Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 1: Basic Set Theory

Lecturer: Krishna Jagannathan *Scribe: Arjun Bhagoji*

We will begin with an informal and intuitive approach to set theory known as "Naive Set Theory".

1.1 What is a set?

A set can be thought of as a collection of *well-defined* objects. By well-defined, we mean that an object either belongs to a set or it does not. Objects belonging to a set are known as *elements* of the set. Sets can be specified in 2 ways:

1. *Extensional definition*-All the elements of the set are listed out explicitly and enclosed within curly brackets. E.g., the set of all natural numbers from 1 to 5 may be specified as $A = \{1, 2, 3, 4, 5\}$.
2. *Intensional definition*-Here, a set is defined in terms of the property which is satisfied by all its members. This is also known as the *set builder notation*. E.g., the set A above may also be defined as $A = \{x | x : x \leq 5, x \in \mathbb{N}\}$. In general, some set C may defined as $C = \{x | P(x)\}$, where $P(x)$ is some property.

We now define the notion of a subset and use the idea of subsets to define when two sets are identical.

Definition 1.1 *(i) A set A is said to be a subset of (or contained in) another set B if every element of A is also an element of B. This is denoted as $A \subseteq B$. Here, B is said to be a superset of A.*

(ii) A is a proper *subset of B (denoted $A \subset B$) if A is a subset of B and there is at least one element in B which does not belong to A.*

(iii) Two sets A and B are said to be identical (or equal) if $A \subseteq B$ and $B \subseteq A$. In other words, every element of A is an element of B, and vice versa.

Two special sets of interest are:

1. The *universal set* U, a set which contains all elements[1]
2. The *empty set* $\emptyset$, which has, as its name indicates, no elements. It is a subset of every set including itself and a proper subset of every set excluding itself.

1.2 Operations on sets

1.2.1 Complement

Taking the complement of a set is a unary operation (i.e., only one set is operated upon) defined as

[1]However, in the usual formulations of set theory, the concept of a universal set leads to a paradox known as Russell's paradox. The interested student may look up this famous paradox,'*en.wikipedia.org/wiki/Russell's_paradox*'.

Definition 1.2 *For a set A, its complement is defined as $A^c \triangleq \{x | x \notin A, x \in U\}$.*

The context for the complement of a set is provided by the universal set U. The Venn diagram representation of a set's complement is

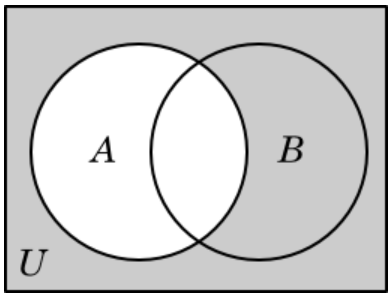

Figure 1.1: Complement (gray area) of a set A

1.2.2 Union and Intersection

Let $\mathcal{I}$ be an abstract index set. Consider a family of sets $\{A_i,\ i \in \mathcal{I}\}$ indexed by $\mathcal{I}$.

Definition 1.3 Union*: The union of $\{A_i,\ i \in \mathcal{I}\}$ is defined as*

$$\bigcup_{i \in \mathcal{I}} A_i = \{x | x \in A_j \text{ for some } j \in \mathcal{I}\}.$$

In words, the union $\bigcup_{i \in \mathcal{I}} A_i$ is a set consisting of those elements which are elements of *at least one* of the A_i's.

Definition 1.4 Intersection*: The intersection of $\{A_i,\ i \in \mathcal{I}\}$ is defined as*

$$\bigcap_{i \in \mathcal{I}} A_i = \{x | x \in A_j \text{ for every } j \in \mathcal{I}\}.$$

In words, the intersection $\bigcap_{i \in \mathcal{I}} A_i$ is a set consisting of those elements which are elements of *all* the A_i's.

Remark: 1.5 *When the index set $\mathcal{I}$ is a finite set, say $\mathcal{I} = \{1, 2, 3\}$ the definition of union given above coincides with the "middle-school" understanding of unions, i.e., taking the union of sets one-at-a-time. For example, $\bigcup_{i=1}^{3} A_i = A_1 \bigcup A_2 \bigcup A_3$. However, this "one-by-one" interpretation completely breaks down when the index set $\mathcal{I}$ is infinite. For example when $\mathcal{I} = \mathbb{N}$, the union $\bigcup_{i=1}^{\infty} A_i$ does not have any interpretation in terms of taking unions one by one, till infinity. After all, there is no A_∞ in the family $\{A_i, i \in \mathbb{N}\}$, and there is no notion of "limiting unions". Thus, $\bigcup_{i=1}^{\infty} A_i$ should be interpreted just as Definition 1.3 says: it is the set of all elements contained in at least one of the $A_i,\ i \in \mathbb{N}$.*

In order to avoid the (dangerous) temptation to interpret $\bigcup_{i=1}^{\infty} A_i$ as some sort of a limit of finite, "one-by-one" unions, a better notation would be to use $\bigcup_{i \in \mathbb{N}} A_i$, instead of the potentially misleading but more commonly used notation $\bigcup_{i=1}^{\infty} A_i$.

The following useful identities related to unions and intersections can be proven easily (*do it!*) from the definitions.

$$\left(\bigcap_{i\in\mathcal{I}} A_i\right)\bigcup B = \bigcap_{i\in\mathcal{I}}(A_i\bigcup B), \tag{1.1}$$

and

$$\left(\bigcup_{i\in\mathcal{I}} A_i\right)\bigcap B = \bigcup_{i\in\mathcal{I}}(A_i\bigcap B), \tag{1.2}$$

An especially important set of laws regarding the interchangeability of unions and intersections under the complement operation are *De Morgan's laws.* The two laws are (prove them!):

1. $(\bigcap_{i\in\mathcal{I}} A_i)^c = \bigcup_{i\in\mathcal{I}} A_i^c$, that is, the complement of the intersection is the union of the complements.
2. $(\bigcup_{i\in\mathcal{I}} A_i)^c = \bigcap_{i\in\mathcal{I}} A_i^c$, that is, the complement of the union is the intersection of the complements.

Finally, the relative complement operation on two sets allows us to "subtract" one from the other.

Definition 1.6 Relative complement*: The relative complement of B in A is defined as $A\backslash B \triangleq \{x|x \in A, x \notin B\} = A \cap B^c$. Similarly, the relative complement of A in B is defined as $B\backslash A \triangleq \{x|x \in B, x \notin A\} = B \cap A^c$.*

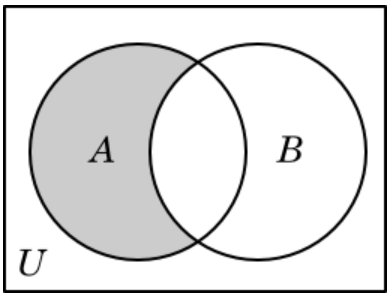

Figure 1.2: Relative complement of B in A

The unary complement operation for a set A can also be understood as the relative complement of A in U, the universal set.

1.2.3 Cartesian products

A Cartesian product is an operation on sets which returns a product set from multiple sets.

