2 + Y^2})$ is uniform over $(-\pi/2, \pi/2)$.

Proof. (a) It follows from applying the convolution formula for continuous random variables, and doing lots of algebra. The whole thing is even in Wikipedia:

https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables

- (b) It also follows from applying the convolution formula, and doing some algebra. For the sake of simplicity, we prove it for the case $\theta = 1$.

$$\begin{aligned}
f_{G_1+G_2}(z) &= \int_0^z f_{G_1}(x)f_{G_2}(z-x) dx \\
&= \int_0^z \frac{x^{k_1-1}e^{-x}}{\Gamma(k_1)} \frac{(z-x)^{k_2-1}e^{-(z-x)}}{\Gamma(k_2)} dx \\
&= e^{-z} \int_0^z \frac{x^{k_1-1}(z-x)^{k_2-1}}{\Gamma(k_1)\Gamma(k_2)} dx \quad \text{variable change: } x=zt \\
&= e^{-z} z^{k_1+k_2-1} \int_0^1 \frac{t^{k_1-1}(1-t)^{k_2-1}}{\Gamma(k_1)\Gamma(k_2)} dt \quad \text{almost the density of a Beta}(k_1, k_2) \text{ r.v.} \\
&= \frac{e^{-z} z^{k_1+k_2-1}}{\Gamma(k_1+k_2)}
\end{aligned}$$

(c) We have

$$\mathbb{P}(N^2 \leq z) = \mathbb{P}(|N| \leq \sqrt{z}) = 2 \int_0^{\sqrt{z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Then, differentiating with respect to z , we obtain

$$f_{N^2}(z) = \frac{z^{\frac{1}{2}-1} e^{-\frac{z}{2}}}{\sqrt{2\pi}},$$

which is the density of a $Gamma(1/2, 2)$.

- (d) From (c), we know that X^2 and Y^2 are $Gamma(1/2, 2)$. Then, applying (b) we get that $X^2 + Y^2$ is $Gamma(1, 2)$, which is the same as $Exp(2)$.
- (e) Note that $R = \sqrt{X^2 + Y^2}$ and $\Theta = \arcsin(Y/\sqrt{X^2 + Y^2})$ correspond to the radius and angle in polar coordinates. As a result, the probability of the event $\{0 \leq \Theta \leq \theta_0\} \cap \{R \leq r_0\}$ can be computed using polar coordinates as follows:

$$\begin{aligned}
\mathbb{P}\left(\{0 \leq \Theta \leq \theta_0\} \cap \{R \leq r_0\}\right) &= \int_{\{0 \leq \Theta \leq \theta_0\} \cap \{R \leq r_0\}} \frac{1}{2\pi} e^{\frac{x^2+y^2}{2}} dxdy \\
&= \int_0^{\theta_0} \int_0^{r_0} \frac{1}{2\pi} e^{\frac{r^2}{2}} r dr d\theta \\
&= \theta_0 \int_0^{r_0} \frac{1}{2\pi} e^{\frac{r^2}{2}} r dr \\
&= \mathbb{P}(0 \leq \Theta \leq \theta_0) \mathbb{P}(R \leq r_0).
\end{aligned}$$

Thus, they are independent, and Θ is uniform.

□

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

The Jacobian Formula: functions are linear if you look really close

Notational remark: The bolded variables are either matrices or vectors; I like to do that to visually remind myself what is what exactly. This will be a little confusing because usually bolded uppercase letters are matrices, lower case are vectors, but here I'm also adding *random vectors* as bolded upper-case letters. Also, $|\cdot|$, when applied to a matrix, is the *absolute value of the determinant*.

The multivariate derived-distribution problem is set up as follows: $\mathbf{X} = (X_1, \dots, X_n)$ are jointly continuous with density function $f_{\mathbf{X}}$ over \mathbb{R}^n . We also have a measurable function $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ and we define the random variable $\mathbf{Y} = g(\mathbf{X})$. Our goal is to find a good means of finding the distribution of \mathbf{Y} in terms of the distribution of \mathbf{X} .

In particular, we will make an assumption about g which is “well-behaved” in a few ways – and allows us to use the *Jacobian formula*. We will assume the following:

Assumption 0.1. Let $U \subset \mathbb{R}^n$ be an open set, and let $g : U \rightarrow \mathbb{R}^n$ be

- continuously differentiable
- an injection; and
- has non-vanishing determinant of the Jacobian, i.e. $\frac{\partial g}{\partial \mathbf{x}} \neq 0$.

We also have the following fact, which is super useful:

Fact 0.1. Define V as the image $g(U)$. Then if $g : U \rightarrow V$ satisfies the assumption: (i) V is open; (ii) $g^{-1} : V \rightarrow U$ is well-defined; (iii) g^{-1} satisfies the assumption as well.

Let us define $\mathbf{J}(\mathbf{y})$ to be the *Jacobian* (first-derivative, basically) of g^{-1} at \mathbf{y} . Basically, around any point \mathbf{y} , we consider a tiny cube A of volume δ^n and note that the probability mass inside came from the parallelepiped $B = g^{-1}(A) \approx \mathbf{J}(\mathbf{y})A$. The volume of it is then $\approx |\mathbf{J}(\mathbf{y})|\delta^n$ (linear algebra fact), and the density inside is approximately $f_{\mathbf{X}}(g^{-1}(\mathbf{y}))$. Thus, the mass ($\sim \mathbf{Y}$) inside A should be equal to the mass ($\sim \mathbf{X}$) in B , giving:

$$f_{\mathbf{y}}(\mathbf{y}) \cdot \delta^n \approx f_{\mathbf{X}}(g^{-1}(\mathbf{y})) \cdot |\mathbf{J}(\mathbf{y})|\delta^n$$

Dividing both sides by δ^n and then taking $\delta \searrow 0$ (which turns the \approx into $=$), we get the actual **Jacobian formula**:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y})) \cdot |\mathbf{J}(\mathbf{y})|$$

For convenience, we will also be using the matrix $\mathbf{M} := \frac{\partial g}{\partial \mathbf{x}}(g^{-1}(\mathbf{y}))$ (forward Jacobian of g measured at $\mathbf{x} = g^{-1}(\mathbf{y})$). We will use the fact that $|\mathbf{J}(\mathbf{y})| = |\mathbf{M}|^{-1}$.

An innocent little problem using the Jacobian formula

Problem 0.1. Let $\mathbf{X} = (X_1, X_2)$ be jointly continuous with PDF $f_{\mathbf{X}}(x_1, x_2) = \exp(-x_1 - x_2)$ for $x_1, x_2 > 0$, and let

$$\mathbf{Y} = (Y_1, Y_2) = (X_1 + X_2, X_1 X_2)$$

We want to know: (a) what is the joint PDF of \mathbf{Y} , and (b) are Y_1, Y_2 independent?

Well, to (b) we can already answer “no” because if $Y_2 \geq 100$, then Y_1 has to be bigger than 1 and that basically settles it.

(Formally, we say $\mathbb{P}[(Y_2 \geq 100) \cap (Y_1 \leq 1)] = 0 \neq \mathbb{P}[Y_2 \geq 100] \cdot \mathbb{P}[Y_1 \leq 1]$)

But let's do this in the principled way.

First, we have an issue that g is not one-to-one (note that $g(x_1, x_2) = g(x_2, x_1)$); we will solve this by means of *order statistics*. We can assume that $x_1 \neq x_2$ because $\{\mathbf{x} : x_1 = x_2\}$ has Lebesgue measure 0. Define:

$$Z_1 = \min(X_1, X_2) \text{ and } Z_2 = \max(X_1, X_2)$$

From the order-statistics problem in the homework, we know that the PDF $f_{\mathbf{Z}}$ is

$$f_{\mathbf{Z}}(z_1, z_2) = \begin{cases} 2 \exp(-z_1 - z_2) & \text{if } 0 < z_1 < z_2 \\ 0 & \text{otherwise} \end{cases}$$

Note here that our set $U \subset \mathbb{R}^2$ is now

$$U = \{\mathbf{z} : 0 < z_1 < z_2\}$$

which is indeed open, and g remains the same and is therefore still continuously differentiable. Finally, if we look at the Jacobian of g , we find that

$$\frac{\partial g}{\partial \mathbf{z}} = \begin{bmatrix} 1 & 1 \\ z_2 & z_1 \end{bmatrix} \text{ and so } \frac{\partial g}{\partial \mathbf{z}} = z_2 - z_1$$

whose determinant is not 0 since $z_2 \neq z_1$.

Ok, let's take a deep breath and remind ourselves of the Jacobian formula:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Z}}(g^{-1}(\mathbf{y})) |\mathbf{J}(\mathbf{y})|$$

(hidden is a $\mathbf{1}_V(\mathbf{y})$ term, i.e. this only works on the range of g). We'll need to find these two parts, $f_{\mathbf{Z}}(g^{-1}(\mathbf{y}))$ and $|\mathbf{J}(\mathbf{y})|$.