Definition 1.7 Cartesian product*: The Cartesian product of 2 sets A and B is defined as $A \times B \triangleq \{(x, y) : x \in A, y \in B\}$, that is, it is the set of all ordered pairs of elements from the two sets, such that the first component belongs to A and the second to B.*

For example, if $A = \{1, 2\}$ and $B = \{a\}$, then $A \times B = \{(1, a), (2, a)\}$ and $B \times A = \{(a, 1), (a, 2)\}$. Clearly, this operation is not commutative. The Cartesian product of n sets $A_1, A_2 \cdots A_n$ is

$$A_1 \times A_2 \cdots \times A_n = \{(a_1, a_2 \cdots, a_n) : a_i \in A_i\}$$

If all the n sets are identical, then we get

$$A^n = \{(a_1, a_2 \cdots, a_n) : a_i \in A\}$$

1.3 Power sets

Definition 1.8 Power set*: The power set of a set A, denoted as $\mathcal{P}(A)$ or 2^A, is the set of all subsets of A including the null set $\emptyset$ and A itself.*

For example, the power set of $A = \{1, 2\}$ is

$$\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$$

A power set is an example of a class, which is a collection of sets and is usually denoted by a script letter, like so: $\mathcal{A}$. The union and intersection operations extend to classes, as does the idea of subsets, in a suitably modified form.

1.4 Functions

Definition 1.9 *A function f from a set A to another set B is a subset of the Cartesian product $(A \times B)$ of the sets such that every element of A is the first component of one and only one ordered pair in the subset. In simple terms, it is a rule that maps every element from set A to a unique element in set B. It is commonly denoted as $f : A \to B$ and A is known as the* domain *while B is known as the* codomain*.*

The element in the codomain (say, b) which is associated with an element in the domain (say, a) is known as the *image* of the element 'a', and 'a' by itself is called the *argument* of the function 'f' and is also termed as *pre-image* of the element 'b'. Then, we say f maps a to b and is represented as $b = f(a)$. The *range* of a function is the set of all elements in the co-domain which are images for elements in the domain, hence, it is the subset (not necessarily proper subset) of the codomain . Functions can be classified as follows:

1. *Injective*: An injective or one-to-one function is one where $a \neq b \Rightarrow f(a) \neq f(b), \forall a, b \in \text{domain}(f)$. E.g., function $f : \mathbb{N} \to \mathbb{R}$ defined as $f(x) = x$, $\forall x \in \mathbb{N}$, is an injective function.
2. *Surjective*: A surjective or onto function is one where $\forall b \in \text{codomain}(f), \exists a \in \text{domain}(f)$ *such that* $f(a) = b$. For example, the following are surjective functions:
   (i) Let $A = \{1, 2, 3\}$ and $B = \{0, 1\}$. The function $g : A \to B$ defined as $g(1) = 0, g(2) = 0$ and $g(3) = 1$ is a surjective function.
   (ii) The function $h : \mathbb{R} \to \mathbb{R}$ defined as $h(x) = x$, $\forall x \in \mathbb{R}$ is also surjective.

A function which is both injective and surjective is known as a bijective function (or a *bijection*). The example of function 'h' stated above is also a bijective function. An inverse function can be defined for a bijection since the mapping is unique and the entire codomain is covered. The notion of a bijection can be used to understand the equicardinality of infinite sets, i.e., when can we say that the "size" of two infinte sets is equal? This question will be answered in the subsequent lectures.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 2: A crash course in Real Analysis

Lecturer: Dr. Krishna Jagannathan *Scribe: Sudharsan Parthasarathy*

This lecture is an introduction to Real Analysis. Here we introduce the important concepts and theorems from Real Analysis that will be useful in the rest of the course. Interested readers may refer the book listed in the References section to learn the proofs of the theorems.

2.1 Notations

$\in$ - belongs to.
$\exists$ - there exists.
$\forall$ - for all.
$\implies$ - implies.
$\mathbb{R}$ - set of real numbers.
$\mathbb{Q}$- set of rational numbers.
$\wedge$ - and.
$\mathbb{N}$ - set of natural numbers.
$\rightarrow$ - converges to.
iff- if and only if.
$\subseteq$ - is a subset of.
ϕ -null set.
$\cap$ - intersection.
i.e. - that is.

2.2 Field

A set X is a field if it satisfies the six properties listed below under the two abstract operations '+' and '.'.

Closure: If a and $b \in X$, then $a+b \in X$ and $a.b \in X$. Hence X is closed under addition and multiplication.

Commutativity: If a and $b \in X$, then $a+b = b+a$ and $a.b = b.a$. Hence X is commutative under addition and multiplication.

Associativity: If a, b and $c \in X$, then $(a+b)+c = a+(b+c)$ and $(a.b).c = a.(b.c)$. Hence X is associative under addition and multiplication.

Identity: If $a \in X$, then $\exists$ elements 0 and 1 in X such that $a+0=a$ and $a.1=a$.

Inverse: If $a \in X$, then $\exists$ elements -a and a^{-1} in X such that $a+(-a)=0$ and $a.a^{-1}=1$. Multiplicative inverse does not exist if $a=0$.

Distributivity: If a, b and $c \in X$, then multiplication is distributive with respect to addition. $a.(b+c)=a.b+a.c$.

Note that, the elements 0 and 1 are unique. If $X \subseteq$ real numbers $\mathbb{R}$, then the elements in the field can also be compared. A field whose elements can be compared is called an ordered field. Another example for an

ordered field is a set of rational numbers $\mathbb{Q}$. Henceforth we will concentrate only on the real field $\mathbb{R}$.

2.2.1 Order axioms

Law of trichotomy: If a, $b \in \mathbb{R}$, then a=b or $a > b$ or $a < b$.

Transitivity: Let a, b, $c \in \mathbb{R}$. If $a > b$ and $b > c$, then $a > c$.

Ordering and addition operator: Let a, b, $c \in \mathbb{R}$. $a > b \implies a+c > b+c$.

Ordering and product operator: Let a, b, $c \in \mathbb{R}$. $a > b \implies ac > bc$, if $c > 0$.

2.3 Boundedness

A subset S of $\mathbb{R}$ is bounded above if $\exists$ a real number M such that $x \leq M$, $\forall\ x \in S$. Here, M is called an upper bound of S. Similarly S is bounded below if $\exists$ a real number m such that $x \geq m$, $\forall\ x \in S$. Here, m is called a lower bound of S. A set is bounded if it is both bounded above and below. Any element greater than M and lesser than m are also upper and lower bounds of S respectively.

Supremum: *The supremum of S is the least upper bound of the set S.* More precisely, K is a supremum of S if

- K is an upper bound of S, *i.e.* , $x \leq K$, $\forall\ x \in S$.
- There exists no number less than K which is an upper bound of S, *i.e.* for any $\delta > 0$, $\exists\ z \in S$ such that $z > K$-δ.

Similarly one can define the infimum, as the greatest lower bound of a set. It is important to note that supremum and infimum need not be elements of the set. For instance, 1 is the supremum of the set (0,1), but is not an element of the set. Also, if the supremum is an element of the set itself, then it is the maximum of that set.

2.4 Completeness property

The completeness axiom or the least upper bound property is one of the fundamental properties of the real field $\mathbb{R}$.

Completeness Axiom: *Any non empty subset A of $\mathbb{R}$ which is bounded above has a supremum in $\mathbb{R}$.*

In other words, the Completeness Axiom guarantees that, for any nonempty set of $\mathbb{R}$ that is bounded above, a supremum exists. Although $\mathbb{R}$ and $\mathbb{Q}$ are ordered fields, we will see in the exercise below that the latter does not satisfy the completeness property. Indeed, completeness along with the ordered field property characterizes $\mathbb{R}$. Thus, $\mathbb{R}$ is also referred to as a complete ordered field.

We now list a few important theorems (without proofs), which are consequences of the completeness property.

Theorem 2.1 *If x and y are any two positive real numbers, then there exists a positive integer m such that $mx > y$. This is called as the Archimedean property of real numbers.*

Theorem 2.2 *Every open interval contains a rational number.*

Theorem 2.3 *Let $x \in \mathbb{R}$, $n \geq 2$, $n \in \mathbb{N}$, then*

- *If $x \geq 0$ and n is even then $\exists$ a unique $y \geq 0$ such that $y^n = x$.*
- *If $x \in \mathbb{R}$ and n is odd then $\exists$ a unique $y \in \mathbb{R}$ such that $y^n = x$.*

2.5 Sequences

A (real) sequence is a function from $\mathbb{N}$ to $\mathbb{R}$. A sequence $\{x_n\}$ of real numbers is said to converge to $x \in \mathbb{R}$ if for every $\epsilon > 0$, $\exists$ a natural number n_0 such that $|x_n - x| < \epsilon \; \forall \; n \geq n_0$.

Theorem 2.4 *Let $\{x_n\}$ be a monotonically increasing sequence such that $x_n \leq \alpha$ for some $\alpha \in \mathbb{R}$, and all $n \geq 1$. Then $\{x_n\}$ converges to a real number.*

In other words, the above theorem can be stated as: a monotonically non-decreasing sequence which is bounded above converges. The proof again uses the completeness property. The student is encouraged to attempt a proof of this theorem, before referring to a text.

Corollary 2.5 *A convergent sequence is bounded.*

Of course, a bounded sequence need not converge: consider for example, the sequence $x_n = \{(-1)^n\}$. The sequence is bounded, but does not converge. Next, we list some elementary properties of limits. Let $\{x_n\}$ and $\{y_n\}$ be two sequences that converge to x and y, respectively.

- $x_n + y_n \to x + y$.
- $\alpha x_n \to \alpha x$, $\forall \; \alpha \in \mathbb{R}$.
- $x_n y_n \to xy$.
- $x_n \geq 0 \; \forall \; n \implies x \geq 0$.
- $x_n \leq y_n \; \forall \; n \implies x \leq y$.
- If $y \neq 0$, $\frac{x_n}{y_n} \to \frac{x}{y}$.

Theorem 2.6 Sandwich/ Two policeman theorem:
If $x_n \leq z_n \leq y_n \; \forall \; n$ and if x_n and y_n converge to x, then z_n also converges to x (Prove it!).

Examples of some important sequences are as follows:

Cauchy Sequences: A sequence $\{x_n\}$ is called a Cauchy sequence if $\forall \; \epsilon > 0$, $\exists \; n_0 \in \mathbb{N}$ such that $|x_n - x_m| < \epsilon \; \forall \; n, \; m \geq n_0$.

Subsequences: A subsequence of a sequence is an infinite ordered subset of that sequence. Here are a few basic theorems about subsequences.

Theorem 2.7 *Every real sequence has a monotonic subsequence.*

Theorem 2.8 Bolzano-Weistrass Theorem: *Every bounded sequence has a convergent subsequence.*

Theorem 2.9 *A sequence $\{x_n\}$ is convergent iff $\{x_n\}$ is bounded and every convergent subsequence of $\{x_n\}$ converges to the same limit.*

Theorem 2.10 *A real sequence is convergent iff it is a Cauchy sequence.*

2.6 Metric Spaces

A set X is a metric space if we can associate a real number $d(a, b)$ with any two elements a and b of the set X such that

- $d(a, b) > 0$ if $a \neq b$; $d(a, a)$=0.
- $d(a, b)$=$d(b, a)$.
- Triangle inequality: $d(a, b) \leq d(a, c) + d(b, c)$ for any $c \in X$.

Any function d that satisfies these properties on a set is called a metric.

2.6.1 Open set

Let (X,d) be a metric space. The *open ball* $B(x, r)$ centred at x of radius r is defined as $B(x, r) = \{y \in X : d(x, y) < r\}$. A set $A \subseteq X$ is said to be *open* in X if for every $x \in A$, $\exists\ r > 0$ such that $B(x, r) \subseteq A$.

Theorem 2.11 *Let (X,d) be a metric space, then*

- *X and the null set ϕ are open in X.*
- *An arbitrary union of open sets is open.*
- *A finite intersection of open sets is open.*

Definition 2.12 *Interior point: Let (X,d) be a metric space and $A \subseteq X$. A point $x \in X$ is called an interior point of A if there exists $r > 0$, such that $B(x, r) \subseteq A$.*

Let A^0 denote the set of all interior points of A. Clearly, $A^0 \subseteq A$.

Lemma 2.13 *Let (X,d) be a metric space, then*

- *A^0 is open in X.*
- *A^0 is the largest open set contained in A.*
- *A^0=A iff A is open.*

2.6.2 Closed set

Let (X,d) be a metric space and $A \subseteq X$. A is said to be *closed* in X if A^c is open in X.

Theorem 2.14 *Let (X,d) be a metric space, then*

- *X and the null set ϕ are closed in X.*
- *An arbitrary intersection of closed sets is closed.*
- *A finite union of closed sets is closed.*

Definition 2.15 *Limit point: Let (X,d) be a metric space and $A \subseteq X$. A point $x \in X$ is called a limit point of A, if for every $r > 0$, $B(x,r)$ contains at least one point of A.*

The *closure* of A, denoted $\overline{A}$, is defined as the set of all limit points of A. Clearly, $A \subseteq \overline{A}$.

Lemma 2.16 *Let (X,d) be a metric space, then*

- *$\overline{A}$ is closed in X.*
- *$\overline{A}$ is the smallest closed set containing A.*
- *$\overline{A}$=A iff A is closed.*

2.6.3 Compact set

A subset A of a metric space X is compact if every sequence in A has a convergent subsequence in A.

Theorem 2.17 Heine-Borel Theorem:
In any Euclidean space $\mathbb{R}^d$, a set A is compact iff it is closed and bounded.

2.7 Functions

A function f maps every element in set A to a unique element in set B. Let $(X,\ d_X)$ and $(Y,\ d_Y)$ be two metric spaces. Let A be a subset of X and $a \in A$, and let f be a function from A to Y. The function f is said to be continuous at a if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $d_Y(f(x), f(a)) < \epsilon$ for all points $x \in A$ for which $d_X(x,a) < \delta$. If f is continuous at every point on X, then f is continuous on X.

Theorem 2.18 *f is continuous iff $f(x_n)$ converges to $f(x)$ in Y whenever the sequence x_n converges to x in X.*

Theorem 2.19 *A function f that maps a metric space X into a metric space Y is continuous on X iff $f^{-1}(B)$ is open in X for every open set B in Y. ($f^{-1}(B)$ is the inverse image of set B. f^{-1} does not mean inverse function here.)*

Theorem 2.20 *A function f that maps a metric space X into a metric space Y is continuous on X iff $f^{-1}(B)$ is closed in X for every closed set B in Y.*

Theorem 2.21 *If function f is a continuous mapping of a compact metric space X into a metric space Y, then $f(X)$ is compact.*

2.8 Exercises

1. Prove the uniqueness of the supremum and infimum of a set.

2. Let $S = \{x : x \in \mathbb{Q}, x > 0 \wedge x^2 < 2\}$ be a subset of $\mathbb{Q}$. Show that S has no rational supremum. This shows that the completeness axiom does not hold for $\mathbb{Q}$.

3. Let f and g be continuous functions on metric space X, then $f + g$ and fg are continuous on X.

References

[WR] Walter Rudin, "Principles of Mathematical Analysis," *McGraw Hill International Series*, Third Edition.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 3: Cardinality and Countability

Lecturer: Dr. Krishna Jagannathan *Scribe: Ravi Kiran Raman*

3.1 Functions

We recall the following definitions.