The Density at the Inverse: This luckily turns out to be quite easy, since by definition $z_1 + z_2 = y_1$ when $\mathbf{y} = g(\mathbf{z})$. Therefore, the density can just be computed:

$$f_{\mathbf{Z}}(g^{-1}(\mathbf{y})) = 2 \exp(-y_1)$$

The Determinant: For this, we gotta look at g^{-1} . Given \mathbf{y} , what is \mathbf{z} ? Well, solving gives

$$y_2 = z_1(y_1 - z_1) = z_2(y_1 - z_2)$$

which can be solved quadratically. z_2 is the max, so

$$z_1 = \frac{y_1 - \sqrt{y_1^2 - 4y_2}}{2} \quad \text{and} \quad z_2 = \frac{y_1 + \sqrt{y_1^2 - 4y_2}}{2}$$

As a bit of a sanity check, let's look at $y_1^2 - 4y_2$, and hope that it's positive. We know

$$y_1^2 - 4y_2 = (x_1 + x_2)^2 - 4x_1x_2 \geq 0 \text{ because it's the square of AM-GM}$$

So our receiving set V is just

$$V = \{\mathbf{y} : y_1^2 - 4y_2 \geq 0\}$$

Alright, enough putting it off: what about the Jacobian $\mathbf{J}(\mathbf{y})$ of g^{-1} ? To make things super-simple, however, note that we already have the determinant of the matrix \mathbf{M} , which is $z_1 - z_2$ (the *absolute value* of $\det(\mathbf{M})$ (at \mathbf{z}) is $z_2 - z_1$); and we know z_1 and z_2 in terms of y_1 and y_2 . Thus, we get

$$\det(\mathbf{M}) = z_1 - z_2 = \frac{y_1 - \sqrt{y_1^2 - 4y_2}}{2} - \frac{y_1 + \sqrt{y_1^2 - 4y_2}}{2} = \sqrt{y_1^2 - 4y_2}$$

and therefore

$$\det(\mathbf{J}(\mathbf{y})) = \det(\mathbf{M})^{-1} = -\frac{1}{\sqrt{y_1^2 - 4y_2}}$$

Now, we take the absolute value of this to get what we needed:

$$|\mathbf{J}(\mathbf{y})| = \frac{1}{\sqrt{y_1^2 - 4y_2}}$$

Finally, we can put everything together that we needed – not forgetting the term that we hid (indicator of V) – to get

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Z}}(g^{-1}(\mathbf{y})) |\mathbf{J}(\mathbf{y})| \mathbf{1}_V(\mathbf{y}) = \frac{2 \exp(-y_1)}{\sqrt{y_1^2 - 4y_2}} \mathbf{1}_{\{y_1^2 - 4y_2 > 0\}}$$

As an afterthought, we get part (b) – are they indepedent? – is “no” (as we already knew) because this PDF does not factor nicely into a y_1 term and a y_2 term.

Conditional probability example

Problem 0.2. Alice sends a bit to Bob; this is some $X \in \{-1, 1\}$, and the probability of $X = -1$ or 1 is $1/2$ for each. However, the communication channel is noisy - in particular, it introduces some Gaussian noise $N \sim \mathcal{N}(0, 1)$ (which is independent from the transmitted bit). Bob then receives $Y = X + N$, and wants to remove the noise and recover the original bit.

Bob finds that $Y = y$, for some $y \in \mathbb{R}$. Compute the probability $\mathbb{P}[X = 1 | Y = y]$.

This is a problem about conditioning with probability densities. Let $f_{Y|X}$ be the conditional density of Y given X , and let f_Y be the marginal density of Y . In this problem we want something of the form $\mathbb{P}[X|Y]$ but are really given things of the form $\mathbb{P}[Y|X]$ (and $\mathbb{P}[X]$) – so a natural approach is to use Bayes' formula.

Defining p_X to be the *probability mass function* of X , we get

$$\mathbb{P}[X = 1 | Y = y] = \frac{p_X(1) \cdot f_{Y|X}(y | 1)}{f_Y(y)}$$

Note that because the noise is $\mathcal{N}(0, 1)$ (and independent of X), note that $Y \sim \mathcal{N}(X, 1)$ for whatever X is. Therefore, the density

$$f_{Y|X}(y | x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}}$$

Furthermore, f_Y is built as an average of these (recalling that X can only take two values):

$$f_Y(y) = \sum_x p_X(x) \cdot f_{Y|X}(y | x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(y+1)^2}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}}}{2} = \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{(y+1)^2}{2}} + e^{-\frac{(y-1)^2}{2}}}{2}$$

because $p_X(x) = 1/2$ for $x = -1, 1$. Plugging in all of these into the formula above yields (after a bunch of cancellations with the $1/2$ and the $1/\sqrt{2\pi}$):

$$\mathbb{P}[X = 1 | Y = y] = \frac{p_X(1) \cdot f_{Y|X}(y | 1)}{f_Y(y)} = \frac{e^{-\frac{(y-1)^2}{2}}}{e^{-\frac{(y+1)^2}{2}} + e^{-\frac{(y-1)^2}{2}}} = \frac{e^y}{e^{-y} + e^y}$$

(the last step is just an algebraic simplification, cancelling out the $e^{-\frac{y^2+1}{2}}$ on the top and bottom).

Notably, this function has the following natural properties for this problem (sanity check):

- $\lim_{y \rightarrow -\infty} \mathbb{P}[X = 1 | Y = y] = 0$ and $\lim_{y \rightarrow \infty} \mathbb{P}[X = 1 | Y = y] = 1$.
- $\mathbb{P}[X = 1 | Y = y]$ is (strictly) monotonically increasing.
- $\mathbb{P}[X = 1 | Y = 0] = 1/2$.

Borel-Cantelli example

Problem 0.3. Suppose we have a sequence of nonnegative random variables X_n (not necessarily independent) such that for any constant $c > 0$, the following holds:

$$0 < \mathbb{P}[X_n > c] \leq \frac{1}{c^2}$$

We want to show the following two things:

- (a) For any constant $b > 0$, there is 0 probability that $\limsup_{n \rightarrow \infty} \frac{X_n}{n} > b$.
- (b) With probability 1, $\lim_{n \rightarrow \infty} \frac{X_n}{n} = 0$.

For part (a), this is all about getting the thing we want to prove into a format where we can hit it with the given inequality. Furthermore, recall that \limsup is basically an “infinitely often” thing, which suggests that we might want to apply *Borel-Cantelli*. This means:

$$\limsup_{n \rightarrow \infty} \frac{X_n}{n} > b \iff \left\{ \frac{X_n}{n} > b \text{ i.o.} \right\}$$

(CAUTION! Need to be careful about the inequalities - if it's \geq it becomes more complicated, see Grading Notes 1 and 3.) Furthermore, we can re-write it to make the given inequality applicable. Define:

$$A_n := \left\{ \omega : \frac{X_n(\omega)}{n} > b \right\} = \left\{ \omega : X_n(\omega) \geq bn \right\}$$

Then, applying the inequality, we get

$$\mathbb{P}[A_n] = \mathbb{P}[X_n > bn] \leq \frac{1}{b^2 n^2}$$

Therefore, summing up these probabilities gives, for any $b > 0$,

$$\sum_n \mathbb{P}[A_n] = \sum_n \frac{1}{b^2 n^2} = \frac{\pi^2}{9b^2} < \infty$$

Therefore, we can apply Borel-Cantelli to conclude that $\limsup_{n \rightarrow \infty} \frac{X_n}{n} > b$ has probability 0.

For part (b), there are two options available (both basically the same concept). First, note that because X_n is *nonnegative*, we know that $0 \leq \liminf_{n \rightarrow \infty} X_n \leq \limsup_{n \rightarrow \infty} X_n$. Therefore, if $\limsup_{n \rightarrow \infty} X_n = 0$, we know that $\limsup_{n \rightarrow \infty} X_n = 0 = \liminf_{n \rightarrow \infty} X_n$, and therefore $\lim_{n \rightarrow \infty} X_n$ exists and is 0. Thus,

$$\lim_{n \rightarrow \infty} \frac{X_n}{n} = 0 \iff \limsup_{n \rightarrow \infty} \frac{X_n}{n} = 0$$

So now we really need to write “ $\limsup_{n \rightarrow \infty} X_n = 0$ ” (as an event) in terms of events we already have - and a countable number of them too. Defining

$$C := \left\{ \omega : \limsup_{n \rightarrow \infty} \frac{X_n(\omega)}{n} = 0 \right\} \text{ and } C_k := \left\{ \omega : \limsup_{n \rightarrow \infty} \frac{X_n(\omega)}{n} \leq \frac{1}{k} \right\}$$

We then just see that (by the *union bound*, and part (a))

$$\begin{aligned} C = \bigcap_k C_k &\implies C^c = \bigcup_k C_k^c \implies \mathbb{P}[C^c] \leq \sum_k \mathbb{P}[C_k^c] \\ &= \sum_k 0 = 0 \implies \mathbb{P}[C] = 1 - \mathbb{P}[C^c] = 1 \end{aligned}$$

Alternately, it can be observed that $C_k \searrow C$, and $\mathbb{P}[C_k] = 1$ for all k ; therefore, by continuity of probability we can conclude that $\mathbb{P}[C] = \lim_{k \rightarrow \infty} \mathbb{P}[C_k] = 1$.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

Exercise 1. Let X_1 and X_2 be independent random variables, uniform over the interval $(0, 1)$. Find the PDF of $X_1 X_2$.