Definition 3.1 *A* **function** $f : A \to B$ *is a rule that maps every element of set* A *to a unique element in set* B*.*

In other words, $\forall x \in A, \exists y \in B$ and only one such element, such that, $f(x) = y$. Then y is called the image of x and x, the pre-image of y under f. The set A is called the domain of the function and B, the co-domain. $\mathcal{R} = \{y : \exists x \in A,\ s.t.\ f(x) = y\}$ is called as the range of the function f.

Definition 3.2 *A function* $f : A \to B$ *is said to be an* **injective (one-to-one) function***, if every element in the range* $\mathcal{R}$ *has a unique pre-image in* A*.*

Definition 3.3 *A function* $f : A \to B$ *is said to be a* **surjective (onto) function***, if* $\mathcal{R} = B$*, i.e,* $\forall y \in B,\ \exists x \in A,\ s.t.\ f(x) = y$*.*

Definition 3.4 *A function* $f : A \to B$ *is a* **bijective function** *if it is both injective and surjective.*

Hence, in a bijective mapping, every element in the co-domain has a pre-image and the pre-images are unique. Thus, we can define an inverse function, $f^{-1} : B \to A$, such that, $f^{-1}(y) = x$, if $f(x) = y$. In simple terms, bijective functions have well-defined inverse functions.

3.2 Cardinality and Countability

In informal terms, the cardinality of a set is the number of elements in that set. If one wishes to compare the cardinalities of two finite sets A and B, it can be done by simply counting the number of elements in each set, and declare either that they have equal cardinality, or that one of the sets has more elements than the other. However, when sets containing infinitely many elements are to be compared(for example, $\mathbb{N}$ versus $\mathbb{Q}$), this elementary approach is not efficient to do it. In the late nineteenth century, Georg Cantor introduced the idea of comparing the cardinality of sets based on the nature of functions that can be possibly defined from one set to another.

Definition 3.5 *(i) Two sets* A *and* B *are* **equicardinal** *(notation* $|A| = |B|$*) if there exists a bijective function from* A *to* B*.*

(ii) B *has cardinality greater than or equal to that of* A *(notation* $|B| \geq |A|$*) if there exists an injective function from* A *to* B*.*

(iii) B has cardinality strictly greater than that of A (notation $|B| > |A|$) if there is an injective function, but no bijective function, from A to B.

Having stated the definitions as above, the definition of countability of a set is as follow:

Definition 3.6 *A set E is said to be* **countably infinite** *if E and $\mathbb{N}$ are equicardinal. And, a set is said to be* **countable** *if it is either finite or countably infinite.*

The following are some examples of countable sets:
1. *The set of all integers $\mathbb{Z}$ is countably infinite.*
We can define the bijection $f : \mathbb{Z} \to \mathbb{N}$ as follows :

$n = f(z) \in \mathbb{N}$	$z \in \mathbb{Z}$
1	0
2	+1
3	-1
4	+2
5	-2
.	.
.	.
.	.

The existence of this bijective map from $\mathbb{Z}$ to $\mathbb{N}$ proves that $\mathbb{Z}$ is countably infinite.

2. *The set of all rationals in $[0, 1]$ is countable.*
Consider the rational number $\frac{p}{q}$ where $q \neq 0$. Increment q in steps of 1 starting with 1. For each such q and $0 \leq p \leq q$, add the rational number $\frac{p}{q}$ to the set, if it not already present. By this way, the set of rational numbers in $[0, 1]$ can be explicitly listed as: $\{0, 1, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, ...\}$
Clearly, we can define a bijection from $\mathbb{Q} \cap [0, 1] \to \mathbb{N}$ where each rational number is mapped to its index in the above set. Thus the set of all rational numbers in $[0, 1]$ is countably infinite and thus countable.

3. *The set of all Rational numbers, $\mathbb{Q}$ is countable.*
In order to prove this, we state an important theorem, whose proof can be found in [1].

Theorem 3.7 *Let $\mathcal{I}$ be a countable index set, and let E_i be countable for each $i \in \mathcal{I}$. Then $\bigcup_{i \in \mathcal{I}} E_i$ is countable.* More glibly, it can also be stated as follows: *A countable union of countable sets is countable.*

We will now use this theorem to prove the countability of the set of all rational numbers. It has been already proved that the set $\mathbb{Q} \cap [0, 1]$ is countable. Similarly, it can be showed that $\mathbb{Q} \cap [n, n+1]$ is countable, $\forall n \in \mathbb{Z}$. Let $Q_i = \mathbb{Q} \cap [i, i + 1]$. Thus, clearly, the set of all rational numbers, $\mathbb{Q} = \cup_{i \in \mathbb{Z}} Q_i$ – a countable union of countable sets – is countable.

Remark: For two finite sets A and B, we know that if A is a strict subset of B, then B has cardinality greater than that of A. As the above examples show, this is not true for infinite sets. Indeed, $\mathbb{N}$ is a strict subset of $\mathbb{Q}$, but $\mathbb{N}$ and $\mathbb{Q}$ are equicardinal!

4. The set of all *algebraic* numbers (numbers which are roots of polynomial equations with rational coefficients) is countable.

5. The set of all computable numbers, i.e., real numbers that can be computed to within any desired precision by a finite, terminating algorithm, is countable (see Wikipedia article for more details).

Definition 3.8 *A set F is* **uncountable** *if it has cardinality strictly greater than the cardinality of* $\mathbb{N}$.

In the spirit of Definition 3.5, this means that F is uncountable if an injective function from $\mathbb{N}$ to F exists, but no such bijective function exists.

An interesting example of an uncountable set is the set of all infinite binary strings. The proof of the following theorem uses the celebrated 'diagonal argument' of Cantor.

Theorem 3.9 (Cantor) *: The set of all infinite binary strings, $\{0,1\}^\infty$, is uncountable.*

Proof: It is easy to show that an injection from $\mathbb{N}$ to $\{0,1\}^\infty$ exists (exercise: produce one!). We need to show that no such bijection exists.

Let us assume the contrary, i.e, let us assume that the set of all binary strings, $A = \{0,1\}^\infty$ is countably infinite. Thus there exists a bijection $f : A \to \mathbb{N}$. In other words, we can order the set of all infinite binary strings as follows:

$$\begin{matrix} a_{11} & a_{12} & a_{13} & . & . & . \\ a_{21} & a_{22} & a_{23} & . & . & . \\ a_{31} & a_{32} & a_{33} & . & . & . \\ . & . & . & & & \\ . & . & . & & & \\ . & . & . & & & \end{matrix} \quad ,$$

where, a_{ij} is the j^{th} bit of the i^{th} binary string, $i, j \geq 1$.

Consider the infinite binary string given by $\bar{a} = \bar{a_{11}}\bar{a_{22}}\bar{a_{33}}...$, where $\bar{a_{ij}}$ is the complement of the bit a_{ij}.

Since our list contains *all* infinite binary strings, there must exist some $k \in \mathbb{N}$ such that the string $\bar{a}$ occurs at the k position in the list, i.e., $f(\bar{a}) = k$. The k^{th} bit of this specific string is $\bar{a_{kk}}$. However, from the above list, we know that the k^{th} bit of the k^{th} string is a_{kk}. Thus, we can conclude that the string $\bar{a}$ cannot occur in any position $k \geq 1$ in our list, contradicting our initial assumption that our list exhausts all possible infinite binary strings.

Thus, there cannot possibly exist a bijection from $\mathbb{N}$ to $\{0,1\}^\infty$, proving that $\{0,1\}^\infty$ is uncountable. ■

Now using Cantor's theorem, we will prove that the set of irrational numbers is uncountable.