Solution:

1. **The Jacobian approach:** We wish to derive the PDF of $Y_1 = g(X_1, X_2) = X_1 X_2$. Thus, we define $Y_2 = X_2$ and use find the Jacobian. From the relation $x_1 = y_1/x_2$ we see that $h(y_1, y_2) = y_1/y_2$. The partial derivative $\partial h/\partial y_1$ is $1/y_2$. We obtain

$$f_{Y_1}(y_1) = \int f_X(y_1/y_2, y_2) \frac{1}{y_2} dy_2 = \int f_X(y_1/x_2, x_2) \frac{1}{x_2} dx_2.$$

Recall that $X_1, X_2 \stackrel{d}{=} U(0, 1)$, and independent. Then, their common PDF is $f_{X_i}(x_i) = 1$, for $x_i \in [0, 1]$. Note that $f_{Y_1}(y_1) = 0$ for $y \notin (0, 1)$. Furthermore, $f_{X_1}(y_1/x_2)$ is positive (and equal to 1) only in the range $x_2 \geq y_1$. Also $f_{X_2}(x_2)$ is positive, and equal to 1, iff $x_2 \in (0, 1)$. In particular,

$$f_X(y_1/x_2, x_2) = f_{X_1}(y_1/x_2)f_{X_2}(x_2) = 1, \quad \text{for } x_2 \geq y_1.$$

We then obtain

$$f_{Y_1}(y_1) = \int_{y_1}^1 \frac{1}{x_2} dx_2 = -\log y_1, \quad y_1 \in (0, 1).$$

2. **The direct approach:** The direct approach to this problem would first involve the calculation of $F_{Y_1}(y_1) = \mathbb{P}(X_1 X_2 \leq y_1)$. It is actually easier to calculate

$$\begin{aligned} 1 - F_{Y_1}(y_1) &= \mathbb{P}(X_1 X_2 \geq y_1) = \int_{y_1}^1 \int_{y_1/x_1}^1 dx_2 dx_1 \\ &= \int_{y_1}^1 \left(1 - \frac{y_1}{x_1}\right) dx_1 \\ &= (x_1 - y_1 \log x_1) \Big|_{y_1}^1 = (1 - y_1) + y_1 \log y_1. \end{aligned}$$

Thus, $F_{Y_1}(y_1) = y_1 - y_1 \log y_1$. Differentiating, we find that $f_{Y_1}(y_1) = -\log y_1$.

3. **The easiest approach:** An even easier solution for this particular problem (along the lines of the stick example in Lecture 9) is to realize that conditioned on $X_1 = x_1$, the random variable $Y_1 = X_1 X_2$ is uniform on $[0, x_1]$, and using the total probability theorem,

$$f_{Y_1}(y_1) = \int_{y_1}^1 f_{X_1}(x_1) f_{Y_1|X_1}(y_1 | x_1) dx_1 = \int_{y_1}^1 \frac{1}{x_1} dx_1 = -\log y_1.$$

Exercise 2. Let $\{X_n\}$ be a sequence of i.i.d. random variables, with $X_1 \sim \exp(\lambda)$, and let $N \sim \text{Geom}(\beta)$ be an independent geometric random variable. Show that $T = X_1 + \cdots + X_N \sim \exp(\lambda\beta)$.

Solution: It is enough to show that its mgf is

$$E[e^{sT}] = \frac{\beta\lambda}{\beta\lambda - s}$$

Taking conditional expectation, we have

$$E[e^{sT}] = E\left[e^{s \sum_{n=1}^N X_n}\right] = E\left[E\left[e^{s \sum_{n=1}^i X_n} \mid N = i\right]\right].$$

For a fixed i , we know that

$$E\left[e^{s \sum_{n=1}^i X_n} \mid N = i\right] = \left(\frac{\lambda}{\lambda - s}\right)^i$$

Combining this with what we had before, we obtain

$$E[e^{sT}] = E\left[\left(\frac{\lambda}{\lambda - s}\right)^N\right] = \sum_{n=1}^{+\infty} \left(\frac{\lambda}{\lambda - s}\right)^n \beta(1 - \beta)^{n-1} = \frac{\beta}{1 - \beta} \sum_{n=1}^{+\infty} \left[\frac{\lambda(1 - \beta)}{\lambda - s}\right]^n,$$

and thus

$$\frac{\beta}{1 - \beta} \left[\frac{1}{1 - \frac{\lambda(1 - \beta)}{\lambda - s}} - 1 \right] = \frac{\beta}{1 - \beta} \left(\frac{\lambda - s}{\lambda - s - \lambda + \lambda\beta} - 1 \right) = \frac{\beta}{1 - \beta} \left(\frac{\lambda - s + s - \lambda\beta}{\lambda\beta - s} \right) = \frac{\lambda\beta}{\lambda\beta - s}.$$

Exercise 3. (Discrete-continuous Bayes rule) As part of a clinical trial, a patient undergoes either medical treatment A or medical treatment B . The treatment is chosen randomly, and each treatment has equal probability of being chosen. After the treatment, some health index X is observed for the patient. If treatment A is selected, the PDF of X is

$$f_{X|A}(x) = \begin{cases} 1 & \text{if } 0 < x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

If treatment B is selected, the PDF of X is

$$f_{X|B}(x) = \begin{cases} 3 & \text{if } 0 < x \leq 1/3, \\ 0 & \text{otherwise.} \end{cases}$$

If we are told that the value of X was less than $1/4$, what is the conditional probability that treatment A was the one selected?

Solution: We have

$$\begin{aligned} \mathbb{P}(A | X < 1/4) &= \frac{\mathbb{P}(A)\mathbb{P}(X \leq 1/4 | A)}{\mathbb{P}(A)\mathbb{P}(X \leq 1/4 | A) + \mathbb{P}(B)\mathbb{P}(X \leq 1/4 | B)} \\ &= \frac{\mathbb{P}(A) \int_0^{1/4} f_{X|A}(x) dx}{\mathbb{P}(A) \int_0^{1/4} f_{X|A}(x) dx + \mathbb{P}(B) \int_0^{1/4} f_{X|B}(x) dx} \\ &= \frac{0.5 \int_0^{1/4} 1 dx}{0.5 \int_0^{1/4} 1 dx + 0.5 \int_0^{1/4} 3 dx} \\ &= \frac{1}{4}. \end{aligned}$$

Exercise 4. Let X_1, X_2, \dots be a sequence of i.i.d. Bernoulli random variables (coin tosses), such that $\mathbb{P}(X_1 = H) = p \in (0, 1)$. Let

$$L_n = \max\{m \geq 0 : X_n = H, X_{n+1} = H, \dots, X_{n+m-1} = H, X_{n+m} = T\}$$

be the length of the run of heads starting from the n -th coin toss. Prove that

$$\limsup_{n \rightarrow \infty} \frac{L_n}{\log(n)} = \frac{1}{\log(1/p)} \quad \text{a.s..} \quad (1)$$

Solution: First, note that L_n has the same geometric distribution for all n , i.e., we have

$$\mathbb{P}(L_n = k) = (1-p)p^k, \quad \forall k \geq 0,$$

for all n .

For any $\epsilon > 0$, we have

$$\begin{aligned} \sum_{n=1}^{\infty} \mathbb{P}\left(\frac{L_n}{\log(n)} > \frac{1+\epsilon}{\log(1/p)}\right) &\leq \sum_{n=1}^{\infty} p^{(1+\epsilon)\frac{\log(n)}{\log(1/p)}} \\ &= \sum_{n=1}^{\infty} e^{-(1+\epsilon)\log(n)} \\ &= \sum_{n=1}^{\infty} n^{-(1+\epsilon)} \\ &< \infty. \end{aligned}$$

Thus, Borel-Cantelli implies that

$$\mathbb{P}\left(\limsup_n \left\{\frac{L_n}{\log(n)} > \frac{1+\epsilon}{\log(1/p)}\right\}\right) = 0.$$

Since $L_n > (1+\epsilon)\frac{\log(n)}{\log(1/p)}$ only happens finitely many times, we also have

$$\mathbb{P}\left(\limsup_n \frac{L_n}{\log(n)} > \frac{1+\epsilon}{\log(1/p)}\right) = 0.$$

Since this holds for all $\epsilon > 0$, we must have

$$\mathbb{P}\left(\limsup_n \frac{L_n}{\log(n)} \leq \frac{1}{\log(1/p)}\right) = 1.$$

On the other hand, consider the sequence of events

$$A_n = \{X_{r_n} = H, \dots, X_{r_n+d_n-1} = H\},$$

where $r_n = n^n$ and $d_n = \lfloor \log(n)/\log(1/p) \rfloor$. We have

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} p^{d_n} = \sum_{n=1}^{\infty} e^{d_n \log(p)} \geq \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

Furthermore, note that the events A_n are independent. Thus, Borel-Cantelli implies that

$$\mathbb{P}(A_n \text{ i.o.}) = 1.$$

This means that there are runs of at least $\lfloor \log(n)/\log(1/p) \rfloor$ heads infinitely often, and thus

$$\mathbb{P}\left(\limsup_n \frac{L_n}{\log(n)} \geq \frac{1}{\log(1/p)}\right) = 1.$$

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

Taste the Rainbow?

This morning I took out a little fun-size packet of Skittles, and found to my surprise that of the 16 skittles inside not a single one was green. (Skittles come in five flavors - green, yellow, orange, red, purple - and we're going to assume that each skittle is i.i.d. assigned one of these with uniform probability. Incidentally, this story is 100% true.)

This surprised me, so I wondered – what is the probability of getting such a packet, where some flavor is missing? (I assumed that all packets have 16 skittles.) Well, for any given flavor (say, green), the probability that a skittle is not that flavor is $4/5$, and there are 16 in a packet, so

$$\mathbb{P}[\text{packet contains no green}] = (4/5)^{16}$$

But I'm not interested in just “no green” – I want to know what the probability of missing *any* flavor is. This is upper-bounded by using the Union Bound over the 5 flavors, giving

$$\mathbb{P}[\text{packet is missing a flavor}] \leq 5 \cdot (4/5)^{16}$$

This is actually a fairly close bound, because it's only due to the possibility that *two* flavors might be missing which makes it a bound and not an equality. But missing two flavors is phenomenally unlikely – and from Problem 2 on the midterm we know that

$$\mathbb{P}[\text{packet is missing a flavor}] \geq 5 \cdot (4/5)^{16} - \binom{5}{2} (3/5)^{16}$$

We can then give both upper- and lower-bounds:

$$0.14 \leq 5 \cdot (4/5)^{16} - \binom{5}{2} (3/5)^{16} \leq \mathbb{P}[\text{packet is missing a flavor}] \leq 5 \cdot (4/5)^{16} \approx 0.14$$

This is really surprising! This means that if everything is uniform and independent, roughly *one out of every seven* packs is missing a flavor. Incidentally, the probability of there being a missing-flavor packet out of *five* random packets is

$$\mathbb{P}[\text{at least one is missing a flavor}] = 1 - \mathbb{P}[\text{no packet is missing a flavor}] \geq 1 - (0.86)^5 \approx 0.53$$

This means you have a *slightly better than 1/2 chance* of getting such a pack in a group of five.

I feel like there's a fortune in bet winnings just waiting here.

Characteristic Functions

First things first – make sure you are comfortable with (a) complex numbers in general, and (b) especially with expressions of the form e^{it} , notably the Euler formula

$$e^{it} = \cos(t) + i \sin(t) \quad (\text{note that this has L2-norm of 1})$$

(and its extension $e^{it+s} = e^s(\cos(t) + i \sin(t))$).

Limitations of the MGF, and how to get around them

The MGF is a very useful tool, but it has the notable limitation of sometimes not existing. For instance, consider the *Cauchy distribution*:

Definition 0.1. The Cauchy distribution of location μ and scale γ is the continuous distribution on \mathbb{R} with PDF

$$f_X(x) = \frac{1}{\pi\gamma \left(1 + \left(\frac{x-x_0}{\gamma}\right)^2\right)}$$

This happens to have CDF of the form

$$F_X(x) = \frac{1}{\pi} \arctan\left(\frac{x-x_0}{\gamma}\right) + \frac{1}{2}$$

This is often called *pathological* because its expectation is not defined. Furthermore, the MGF is defined *nowhere* (except at $s = 0$) – we can show this by simply attempting to compute

$$M_X(s) = \mathbb{E}[e^{sX}] = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx = \int_{-\infty}^{\infty} e^{sx} \frac{1}{\pi\gamma \left(1 + \left(\frac{x-x_0}{\gamma}\right)^2\right)} dx$$

For any $s \neq 0$, we have the following for sufficiently big positive x or big negative x :

$$e^{sx} > \pi\gamma \left(1 + \left(\frac{x-x_0}{\gamma}\right)^2\right)$$

This immediately implies that the integral is infinite because it is > 1 on infinitely large measure.

So if we can't use the MGF on Cauchy, what can we do? Use e^{itX} instead of e^{sX} – the expression e^{itX} is always of L2-norm 1 because itX has no real part. We therefore define:

Definition 0.2. The *characteristic function* of a real-valued random variable X is a function $\phi_X : \mathbb{R} \rightarrow \mathbb{C}$ given by

$$\phi_X(t) := \mathbb{E}[e^{it}]$$

Because it has L2-norm of 1 everywhere, both the real and imaginary components of e^{itX} are absolutely bounded by 1 – and therefore by the Bounded Convergence Theorem, the expectation exists and is finite. Even more, we know that $\phi_X(t)$ is always within the unit circle around 0 in the complex plane.

Why is the characteristic function useful?

If you've seen *Fourier analysis*, you might recognize the characteristic function as being super similar to the Fourier transform (but without the -2π constant term in the exponent). Furthermore, we'll use without proof here the following facts (Yury will probably cover them sometime):

Proposition 0.1. X, Y have the same distribution $\iff \phi_X = \phi_Y$ everywhere.

(Note: it is possible for the characteristic functions of different random variables to agree on an interval containing 0, but somehow disagree elsewhere. However, I don't know any examples and they won't be discussed here.)

Theorem 0.1 (Levy's Continuity Theorem). If X_1, X_2, \dots and X are random variables, and $\phi_{X_n} \rightarrow \phi_X$ (pointwise) everywhere, then $X_1, X_2, \dots \rightarrow X$ in distribution.

This makes it a very powerful tool for this sort of thing.

We'll also use the following, which can be proved in the same manner as for MGFs:

Proposition 0.2. The characteristic function satisfies the following properties:

- If a, b are real numbers, $\phi_{aX+b}(t) = e^{itb}\phi_X(at)$.
- If X, Y are independent random variables, $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

Proof. For the first, we just write

$$\phi_{aX+b}(t) = \mathbb{E}[e^{it(aX+b)}] = \mathbb{E}[e^{itb} e^{it(aX)}] = e^{itb} \mathbb{E}[e^{it(aX)}] = e^{itb} \phi_X(at)$$

For the second, we use the fact that X, Y independent $\implies e^{itX}, e^{itY}$ independent. Then:

$$\phi_{X+Y}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX} e^{itY}] = \mathbb{E}[e^{itX}] \mathbb{E}[e^{itY}] = \phi_X(t)\phi_Y(t)$$

concluding the proof. □

Some quick problems using the CF

Problem 0.1. Prove that the sum of two Cauchy's is also Cauchy.

The CF of the Cauchy distribution $f_X(x) = \frac{1}{\pi\gamma(1+(\frac{x-x_0}{\gamma})^2)}$ happens to be $\phi_X(t) = e^{itx_0 - \gamma|t|}$. This is quite difficult to actually compute without complex analysis tools, but we'll use it. The rest is simple: let X, Y have parameters x_0, γ_X and y_0, γ_Y . Then

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) = e^{itx_0 - \gamma_X|t|} e^{ity_0 - \gamma_Y|t|} = e^{it(x_0+y_0) - (\gamma_X + \gamma_Y)|t|}$$

which is also the CF of a Cauchy (with parameters $x_0 + y_0$ and $\gamma_X + \gamma_Y$).

Problem 0.2. Use characteristic functions to show that average of n i.i.d. $\text{Ber}(p)$ converges to a constant (equal to the probability p) as $n \rightarrow \infty$.

We consider $X_k \sim \text{Ber}(p)$ (i.i.d.), and $S_n = \frac{1}{n} \sum_{k=1}^n X_k$. The CF of X_k is

$$\phi_{X_k}(t) = \mathbb{E}[e^{itX_k}] = (1-p) + pe^{it}$$

Furthermore, adding independent random variables multiplies CFs (same as MGFs), giving

$$\phi_{S_n}(t) = \phi_{\sum_{k=1}^n X_k}(t/n) = ((1-p) + pe^{it/n})^n = (1 + p(e^{it/n} - 1))^n$$

Note that as $n \rightarrow \infty$, we have $it/n \rightarrow 0$ – so we'll take the first-order Taylor expansion at 0:

$$e^{it/n} = 1 + it/n + O(n^{-2}) \implies (e^{it/n} - 1) = it/n + O(n^{-2})$$

(Why the first-order? Because the $O(n^{-2})$ term is too small to affect the result in the limit, even with the outer power-of- n .) This gives

$$\lim_{n \rightarrow \infty} (1 + p(e^{it/n} - 1))^n = \lim_{n \rightarrow \infty} (1 + (itp)/n)^n = e^{itp}$$

But we can easily recognize that e^{itp} is just the CF of the distribution which returns p with probability 1. Therefore, the S_n 's converge (in distribution) to that distribution.