Claim 3.10 *The sets $[0,1]$, $\mathbb{R}$ and $\{\mathbb{R} \setminus \mathbb{Q}\}$ are uncountable.*

Proof: Firstly, consider the set $[0,1]$. Any number in this set can be expressed by its binary equivalent and thus, there appears to be a bijection from $[0,1] \to \{0,1\}^\infty$. However, this is not exactly a bijection as there is a problem with the dyadic rationals (i.e., numbers of the form $\frac{a}{2^b}$, where a and b are natural numbers, and a is odd). For example, 0.01000.... in binary is the same as 0.001111.... . However we can tweak this "near bijection" to produce an explicit bijection in the following way. For any infinite binary string $x = (x_1, x_2, \dots) \in \{0,1\}^\infty$, let

$$g(x) = \sum_{k=1}^{\infty} x_k 2^{-k}.$$

The function g maps $\{0,1\}^\infty$ "almost bijectively" to $[0,1]$, but unfortunately, the dyadic rationals have two pre-images. For example we have $g(1000...) = g(0111...) = \frac{1}{2}$. To fix this let the the set of dyadic rationals be diven by the list

$$\mathcal{D} = \left\{ d_1 = \frac{1}{2}, d_2 = \frac{1}{4}, d_3 = \frac{3}{4}, d_4 = \frac{1}{8}, d_5 = \frac{3}{8}, d_6 = \frac{5}{8}, d_7 = \frac{7}{8}, ... \right\}$$

Note that the dyadic rationals can be put in a list as given above as they are countable. Next, we define the following bijection $f(x)$ from $\{0,1\}^\infty$ to $[0,1]$.

$$f(x) = \begin{cases} g(x) & \text{if } g(x) \notin \mathcal{D}, \\ d_{2n-1} & \text{if } g(x) = d_n \text{ for some } n \in \mathbb{N} \text{ and } x_k \text{ terminates in 1}, \\ d_{2n} & \text{if } g(x) = d_n \text{ for some } n \in \mathbb{N} \text{ and } x_k \text{ terminates in 0}. \end{cases}$$

This is an explicit bijection from $\{0,1\}^\infty$ to $[0,1]$ which proves that the set $[0,1]$ is uncountable. (Why?)

Next, we can define a bijection from $(0,1) \to \mathbb{R}$, for instance using the function $\tan(\pi x - \frac{\pi}{2}),\ x \in (0,1)$. Thus the set of all real numbers, $\mathbb{R}$ is uncountable.

Finally, we can write, $\mathbb{R} = \mathbb{Q} \cup \{\mathbb{R} \setminus \mathbb{Q}\}$. Since $\mathbb{Q}$ is countable and $\mathbb{R}$ is uncountable, we can easily argue that $\{\mathbb{R} \setminus \mathbb{Q}\}$, i.e, the set of all irrational numbers, is uncountable. ■

3.3 Exercises

1. Prove that $2^{\mathbb{N}}$, the power set of the natural numbers, is uncountable. (Hint: Try to associate an infinite binary string with each subset of $\mathbb{N}$.)

2. Prove that the Cartesian product of two countable sets is countable.

3. Let A be a countable set, and B_n be the set of all n-tuples $(a_1, ..., a_n)$, where $a_k \in A (k = 1, 2, ..., n)$ and the elements $a_1, a_2, ..., a_n$ need not be distinct. Show that B_n is countable.

4. Show that an infinite subset of a countable set is countable.

5. A number is said to be an algebraic number if it is a root of some polynomial equation with integer coefficients. For example, $\sqrt{2}$ is algebraic since it is a root of the polynomial $x^2 - 2$. However, it is known that π is not algebraic. Show that the set of all algebraic numbers is countable. Also, a transcendental number is a real number that is not algebraic. Are the transcendental numbers countable?

6. The *Cantor* set is an interesting subset of [0, 1], which we will encounter several times in this course. One way to define the Cantor set C is as follows. Consider the set of all real numbers in $[0,1]$ written down in ternary (base-3) expansion, instead of the usual decimal (base-10) expansion. A real number $x \in [0,1]$ belongs to C iff x admits a ternary expansion without any 1s. Show that C is uncountably infinite, and that it is indeed equi-cardinal with $[0,1]$.

References

[1] Walter Rudin, "Principles of Mathematical Analysis," *McGraw Hill International Series*, Third Edition.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 4: Probability Spaces

Lecturer: Dr. Krishna Jagannathan *Scribe: Jainam Doshi, Arjun Nadh and Ajay M*

4.1 Introduction

Just as a point is not defined in elementary geometry, probability theory begins with two entities that are not defined. These undefined entities are a **Random Experiment** and its **Outcome**. These two concepts are to be understood intuitively, as suggested by their respective English meanings. We use these undefined terms to define other entities.

Definition 4.1 *The* **Sample Space** Ω *of a random experiment is the set of all possible outcomes of a random experiment.*

An outcome (or elementary outcome) of the random experiment is usually denoted by ω. Thus, when a random experiment is performed, the outcome $\omega \in \Omega$ is picked by the Goddess of Chance or Mother Nature or your favourite genie.

Note that the sample space Ω can be finite or infinite. Indeed, depending on the cardinality of Ω, it can be classified as follows:

1. Finite sample space
2. Countably infinite sample space
3. Uncountable sample space

It is imperative to note that for a given random experiment, its sample space is defined depending on what one is interested in observing as the outcome. We illustrate this using an example. Consider a person tossing a coin. This is a random experiment. Now consider the following three cases:

- Suppose one is interested in knowing whether the toss produces a head or a tail, then the sample space is given by, $\Omega = \{H, T\}$. Here, as there are only two possible outcomes, the sample space is said to be finite.

- Suppose one is interested in the number of tumbles before the coin hits the ground, then the sample space is the set of all natural numbers. In this case, the sample space is countably infinite and is given by, $\Omega = \mathbb{N}$.

- Suppose one is interested in the speed with which the coin strikes ground, then the set of positive real numbers forms the sample space. This is an example of an uncountable sample space, which is given by, $\Omega = \mathbb{R}^+$.

Thus we see that for the same experiment, Ω can be different based on what the experimenter is interested in.

Let us now have a look at one more example where the sample space can be different for the same experiment and you can get different answers based on which sample space you decide to choose.

Bertrand's Paradox: Consider a circle of radius r. What is the probability that the length of a chord chosen at random is greater than the length of the side of an equilateral triangle inscribed in the circle?

This is an interesting paradox and gives different answers based on different sample spaces. The entire description of Bertrand's Paradox can be found in [1].

Definition 4.2 *(Informal) An* **event** *is a subset of the sample space, to which probabilities will be assigned.*

An event is a subset of the sample space, but we emphasise that *not all subsets of the sample space are necessarily considered events*, for reasons that will be explained later. Until we are ready to give a more precise definition, we can consider events to be those "interesting" subsets of Ω, to which we will eventually assign probabilities. We will see later that whenever Ω is finite or countable, all subsets of the sample space can be considered as events, and be assigned probabilities. However, when Ω is uncountable, it is often not possible to assign probabilities to all subsets of Ω, for reasons that will not be clear now. The way to handle uncountable sample spaces will be discussed later.