Problem-solving about the MGF

Problem 0.3. Suppose that we know that

$$\limsup_{x \rightarrow \infty} \frac{\log(\mathbb{P}[X > x])}{x} = -t < 0$$

We want to show that the MGF $M_X(s) < \infty$ for all $s \in [0, t)$.

Note that e^{sX} is actually nonnegative. This is very useful because we can now use that nice little formula of computing the expectation of a nonnegative variable using $\mathbb{P}[X > x]$:

$$\mathbb{E}[e^{sX}] = \int_0^\infty \mathbb{P}[e^{sX} > y] dy$$

This is good, so far, but we really want $\mathbb{P}[X > x]$ – so we'll rewrite $y = e^{sx}$. Note that because e^{sx} is (strictly) monotonically increasing, $e^{sx} > e^{sx} \iff X > x$. The transformation takes y on $(0, \infty)$ to x on $(-\infty, \infty)$, and $dy = s e^{sx} dx$, giving

$$\mathbb{E}[e^{sX}] = s \int_{-\infty}^\infty e^{sx} \mathbb{P}[X > x] dx$$

Note the intuition here (**warning - not rigorous!**):

$$\begin{aligned} \frac{\log(\mathbb{P}[X > x])}{x} \leq -t &\implies \mathbb{P}[X > t] \leq e^{-tx} \\ &\implies s \int_{-\infty}^\infty e^{sx} \mathbb{P}[X > x] dx \leq s + s \int_0^\infty e^{(s-t)x} dx < \infty \end{aligned}$$

(taking advantage of the fact that for $x \leq 0$, we have $e^{sx} \mathbb{P}[X > x] \leq 1$).

How do we make this rigorous? Use an ε .

$$\limsup_{x \rightarrow \infty} \frac{\log(\mathbb{P}[X > x])}{x} = -t$$

really means that for all $\varepsilon > 0$, we have some x_ε such that

$$\frac{\log(\mathbb{P}[X > x])}{x} \leq -t + \varepsilon \quad \text{for all } x > x_\varepsilon$$

This condition is equivalent to $\mathbb{P}[X > x] \leq e^{(-t+\varepsilon)x}$ for all $x > x_\varepsilon$. Now let us fix $s \in [0, t)$ and $\varepsilon < t - s$. Now we split the integral:

$$\mathbb{E}[e^{sX}] = s \int_{-\infty}^\infty e^{sx} \mathbb{P}[X > x] dx = s \int_{-\infty}^{x_\varepsilon} e^{sx} \mathbb{P}[X > x] dx + s \int_{x_\varepsilon}^\infty e^{sx} \mathbb{P}[X > x] dx$$

The integral on the left is finite, as it decays exponentially going to $-\infty$ and is bounded above by e^{sx_ε} . The integral on the right is then upper-bounded by our result for $\mathbb{P}[X > x]$, yielding in total (for some constant C)

$$\mathbb{E}[e^{sX}] \leq C + s \int_{x_\varepsilon}^\infty e^{(s-t+\varepsilon)x} dx < \infty$$

because, of course, we chose $\varepsilon > 0$ such that $s - t + \varepsilon < 0$.

Multivariate normal - conditional expectation

Problem 0.4. Suppose that Y_1, Y_2, \dots, Y_n are i.i.d. $\sim \mathcal{N}(0, 1)$; let X_1, \dots, X_n be linear combinations of these

$$X_j = \sum_{r=1}^n C_{j,r} Y_r \text{ for some constants } C_{j,r}$$

What is the conditional expectation $\mathbb{E}[X_j | X_k]$?

Note that all the normals discussed here have expectation 0, which simplifies things. We have the formula (Theorem 1 in Lecture 14 notes)

$$\mathbb{E}[X_j | X_k] = \mu_{X_j} + V_{X_j X_k} V_{X_k X_k}^{-1} (X_k - \mu_{X_k}) = V_{X_j X_k} V_{X_k X_k}^{-1} X_k$$

where $V_{Z_1 Z_2} = \text{Cov}(Z_1, Z_2)$. The zero means also make the covariance calculations simpler:

$$V_{X_j X_k} = \mathbb{E}[X_j X_k] \quad \text{and} \quad V_{X_k X_k} = \mathbb{E}[X_k X_k]$$

Note that if we have Y_{i_1}, Y_{i_2} (for $i_1 \neq i_2$) which are therefore independent, we get

$$\mathbb{E}[Y_{i_1} Y_{i_2}] = \mathbb{E}[Y_{i_1}] \mathbb{E}[Y_{i_2}] = 0 \quad \text{and} \quad \mathbb{E}[Y_i Y_i] = \text{Var}(Y_i) = 1$$

(by definition since $Y_i \sim \mathcal{N}(0, 1)$).

Now we note the following, and use linearity of expectation:

$$\mathbb{E}[X_j X_k] = \mathbb{E}\left[\sum_{r,s} C_{j,r} C_{k,s} Y_r Y_s\right] = \sum_{r,s} C_{j,r} C_{k,s} \mathbb{E}[Y_r Y_s] = \sum_r C_{j,r} C_{k,r}$$

Note that the above holds also if $j = k$. Therefore,

$$V_{X_j X_k} = \sum_r C_{j,r} C_{k,r} \quad \text{and} \quad V_{X_k X_k} = \sum_r C_{k,r}^2$$

Plugging back in, we get

$$\mathbb{E}[X_j | X_k] = V_{X_j X_k} V_{X_k X_k}^{-1} X_k = \left(\frac{\sum_r C_{j,r} C_{k,r}}{\sum_r C_{k,r}^2}\right) X_k$$

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

Exercise 1.

- (a) If $X_1 \sim \text{Cauchy}(0, \gamma_1)$, $X_2 \sim \text{Cauchy}(0, \gamma_2)$, and they are independent, then $X_1 + X_2 \sim \text{Cauchy}(0, \gamma_1 + \gamma_2)$.
- (b) If $X \sim \text{Cauchy}(0, \gamma)$, then $\alpha X \sim \text{Cauchy}(0, \alpha\gamma)$, for all $\alpha > 0$.
- (c) Let $\{X_n\}$ be a sequence of i.i.d. random variables, with $X_1 \sim \text{Cauchy}(0, \gamma)$. Then,

$$\frac{X_1 + \cdots + X_n}{n} \sim \text{Cauchy}(0, \gamma),$$

for all n .

Solution:

- (a) We have

$$\phi_{X_1+X_2}(t) = \phi_{X_1}(t)\phi_{X_2}(t) = \exp(-\gamma_1|t|)\exp(-\gamma_2|t|) = \exp(-(\gamma_1 + \gamma_2)|t|),$$

which corresponds to a $\text{Cauchy}(0, \gamma_1 + \gamma_2)$.

- (b) We have

$$\phi_{\alpha X}(t) = \phi_X(\alpha t) = \exp(-\alpha\gamma|t|),$$

which corresponds to a $\text{Cauchy}(0, \alpha\gamma)$.

- (c) We have

$$\begin{aligned}\phi_{\frac{X_1+\cdots+X_n}{n}}(t) &= \prod_{k=1}^n \phi_{X_k} \left(\frac{t}{n} \right)^n \\ &= \prod_{k=1}^n \exp \left(-\gamma \left| \frac{t}{n} \right| \right)\end{aligned}$$

which corresponds to a $\text{Cauchy}(0, \gamma)$ for all n .

Exercise 2. Let $\{X_n\}$ be a sequence of random variables, such that $\mathbb{E}[X_n] = 0$ and $Var(X_n) \leq \sigma^2$ for all n , and such that $Cov(X_i, X_j) \rightarrow 0$ when $|i - j| \rightarrow \infty$. Then,

$$S_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{i.p.} 0.$$

Solution: For any $\epsilon > 0$, Chebyshev's inequality implies that

$$\mathbb{P}(|S_n| \geq \epsilon) \leq \frac{Var(S_n)}{\epsilon^2} = \frac{1}{n^2\epsilon^2} \sum_{i=1}^n \sum_{j=1}^n Cov(X_i, X_j).$$

Since $Cov(X_i, X_j) \rightarrow 0$ when $|i - j| \rightarrow \infty$, then for every $\delta > 0$, there exists N_δ such that $|Cov(X_i, X_j)| \leq \delta$ for all i, j such that $|i - j| > N_\delta$. Thus, we have

$$\begin{aligned} \frac{1}{n^2\epsilon^2} \sum_{i=1}^n \sum_{j=1}^n Cov(X_i, X_j) &= \frac{1}{n^2\epsilon^2} \sum_{i=1}^n \left(\sum_{j:|i-j|\leq N_\delta} Cov(X_i, X_j) + \sum_{j:|i-j|> N_\delta} Cov(X_i, X_j) \right) \\ &\leq \frac{1}{n^2\epsilon^2} \sum_{i=1}^n \left(\sum_{j:|i-j|\leq N_\delta} \sigma^2 + \sum_{j:|i-j|> N_\delta} \delta \right) \\ &\leq \frac{1}{n^2\epsilon^2} [n(2N_\delta + 1)\sigma^2 + n^2\delta] \\ &\leq \frac{2N_\delta + 1}{n\epsilon^2} + \frac{\delta}{\epsilon^2}. \end{aligned}$$

Taking limit as $n \rightarrow \infty$, we have

$$\lim_{n \rightarrow \infty} \mathbb{P}(|S_n| \geq \epsilon) \leq \frac{\delta}{\epsilon^2}.$$

Finally, since this is true for all $\delta > 0$, we get

$$\lim_{n \rightarrow \infty} \mathbb{P}(|S_n| \geq \epsilon) = 0.$$

Exercise 3. Let $\{X_n\}$ be a sequence of i.i.d. random variables such that $X_1 \sim \mathcal{N}(0, 1)$. Let us define $Y_k = X_1 + \dots + X_k$. Show that

$$\frac{Y_1 + \dots + Y_n}{n^{3/2}} \xrightarrow{d} \mathcal{N}(0, 1/3).$$

Solution: Let us define

$$S_n = \sum_{k=1}^n Y_k.$$

Note that

$$S_n = \sum_{k=1}^n (n - k + 1) X_k.$$

Then, we have

$$\begin{aligned} \phi_{S_n}(t) &= \prod_{k=1}^n \phi_{X_k}(tk) \\ &= \prod_{k=1}^n \exp\left(-\frac{(tk)^2}{2}\right) \\ &= \exp\left(-\frac{t^2}{2} \sum_{k=1}^n k^2\right) \\ &= \exp\left[-\frac{t^2}{2} \left(\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}\right)\right], \end{aligned}$$

and thus

$$\begin{aligned} \phi_{\frac{S_n}{n^{3/2}}}(t) &= \phi_{S_n}\left(\frac{t}{n^{3/2}}\right) \\ &= \exp\left[-\frac{t^2}{2} \left(\frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}\right)\right]. \end{aligned}$$

Finally, if $S = \lim_{n \rightarrow \infty} \frac{S_n}{n^{3/2}}$, we have

$$\begin{aligned} \phi_S(t) &= \lim_{n \rightarrow \infty} \phi_{\frac{S_n}{n^{3/2}}}(t) \\ &= \exp\left(-\frac{t^2}{6}\right), \end{aligned}$$

which corresponds to a $\mathcal{N}(0, 1/3)$.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

SLLN example problem

Theorem 0.1 (SLLN). Suppose X_1, X_2, \dots are i.i.d. random variables with $\mathbb{E}[|X_i|] < \infty$, and define $S_n := \frac{1}{n} \sum_{i=1}^n X_i$ for all n . Then

$$S_n \rightarrow \mathbb{E}[X_1] \text{ almost surely.}$$

Here is an example problem for using the SLLN:

Problem 0.1. Let X_1, X_2, \dots be i.i.d. nonnegative random variables with finite mean $\mathbb{E}[X_1] = \lambda$. Fixing $\varepsilon > 0$, let C_m be the event that

$$\frac{1}{t} \sum_{i=1}^t X_i - \lambda \leq \varepsilon \text{ for all } t \geq m$$

Prove that there exists some m^* for which $\mathbb{P}[C_{m^*}] > 1/2$.

By the SLLN, we know that if we define the event A as

$$A := \omega : \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^t X_i(\omega) - \lambda = 0$$

then $\mathbb{P}[A] = 1$. Let us also define the event B as

$$B := \bigcup_{m=1}^{\infty} C_m$$

Note that for any $\omega \in A$, there exists some integer $m(\omega)$ such that $C_{m(\omega)}$ happens (by definition). Therefore, for any $\omega \in A$, we know that $\omega \in B$ – and therefore $A \subseteq B$. But this implies $\mathbb{P}[\cup_m C_m] = \mathbb{P}[B] = 1$ (since $\mathbb{P}[A] = 1$).

But C_m are a nondecreasing sequence of events and therefore by continuity of probability $\lim_{m \rightarrow \infty} \mathbb{P}[C_m] = 1$. But that immediately means that for sufficiently large m^* , we have $\mathbb{P}[C_{m^*}] > 1/2$.

Convergence of empirical estimates, done three ways

Problem 0.2. Let X_1, X_2, \dots be i.i.d. $\sim \text{Bern}(p)$, and define $S_n := \frac{1}{n} \sum_{i=1}^n X_i$. We know by SLLN that $\lim_{n \rightarrow \infty} S_n \rightarrow p$ almost surely.

We now ask: given accuracy ε and confidence $1 - \delta$, how many samples of X_i do we need to estimate p to that accuracy and confidence? Formally, given $\varepsilon, \delta > 0$, we want to know the smallest n such that

$$\mathbb{P} |S_n - p| \leq \varepsilon \geq 1 - \delta$$

This can be done three ways:

- Using Hoeffding to first show that

$$\mathbb{P} |S_n - p| > \varepsilon \leq 2e^{-\frac{n\varepsilon^2}{3}}$$

- Using Chebyshev's Inequality to first show that

$$\mathbb{P} |S_n - p| > \varepsilon \leq \frac{1}{4n\varepsilon^2}$$

- Using the Central Limit Theorem, show that

$$\mathbb{P} |S_n - p| > \varepsilon = 2 - 2\Phi(2\varepsilon\sqrt{n})$$

where Φ is the CDF of $\mathcal{N}(0, 1)$ (to make this rigorous, use the Berry-Esseen Theorem).

We need the following theorems of course:

Theorem 0.2 (Hoeffding's Theorem). If X_i are i.i.d. Bernoulli random variables, and $X := \sum_{i=1}^n X_i$,

$$\mathbb{P} |X - \mathbb{E}[X]| \geq \alpha \mathbb{E}[X] \leq 2e^{-\frac{\alpha^2}{3}\mathbb{E}[X]}$$

Theorem 0.3 (Chebyshev's Inequality). If X is a r.v. with finite variance, then

$$\mathbb{P} |X - \mathbb{E}[X]| \geq \alpha \leq \frac{\text{Var}(X)}{\alpha^2}$$

Theorem 0.4 (Central Limit Theorem). Let X_1, X_2, \dots be i.i.d. with finite mean $\mathbb{E}[X_1] = \mu$ and finite variance $\text{Var}(X_1) = \sigma^2$. Then, defining $S_n := \frac{1}{n} \sum_{i=1}^n X_i$,

$$\frac{nS_n - n\mu}{\sqrt{n}\sigma} \rightarrow \mathcal{N}(0, 1) \text{ (in distribution)}$$

By Hoeffding

We multiply everything by n since

$$\mathbb{P} |S_n - p| > \varepsilon = \mathbb{P} |nS_n - np| > n\varepsilon$$

nS_n is just the sum of the X_i 's, and np their expected value. So we just plug into the Hoeffding theorem (with $\alpha = \varepsilon$) to get the result given. Now we need n large enough that

$$2e^{-\frac{n\varepsilon^2}{3}} \leq \delta$$

A little algebra reveals that this is equivalent to $n \geq \frac{3\log(2/\delta)}{\varepsilon^2}$

By Chebyshev

Again we multiply everything by n . Because the X_i are independent,

$$\text{Var}(nS_n) = n\text{Var}(X_i) = np(1-p) \leq \frac{n}{4}$$

Plugging in $\alpha = n\varepsilon$ and the variance above, we get

$$\mathbb{P} |S_n - p| > \varepsilon \leq \frac{1}{4n\varepsilon^2}$$

exactly as we wanted it. Again, we want n large enough that this is $\leq \delta$. Algebra gives: $n \geq \frac{1}{4\delta\varepsilon^2}$

By Central Limit Theorem

Once again, multiply everything by n (and move the denominator over) using $\mu = p$, and define $Y_n := \frac{nS_n - n\mu}{\sqrt{n}\sigma}$. Then we get

$$|Y_n| > \frac{\varepsilon\sqrt{n}}{\sigma} \iff \frac{nS_n - n\mu}{\sqrt{n}\sigma} > \frac{\varepsilon\sqrt{n}}{\sigma} \iff S_n - p > \varepsilon$$

But by the CLT, Y_n is approximately distributed according to $\mathcal{N}(0, 1)$. Therefore, by symmetry,

$$\mathbb{P} |S_n - p| > \varepsilon = \mathbb{P} |Y_n| > \frac{\varepsilon\sqrt{n}}{\sigma} \approx 2(1 - \Phi(\frac{\varepsilon\sqrt{n}}{\sigma})) = 2 - 2\Phi(\frac{\varepsilon\sqrt{n}}{\sigma}) \leq 2 - 2\Phi(2\varepsilon\sqrt{n})$$

The last step happens because $\sigma^2 = p(1-p) \leq 1/4$, which implies $\sigma \leq 1/2$. Note that this is highly non-rigorous, so when in doubt double-check with Berry-Esseen.