Definition 4.3 *An event A is said to* **occur** *if the outcome ω, of the random experiment is an element of A, i.e., if $\omega \in A$.*

Let us take an example. Say the random experiment is choosing a card at random from a pack of playing cards. What is the sample space in this case? It is a 52 element set as each card is a possible outcome. As the sample space is finite, any subset of the sample space can be considered as an event. As a result there will be 2^{52} events (Power set of n elements has 2^n elements). An event can be any subset of the sample space which includes the empty set, all the singleton sets (containing one outcome) and collection of more than one outcomes. Listed below are a few events:

- The 7 of Hearts (1 element)
- A face Card (12 elements)
- A 2 and a 7 at the same time (0 element)
- An ace of any color (4 elements)
- A diamond card (13 elements)

Next, let us look at some nice properties that we would expect events to satisfy:

- Since the sample space Ω always occurs, we would like to have Ω as an event.
- If A is an event (i.e., a "nice" subset of the sample space to which we would like to assign a probability), it is reasonable to expect A^c to be an event as well.
- If A and B are two events, we are interested in the occurrence of at least one of them (A or B) as well as the occurrence of both of them (A and B). Hence, we would like to have $A \cup B$ and $A \cap B$ to be events as well.

The above three properties motivate a mathematical structure of subsets, known as an *algebra.*

4.2 Algebra, $\mathcal{F}_0$

Let Ω be the sample space and let $\mathcal{F}_0$ be a collection of subsets of Ω. Then, $\mathcal{F}_0$ is said to be an **algebra** (or a field) if

i. $\emptyset \in \mathcal{F}_0$.

ii. $A \in \mathcal{F}_0$, implies $A^c \in \mathcal{F}_0$.

iii. $A \in \mathcal{F}_0$ and $B \in \mathcal{F}_0$ implies $A \cup B \in \mathcal{F}_0$.

It can be shown that an algebra is closed under finite union and finite intersection (*see Exercise 1(a)).* However, a natural question that arises at this point is "Is the structure of an algebra enough to study events of typical interest?" Consider the following example:

Example:- Toss a coin repeatedly until the first heads shows. Here, Ω = {H, TH, TTH, ... }. Let us say that we are interested in determining if the number of tosses before seeing a head is even. It is easy to see that this 'event' of interest will not be included in the algebra. This is because an algebra contains only finite unions of subsets, but the 'event' of interest entails a countably infinite union. This motivates the definition of a *σ-algebra.*

4.3 Sigma Algebra, $\mathcal{F}$

A collection $\mathcal{F}$ of subsets of Ω is called a **σ-algebra** (or σ-field) if

i. $\emptyset \in \mathcal{F}$.

ii. A $\in \mathcal{F}$, implies $A^c \in \mathcal{F}$.

iii. If A_1, A_2, A_3, ... is a countable collection of subsets in $\mathcal{F}$, then $\bigcup_{i=1}^{\infty} Ai \in \mathcal{F}$.

Note that unlike an algebra, a σ-algebra is closed under countable union and countable intersection (*see Exercise 1(b)*). Some examples of σ-algebra are:

i. $\{\emptyset, \Omega\}$

ii. $\{\emptyset, A, A^c, \Omega\}$

iii. Power set of Ω, denoted by 2^{Ω}.

The 2-tuple $(\Omega, \mathcal{F})$ is called a *measurable space.* Also, every member of the σ-algebra $\mathcal{F}$ is called an $\mathcal{F}$-measurable set in the context of measure theory. In the specific context of probability theory, $\mathcal{F}$-measurable sets are called *events.* Thus, whether or not a subset of Ω is considered an event depends on the σ-algebra that is under consideration.

4.4 Measure

We now proceed to define measures and measure spaces. We will see that a probability space is indeed a special case of a measure space.

Definition 4.4 *Let $(\Omega, \mathcal{F})$ be a measurable space. A* measure *on $(\Omega, \mathcal{F})$ is a function μ: $\mathcal{F} \longrightarrow [0,\infty]$ such that*

i. $\mu(\emptyset) = 0$.

ii. If $\{A_i, i \geq 1\}$ is a sequence of disjoint sets in $\mathcal{F}$, then the measure of the union (of countably infinite disjoint sets) is equal to the sum of measures of individual sets, i.e.,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \tag{4.1}$$

The second property stated above is known as the *countable additivity* property of measures. From the definition, it is clear that a measure can only be assigned to elements of $\mathcal{F}$. The triplet $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*. μ is said to be a *finite measure* if $\mu(\Omega) < \infty$; otherwise, μ is said to be an *infinite measure*. In particular, if $\mu(\Omega) = 1$, then μ is said to be a *probability measure*. Next, we state this explicitly for pedagogical completeness.

4.5 Probability Measure

A *probability measure* is a function $\mathbb{P}$: $\mathcal{F} \longrightarrow [0,1]$ such that

i. $\mathbb{P}(\emptyset) = 0$.

ii. $\mathbb{P}(\Omega) = 1$.

iii. (*Countable additivity:*) If $\{A_i, i \geq 1\}$ is a sequence of disjoint sets in $\mathcal{F}$, then

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

The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*, and the three properties, stated above, are sometimes referred to as the axioms of probability.

Note:- It is clear from the definition that probabilities are defined only to elements of $\mathcal{F}$, and not necessarily to all subsets of Ω. In other words, probability measures are assigned only to events. Even when we speak of the probability of an elementary outcome ω, it should be interpreted as the probability assigned to the singleton set $\{\omega\}$ (assuming of course, that the singleton is an event).

4.6 Exercises

1. a) Let $A_1, A_2, \ldots, A_n$ be a finite collection of subsets of Ω such that $A_i \in \mathcal{F}_0$ (an algebra), $1 \leq i \leq n$. Show that $\bigcup_{i=1}^{n} A_i \in \mathcal{F}_0$ and $\bigcap_{i=1}^{n} A_i \in \mathcal{F}_0$. Hence, infer that an algebra is closed under finite union and finite intersection.

   b) Suppose A_1, A_2, A_3, ... is a countable collection of subsets in the σ-algebra $\mathcal{F}$, then show that $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

2. [**σ-algebra : Properties and Construction**].
   (a) Show that a σ-algebra is also an algebra.
   (b) Given a sample space Ω and a σ-algebra $\mathcal{F}$ of the subsets of Ω, show that if $A, B \in \mathcal{F}$, $A \setminus B$ and $A \triangle B$, the symmetric difference of A and B are present in $\mathcal{F}$.
   (c) Consider the random experiment of throwing a die. If a statistician is interested in the occurrence of either an odd or an even outcome, construct a sample space and a σ-algebra of subsets of this sample space.
   (d) Let $A_1, A_2, ..., A_n$ be arbitrary subsets of Ω. Describe (explicitly) the smallest σ-algebra $\mathcal{F}$ containing $A_1, A_2, ..., A_n$. How many sets are there in $\mathcal{F}$? (Give an upper bound that is attainable under certain conditions). List all the sets in $\mathcal{F}$ for $n = 2$.
3. Let $\mathcal{F}$ and $\mathcal{G}$ be two σ-algebras of subsets of Ω.
   (a) Is $\mathcal{F} \cup \mathcal{G}$, the collection of subsets of Ω lying in either $\mathcal{F}$ or $\mathcal{G}$ a σ-algebra?
   (b) Show that $\mathcal{F} \cap \mathcal{G}$, the collection of subsets of Ω lying in both $\mathcal{F}$ and $\mathcal{G}$ is a σ-algebra.
   (c) Generalize (b) to arbitrary intersections as follows. Let $\mathcal{I}$ be an arbitrary index set (possibly uncountable), and let $\{\mathcal{F}_i\}_{i \in \mathcal{I}}$ be a collection of σ-algebras on Ω. Show that $\bigcap_{i \in \mathcal{I}} \mathcal{F}_i$ is also a σ-algebra.
4. Let $\mathcal{F}$ be a σ-algebra of subsets of Ω, and let $B \in \mathcal{F}$. Show that
$$\mathcal{G} = \{A \cap B \mid A \in \mathcal{F}\}$$
is a σ-algebra of subsets of B.
5. Let X and Y be two sets and let $f : X \to Y$ be a function. If $\mathcal{F}$ is a σ-algebra over the subsets of Y and $\mathcal{G} = \{ A \mid \exists B \in \mathcal{F}$ such that $f^{-1}(B) = A \}$, does $\mathcal{G}$ form a σ-algebra of subsets of X ?
Note that $f^{-1}(N)$ is the notation used for the pre-image of set N under the function f for some $N \subseteq Y$. That is, $f^{-1}(N) = \{x \in X | f(x) \in N\}$ for some $N \subseteq Y$.
6. Let Ω be an arbitrary set.
   (a) Is the collection $\mathcal{F}_1$ consisting of all finite subsets of Ω an algebra?
   (b) Let $\mathcal{F}_2$ consist of all finite subsets of Ω, and all subsets of Ω having a finite complement. Is $\mathcal{F}_2$ an algebra?
   (c) Is $\mathcal{F}_2$ a σ-algebra?
   (d) Let $\mathcal{F}_3$ consist of all countable subsets of Ω, and all subsets of Ω having a countable complement. Is $\mathcal{F}_3$ a σ-algebra?

References

[1] Sheldon Ross, "A First Course in Probability," *Pearson*, 8th Edition.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 5: Properties of Probability Measures

Lecturer: Dr. Krishna Jagannathan *Scribe: Ajay M, Gopal Krishna Kamath M*

5.1 Properties

In this lecture, we will derive some fundamental properties of probability measures, which follow directly from the axioms of probability. In what follows, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

- **Property 1:-** Suppose A be a subset of Ω such that $A \in \mathcal{F}$. Then,

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A). \tag{5.1}$$

Proof:- Given any subset $A \in \Omega$, A and A^c partition the sample space. Hence, $A^c \cup A = \Omega$ and $A^c \cap A = \emptyset$. By the "*Countable Additivity*" axiom of probability, $\mathbb{P}(A^c \cup A) = \mathbb{P}(A) + \mathbb{P}(A^c)$ $\implies \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c) \implies \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

- **Property 2:-** Consider events A and B such that $A \subseteq B$ and $A, B \in \mathcal{F}$. Then $\mathbb{P}(A) \leq \mathbb{P}(B)$
Proof:- The set B can be written as the union of two disjoint sets A and $A^c \cap B$. Therefore, we have $\mathbb{P}(A) + \mathbb{P}(A^c \cap B) = \mathbb{P}(B) \implies \mathbb{P}(A) \leq \mathbb{P}(B)$ since $\mathbb{P}(A^c \cap B) \geq 0$.

- **Property 3:-** (*Finite Additivity*) If A_1, A_2, ..., A_n are finite number of disjoint events, then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i). \tag{5.2}$$

Proof:- This property follows directly from the axiom of *countable additivity* of probability measures. It is obtained by setting the events A_{n+1}, A_{n+2}, ... as empty sets. LHS will simplify as:

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

RHS can be manipulated as follows:

$$\begin{aligned}
\sum_{i=1}^{\infty} \mathbb{P}(A_i) &\overset{(a)}{=} \lim_{k\to\infty} \sum_{i=1}^{k} \mathbb{P}(A_i) \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i) + \lim_{k\to\infty} \sum_{i=n+1}^{k} \mathbb{P}(A_i) \\
&\overset{(b)}{=} \sum_{i=1}^{n} \mathbb{P}(A_i) + \lim_{k\to\infty} 0 \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i).
\end{aligned}$$

where (a) follows from the definition of an infinite series and (b) is a consequence of setting the events from A_{n+1} onwards to null sets.

- **Property 4:-** For any A, $B \in \mathcal{F}$,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \tag{5.3}$$

In general, for a family of events $\{A_i\}_{i=1}^n \subset \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) + ... + (-1)^{n+1}\mathbb{P}\left(\bigcap_{i=1}^n A_i\right). \tag{5.4}$$

This property is proved using induction on n. The property can be proved in a much more simpler way using the concept of *Indicator Random Variables*, which will be discussed in the subsequent lectures.
Proof of Eq (5.3)**:-** The set $A \cup B$ can be written as $A \cup B = A \cup (A^c \cap B)$. Since A and $A^c \cap B$ are disjoint events, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$. Now, set B can be partitioned as, $B = (A \cap B) \cup (A^c \cap B)$. Hence, $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B)$. On substituting this result in the expression of $\mathbb{P}(A \cup B)$, we will obtain the final result that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

- **Property 5:-** If $\{A_i, i \geq 1\}$ are events, then

$$\mathbb{P}\left(\bigcup_{i=1}^\infty A_i\right) = \lim_{m \to \infty} \mathbb{P}\left(\bigcup_{i=1}^m A_i\right). \tag{5.5}$$

This result is known as *continuity of probability measures.*

Proof:- Define a new family of sets $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, ..., $B_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i$, ...
Then, the following claims are placed:
Claim 1:- $B_i \cap B_j = \emptyset$, $\forall i \neq j$.
Claim 2:- $\bigcup_{i=1}^\infty A_i = \bigcup_{i=1}^\infty B_i$.
Since $\{B_i, i \geq 1\}$ is a disjoint sequence of events, and using the above claims, we get

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

Therefore,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^\infty A_i\right) &= \sum_{i=1}^\infty \mathbb{P}(B_i) \\
&\overset{(a)}{=} \lim_{m \to \infty} \sum_{i=1}^m \mathbb{P}(B_i) \\
&\overset{(b)}{=} \lim_{m \to \infty} \mathbb{P}\left(\bigcup_{i=1}^m B_i\right) \\
&\overset{(c)}{=} \lim_{m \to \infty} \mathbb{P}\left(\bigcup_{i=1}^m A_i\right).
\end{aligned}$$

Here, (a) follows from the definition of an infinite series, (b) follows from *Claim 1* in conjunction with *Countable Additivity* axiom of probability measure and (c) follows from the intermediate result required to prove *Claim 2*.
Hence proved.

- **Property 6:-** If $\{A_i, i \geq 1\}$ is a sequence of increasing nested events i.e. $A_i \subseteq A_{i+1}$, $\forall i \geq 1$, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{m\to\infty} \mathbb{P}\left(A_m\right). \tag{5.6}$$

- **Property 7:-** If $\{A_i, i \geq 1\}$ is a sequence of decreasing nested events i.e. $A_{i+1} \subseteq A_i \forall i \geq 1$, then

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{m\to\infty} \mathbb{P}\left(A_m\right). \tag{5.7}$$

Properties 6 and 7 are said to be corollaries to Property 5.

- **Property 8:-** Suppose $\{A_i, i \geq 1\}$ are events, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{5.8}$$

This result is known as the *Union Bound*. This is bound is trivial if $\sum_{i=1}^{\infty} \mathbb{P}(A_i) \geq 1$ since the LHS of (5.8) is a probability of some event. This is a very widely used bound, and has several applications. For instance, the union bound is used in the probability of error analysis in Digital Communications for complicated modulation schemes.

Proof:- Define a new family of sets $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, ..., $B_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i$, ...

Claim 1:- $B_i \cap B_j = \emptyset$, $\forall i \neq j$.

Claim 2:- $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$.

Since $\{B_i, i \geq 1\}$ is a disjoint sequence of events, and using the above claims, we get

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

Also, since $B_i \subseteq A_i$ $\forall i \geq 1$, $\mathbb{P}(B_i) \leq \mathbb{P}(A_i)$ $\forall i \geq 1$ (using Property 2). Therefore, the finite sum of probabilities follow

$$\sum_{i=1}^{n} \mathbb{P}(B_i) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Eventually, in the limit, the following holds:

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

Finally we arrive at the result,

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

5.2 Exercises

1. a) Prove *Claim 1* and *Claim 2* stated in Property 5.

b) Prove Properties 6 and 7, which are corollaries of Property 5.