Finally, let's see what n is required for a given ϵ, δ . We really want to upper-bound the probability of failure by δ , meaning:

$$2 - 2\Phi(2\varepsilon\sqrt{n}) = \delta \iff \Phi(2\varepsilon\sqrt{n}) = \frac{2-\delta}{2} \iff n = \frac{\Phi^{-1}(\frac{2-\delta}{2})^2}{4\varepsilon^2}$$

Conclusion

Note that the three results have equivalent dependencies on $\frac{1}{\varepsilon}$ (i.e. $\frac{1}{\varepsilon^2}$), but Hoeffding has much better dependence on $\frac{1}{\delta}$ than Chebyshev does; however, Chebyshev has a better constant and can therefore be better when confidence doesn't need to be super large. Meanwhile, the CLT-derived bound is by far the strongest in general but you have to be careful about when you can use it (again, for full rigor, apply Berry-Esseen). For instance, if $\delta = 0.05$ (95% confidence wanted) and $\varepsilon = 0.01$, then:

- Chebyshev proves that 50,000 trials are sufficient.
- Hoeffding proves that 110,667 trials are sufficient.
- CLT proves that (approximately) 9,604 trials are sufficient.

However, if we require much higher confidence, i.e. $\delta = 0.001$, then

- Chebyshev proves that 2,500,000 trials are sufficient.
- Hoeffding proves that 228,028 trials are sufficient.
- CLT proves that (approximately) 27,069 trials are sufficient.

The Chernoff-Union One-Two Punch Combo

Problem 0.3. Let Z_1, \dots, Z_n be uniformly distributed (i.i.d.) in $[0, 1]^2$, and let $L(Z_1, \dots, Z_n)$ be the length of the shortest continuous path which visits all n points. Prove that with high probability, $L_n \propto n^{1/2}$, i.e. that there are constants $0 < b < B$ and a polynomial (or faster-growing function) $q(n)$ such that

$$b n^{1/2} \leq L(Z_1, \dots, Z_n) \leq B n^{1/2}$$

with probability at least $1 - 1/q(n)$ (for all sufficiently large n).

Fun fact: The proof we'll demonstrate also suffices to show that if we are in d -dimensional space, $L \propto n^{1-1/d}$; in fact, it can be generalized to show strong lower bounds for L when the path must satisfy differential constraints as well! (This was part of my master's thesis.)

The upper-bound, in about two seconds

The upper bound is easy and not really the focus here. We can just split up the square into n cells of size $n^{-1/2} \times n^{-1/2}$ and then travel from cell to cell getting all the points before moving on. This takes time proportional to $n^{1/2}$ because each arc has length at most $\sqrt{2}n^{-1/2}$ and there are at most $2n$ of them (point to point and cell to cell). Note that this *always* works.

The lower-bound game plan

For the lower bound, we ask a related question: what is the maximum number of points we can collect with a path of length 1? Clearly if the answer is at most proportional to $n^{1/2}$ we are done because it would then take at least (proportional to) $n^{1/2}$ such paths to get all n points.

We'll begin with the following facts. The proof would work with any constant, but 7 is the smallest constant which works and this makes the constant bounds tighter. The second bound is certainly not the tightest possible, but it doesn't matter (it doesn't even really make the constant bounds worse)!

Fact 0.1. For any ε , a radius- 2ε circle can be covered by 7 radius- ε circles.

Fact 0.2. The unit square can be covered by ε^{-2} circles of radius ε .

Now we are going to follow this game plan:

- Discretize the problem by representing paths by sequences of (appropriately sized) circles.
- Compute the number of such sequences of circles (horrifyingly large!)
- Compute the probability that an arbitrary fixed sequence contains “too many” points (very small thanks to Chernoff!)
- Apply the Union Bound to get the result.

Discretization

Let $B_\ell(z)$ denote the ball of radius ℓ centered at z . We also fix a ‘canonical’ configuration of seven radius- ℓ circles for any radius- 2ℓ circle.

Suppose we let $\epsilon = n^{-1/2}$, and cover the space with $\epsilon^{-2} = n$ circles of radius ϵ . Now we have a path of length ≤ 1 - call it a function $\phi : [0, 1] \rightarrow [0, 1]^2$ which satisfies the *Lipschitz condition* (i.e. $\|\phi(x_1) - \phi(x_2)\|_2 \leq |x_1 - x_2|$) – this is effectively a “speed limit” on the function.

Let us check in on ϕ every ϵ – i.e. we look at $\phi(t\epsilon)$ for all $t = 1, 2, \dots, n^{1/2}$. Specifically, we do the following:

1. Let ψ_0 be the center of the circle $\phi(0)$ falls into (if there is more than one, pick arbitrarily).
2. We build a sequence of points $\psi := \psi_0, \psi_1, \dots, \psi_{n^{1/2}-1}$ as follows:
 - (a) For every $t = 0, 1, \dots, n^{1/2} - 1$, if $\phi(t\epsilon) \in B_\epsilon(\psi_t)$, then $\phi((t+1)\epsilon) \in B_{2\epsilon}(\psi_t)$ (can’t make it out in time because of the Lipschitz condition).
 - (b) Therefore, by Fact 1 from the previous page, it must be in one (at least - if more than one, pick arbitrarily) of the 7 radius- ϵ circles which cover $B_{2\epsilon}(\psi_t)$ – so let ψ_{t+1} be the center of that circle.

Note that this always preserves the condition that $\phi(t\epsilon) \in B_\epsilon(\psi_t)$.

Note also that between $\phi(t\epsilon)$ and $\phi((t+1)\epsilon)$, the path can never leave $B_{2\epsilon}(\psi_t)$ (by the Lipschitz condition) – and therefore the path is entirely contained in

$$S_\psi := \bigcup_{t=0}^{n^{1/2}-2} B_{2\epsilon}(\psi_t)$$

So how many different sequences of this type are possible? Well, denote the set of all such sequences of Ψ :

- There are n choices for the ψ_0 (just the n circles covering $[0, 1]^2$).
- For each of $n^{1/2}$ steps, there are (at most since we might be on the boundary) 7 choices for the next circle.

Therefore $|\Psi| = n \cdot 7^{n^{1/2}}$, which we’ll just round up to $e^{2n^{1/2}}$ because frankly we’re not trying to make this super efficient anyway and $e^2 > 7$ is convenient.

Now what we’ll do is see the maximum number a *sequence of circles* can cover; since every length-1 path is contained in such a sequence this upper-bounds the number of points any length-1 path can collect.

A fixed circle-sequence

Now let's fix an arbitrary sequence $\psi = \psi_0, \dots, \psi_{n^{1/2}-1}$ before the random targets are assigned and ask: how many random targets does this sequence cover (with circles of radius 2ε)?

Well, the sequence consists of $n^{1/2}$ circles of area $\pi(2\varepsilon)^2 = 4\pi n^{-1}$. Thus, the area is at most $4\pi n^{-1/2}$ (again, we can make this more efficient by noting that the circles must overlap and yada yada, but that's a lot of work and doesn't really change the point). Let's define the random variables

$$X_i := \begin{cases} 1 & \text{if } Z_i \in S_\psi \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad X_\psi := \sum_{i=1}^n X_i$$

Since the points are dropped in uniformly, we get

$$\mathbb{E}[X_\psi] := \mathbb{E}[\#\{i : Z_i \text{ in } S_\psi\}] = \sum_{i=1}^n \mathbb{E}[X_i] \leq 4\pi n^{-1/2} \cdot n = 4\pi n^{1/2}$$

Note that equality is “worst case” here, so if we treat the area as equal to this, we'll get an upper-bound, as we want. Since the points are independent, we get to apply the *Chernoff Bound*. Specifically, we use this version:

Theorem 0.5 (Hoeffding Upper Bound). If X_i are i.i.d. Bernoulli random variables, and $X := \sum_{i=1}^n X_i$,

$$\mathbb{P}[X \geq (1 + \delta)\mathbb{E}[X]] \leq e^{-\frac{\delta^2}{3}\mathbb{E}[X]}$$

Plugging in our $\mathbb{E}[X]$, we get that

$$\mathbb{P}[X_\psi \geq (1 + \delta)4\pi n^{1/2}] \leq e^{-\frac{\delta^2}{3}4\pi n^{1/2}} = e^{-\frac{4\pi}{3}\delta^2 n^{1/2}}$$

We will determine what δ we should use later - for now, let's just keep it as a variable.

Union-bounding, and choosing δ

Okay, now we simply combine the two results from above with the Union Bound, define an appropriate δ , and finish. We are interested in bounding

$$\mathbb{P}[\max_{\psi \in \Psi} X_\psi \geq (1 + \delta)4\pi n^{1/2}]$$

We have $e^{2n^{1/2}}$ sequences in Ψ ; each has (at most) a probability of $e^{-\frac{4\pi}{3}\delta^2 n^{1/2}}$ to break the bound. Therefore, the above is just

$$\mathbb{P}[\max_{\psi \in \Psi} X_\psi \geq (1 + \delta)4\pi n^{1/2}] \leq e^{2n^{1/2}} e^{-\frac{4\pi}{3}\delta^2 n^{1/2}}$$

Let's pick $\delta = 0.7$, giving us

$$\mathbb{P}[\max_{\psi \in \Psi} X_\psi \geq 1.7 \cdot 4\pi \cdot n^{1/2}] \leq e^{2n^{1/2}} e^{-\frac{4\pi}{3}(0.7)^2 n^{1/2}} < 2e^{-0.052 \cdot n^{1/2}}$$

Finishing the argument

Since $\max_{\psi \in \Psi} X_\psi$ is an upper bound for how many points a length-1 path can collect, we get the result that

Proposition 0.1. With probability at least $1 - e^{-0.052 \cdot n^{1/2}}$, there is no length-1 path which collects more than $1.7 \cdot 4\pi \cdot n^{1/2} < 22n^{1/2}$ of the n random points.

This now finishes what we wanted to show about the TSP, because if it takes length 1 to get $22n^{1/2}$ points, it will take at least a path of length $\frac{1}{22}n^{1/2}$ to get all n of them giving our final theorem (combined with the upper bound at the top):

Theorem 0.6. Let Z_1, \dots, Z_n be uniformly distributed (i.i.d.) in $[0, 1]^2$, and let $L(Z_1, \dots, Z_n)$ be the length of the shortest continuous path which visits all n points. Then

$$\frac{1}{22} n^{1/2} \leq L(Z_1, \dots, Z_n) \leq 2\sqrt{2} n^{1/2}$$

with probability at least $1 - e^{-0.052 \cdot n^{1/2}}$ (for all sufficiently large n).

Note that the constant factors here are rather horrendous. They can be tightened quite a bit on both ends; but fundamentally this technique will suffer from this problem because of how generous we're prepared to be to allow the paths to get *all* the points in their associated sets. But it does show rate-of-growth quite nicely and can be generalized to a much wider class of problems in which the paths have to satisfy differential constraints.

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J

Recitation 11

Fall 2018

Exercise 1. Consider a discrete-time, finite-state Markov chain $\{X_t\}$, with states $\{1, \dots, n\}$, and transition probabilities p_{ij} . States 1 and n are absorbing, that is, $p_{11} = 1$ and $p_{nn} = 1$. All other states are transient. Let A_1 be the event that the state eventually becomes 1. For any possible starting state i , let $a_i = \mathbf{P}(A_1 \mid X_0 = i)$ and assume that $a_i > 0$ for every $i \neq n$. Conditional on the information that event A_1 occurs, is the process X_n necessarily Markov? If yes, provide a proof, together with a formula for its transition probabilities. If not, provide a counterexample.

Solution: The answer is yes. Let B be an event of the form

$$B = \{X_0 = i_0, X_1 = i_1, \dots, X_{t-1} = i_{t-1}\}.$$

It suffices to show that the transition probability $\mathbb{P}(X_{t+1} = j \mid X_t = i, A_1, B)$ is unaffected by the past history (the event B). We have

$$\mathbb{P}(X_{t+1} = j \mid X_t = i, A_1, B) = \frac{\mathbb{P}(X_{t+1} = j, A_1 \mid X_t = i, B)}{\mathbb{P}(A_1 \mid X_t = i, B)}.$$

By the Markov property of the process $\{X_t\}$ (the future is independent of the past, given the present), we have

$$\mathbb{P}(X_{t+1} = j, A_1 \mid X_t = i, B) = \mathbb{P}(X_{t+1} = j, A_1 \mid X_t = i),$$

and

$$\mathbb{P}(A_1 \mid X_t = i, B) = \mathbb{P}(A_1 \mid X_t = i),$$

from which the desired result follows.

Furthermore,

$$\begin{aligned} \mathbb{P}(X_{t+1} = j \mid X_t = i, A_1) &= \frac{\mathbb{P}(X_{t+1} = j, A_1 \mid X_t = i)}{\mathbb{P}(A_1 \mid X_t = i)} \\ &= \frac{\mathbb{P}(A_1 \mid X_t = i, X_{t+1} = j)\mathbb{P}(X_{t+1} = j \mid X_t = i)}{\mathbb{P}(A_1 \mid X_t = i)} \\ &= \frac{p_{ij}a_j}{a_i}. \end{aligned}$$

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>

A Coupling Example

Recall the Ehrenfest Chain $\{X_t\}$ – we have n particles, each in one of two boxes (left or right). The *state* of the chain is simply the number of particles currently in the left box (and therefore ranges from 0 to n). At every step, a particle is chosen uniformly at random (independent of the past) and then moved to the other box. Therefore, the transition kernel is

$$P(i, j) := \begin{cases} \frac{i}{n} & \text{if } j = i - 1 \\ \frac{n-i}{n} & \text{if } j = i + 1 \\ 0 & \text{otherwise} \end{cases}$$

We define $\mathbb{E}^i[\tau_j]$ to be the expected number of steps to reach state j starting at state i .

Problem 0.1. Use *coupling* to show that $\mathbb{E}^0[\tau_{n/3}]$ is linear in n (so $X_0 = 0$).

Remember that *coupling* is running two stochastic processes off the same randomness in some way, so as to get the desired result. Note that so long as $t \leq \tau_{n/3}$ (where $X_0 = 0$) we have $X_t \leq n/3$ and therefore

$$\mathbb{P}[X_{t+1} = X_t + 1] \geq 2/3 \text{ and } \mathbb{P}[X_{t+1} = X_t - 1] \leq 1/3$$

Therefore, we'll couple this to a weighted random walk $\{Y_t\}$ (with $Y_0 = 0$) with probability $2/3$ to increment every step, and $1/3$ to decrement. Formally, we do this coupling by defining $U_t \stackrel{iid}{\sim} \text{unif}[0, 1]$ and then defining for $t \geq 1$:

$$X_t = X_{t-1} + \mathbf{1}_{U_t \geq 1 - \frac{X_t}{n}} - \mathbf{1}_{U_t < \frac{X_t}{n}} \text{ and } Y_t = Y_{t-1} + \mathbf{1}_{U_t \geq 2/3} - \mathbf{1}_{U_t < 1/3}$$

It's easy to confirm that $\{X_t\}$ and $\{Y_t\}$ behave (marginally) as required; it's also easy to confirm (by induction) that for all $t \leq \tau_{n/3}$ we have $Y_t \leq X_t$, because $X_0 = Y_0 = 0$, and $X_t - X_{t-1} \geq Y_t - Y_{t-1}$ for such t . Therefore, defining $\tau_{n/3}^* := \min_t \{Y_t \geq n/3\}$, we have $\tau_{n/3}^* \geq \tau_{n/3}$; furthermore, $\{Y_t\}$ is much easier to analyze than $\{X_t\}$. Therefore $\mathbb{E}^0[\tau_{n/3}^*] \geq \mathbb{E}^0[\tau_{n/3}]$.

(To intuitively see what $\tau_{n/3}^* \geq \tau_{n/3}$, imagine two runners running a race – if we know that runner 1 is always faster than runner 2, then runner 1 will finish faster than runner 2; and no information about exactly how fast they're running at any particular time is necessary to conclude this.)

Now we can analyze $\mathbb{E}^0[\tau_{n/3}^*]$. Let us define $a_k := \mathbb{E}^{n/3-k}[\tau_{n/3}^*]$. Obviously, $a_0 = 0$; and by the recursion rules we learned give that for $k \geq 1$,

$$a_k = \frac{1}{3}a_{k+1} + \frac{2}{3}a_{k-1} + 1$$

This is solved by $a_k = 3k$. Therefore, $\mathbb{E}^0[\tau_{n/3}^*] = a_{n/3} = n$. Therefore, $\mathbb{E}^0[\tau_{n/3}] \leq n$, so we have a linear upper bound. But also we have a trivial linear lower bound because at least $n/3$ steps are necessary for the Ehrenfest chain to make it from 0 to $n/3$. Therefore

$$n/3 \leq \mathbb{E}^0[\tau_{n/3}] \leq n \implies \mathbb{E}^0[\tau_{n/3}] = \Theta(n)$$

MIT OpenCourseWare

<https://ocw.mit.edu>

6.436J / 15.085J Fundamentals of Probability

Fall 2018

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>