2. A standard card deck (52 cards) is distributed to two persons: 26 cards to each person. All partitions are equally likely. Find the probability that the first person receives all four aces.

3. Consider two events A and B such that $\mathbb{P}(A) > 1 - \delta$ and $\mathbb{P}(B) > 1 - \delta$, for some very small $\delta > 0$. Prove that $\mathbb{P}(A \cap B)$ is close to 1.

4. **[Grimmett]** Given events $A_1, A_2, ...A_n$, prove that,

$$\mathbb{P}(\cup_{1 \le r \le n} A_r) \le \min_{1 \le k \le n} \Big(\sum_{1 \le r \le n} \mathbb{P}(A_r) - \sum_{r : r \ne k} \mathbb{P}(A_r \cap A_k) \Big)$$

5. Consider a measurable space $(\Omega, \mathcal{F})$ with $\Omega = [0, 1]$. A measure $\mathbb{P}$ is defined on the non-empty subsets of Ω (in $\mathcal{F}$), which are all of the form (a, b), $(a, b]$, $[a, b)$ and $[a, b]$, as the length of the interval, i.e., $\mathbb{P}((a, b)) = \mathbb{P}((a, b]) = \mathbb{P}([a, b)) = \mathbb{P}([a, b]) = b - a$.

   a) Show that $\mathbb{P}$ is not just a measure, but its a probability measure.

   b) Let $A_n = [\frac{1}{n+1}, 1]$ and $B_n = [0, \frac{1}{n+1}]$, for $n \ge 1$. Compute $\mathbb{P}(\cup_{i \in \mathbb{N}} A_i)$, $\mathbb{P}(\cap_{i \in \mathbb{N}} A_i)$, $\mathbb{P}(\cup_{i \in \mathbb{N}} B_i)$ and $\mathbb{P}(\cap_{i \in \mathbb{N}} B_i)$.

   c) Compute $\mathbb{P}(\cap_{i \in \mathbb{N}} (B_i^c \cup A_i^c))$.

   d) Let $C_m = [0, \frac{1}{m}]$ such that $\mathbb{P}(C_m) = \mathbb{P}(A_n)$. Express m in terms of n.

   e) Evaluate $\mathbb{P}(\cap_{i \in \mathbb{N}} (C_i \cap A_i))$ and $\mathbb{P}(\cup_{i \in \mathbb{N}} (C_i \cap A_i))$.

6. **[Grimmett]** You are given that at least one of the events A_n, $1 \le n \le N$, is certain to occur. However, certainly no more than two occur. If $\mathbb{P}(A_n) = p$ and $\mathbb{P}(A_n \cap A_m) = q$, $m \ne n$, then show that $p \ge \frac{1}{N}$ and $q \le \frac{2}{N}$.

EE5110: Probability Foundations in Signal Processing **July-November 2015**

Lecture 6: Discrete Probability Spaces

Lecturer: Dr. Krishna Jagannathan *Scribe: Ravi Kolla*

6.1 Discrete Probability Spaces

In this lecture, we discuss discrete probability spaces. This corresponds to the case when the sample space Ω is countable. This is the most conceptually straightforward case, since it is possible to assign probabilities to *all* subsets of Ω.

Definition 6.1 *A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be a discrete probability space if the following conditions hold:*

(a) The sample space Ω is finite or countably infinite,

(b) The σ-algebra is the set of all subsets of Ω, i.e., $\mathcal{F} = 2^{\Omega}$, and

(c) The probability measure, $\mathbb{P}$, is defined for every subset of Ω. In particular, it can be defined in terms of the probabilities $\mathbb{P}(\{\omega\})$ of the singletons corresponding to each of the elementary outcomes ω, and satisfies for every $A \in \mathcal{F}$,

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

and

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

6.1.1 Examples of Discrete Probability Space

1. Let us consider a coin toss experiment with the probability of getting a head as p and the probability of getting a tail as $(1-p)$. Then, the sample space and the σ-algebra are

$$\Omega = \{H, T\} \equiv \{0, 1\}, \qquad \mathcal{F} = 2^{\Omega} = \{\Phi, \{H\}, \{T\}, \{\Omega\}\}.$$

respectively. The probability measure is

$$\mathbb{P}(\{H\}) \equiv \mathbb{P}(\{0\}) = p,$$
$$\mathbb{P}(\{T\}) \equiv \mathbb{P}(\{1\}) = 1 - p.$$

In this case, we say that $\mathbb{P}(.)$ is a Bernoulli measure on $\left(\{0,1\}, 2^{\{0,1\}}\right)$.

2. Let $\Omega = \mathbb{N}$, $\mathcal{F} = 2^{\mathbb{N}}$. Then, we can define the probability of a singleton as

$$\mathbb{P}(\{k\}) = a_k \geq 0, k \in \mathbb{N}$$

under the constraint that

$$\sum_{k \in \mathbb{N}} \mathbb{P}(\{k\}) = 1.$$

For example, $a_k = \frac{1}{2^k},\ \ k \in \mathbb{N}$ is a valid measure, since

$$\sum_{k\in\mathbb{N}} \frac{1}{2^k} = 1.$$

As another example, consider $a_k = (1-p)^{k-1} p,\ \ 0 < p < 1,\ \ k \in \mathbb{N}$. This is known as a geometric measure with parameter p. It is a valid probability measure since

$$\sum_{k\in\mathbb{N}} (1-p)^{k-1} p = 1.$$

3. Let $\Omega = \mathbb{N} \cup \{0\}$, $\mathcal{F} = 2^{\Omega}$. Let us define

$$\mathbb{P}(\{k\}) = \frac{e^{-\lambda}\lambda^k}{k!},\ \ \lambda > 0.$$

This probability measure is called a Poisson measure with parameter λ on $\left(\Omega, 2^{\Omega}\right)$. This is a valid probability measure, since

$$\sum_{k=0}^{\infty} \mathbb{P}(\{k\}) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \underbrace{\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}}_{e^{\lambda}} = 1.$$

4. Let $\Omega = \{0, 1, 2, \cdots, N\},\ \ N \in \mathbb{N}\ \ \mathcal{F} = 2^{\Omega}$. Let us define

$$\mathbb{P}(\{k\}) = \left(\begin{array}{c} N \\ k \end{array} \right) p^k (1-p)^{N-k},\ \ 0 < p < 1.$$

This probability measure is called a Binomial measure with parameters (N, p) on $\left(\Omega, 2^{\Omega}\right)$. This can be verified to be a valid probability measure as follows:

$$\sum_{k\in\Omega} \left(\begin{array}{c} N \\ k \end{array} \right) p^k (1-p)^{N-k} = (p + 1 - p)^N = 1$$

Note that in all the examples above, we have not explicitly specified an expression for $\mathbb{P}(A)$ for every $A \subset \Omega$. Since the sample space is countable, the probability of any subset of the sample space can be obtained as the sum of probabilities of the corresponding elementary outcomes. In other words, for discrete probability spaces, it suffices to specify the probabilities of singletons corresponding to each of the elementary outcomes.

6.2 Exercises

1. An urn contains a number of white balls and b number of black balls. Balls are drawn randomly from the urn without replacement. Find the probability that a white ball is drawn at the kth draw.

2. An urn contains white and black balls. When two balls are drawn without replacement, suppose the probability that both the balls are white is $\frac{1}{3}$.
(a) Find the smallest number of balls in the urn.
(b) How small can the total number of balls be if the number of black balls is even?

3. Consider the sample space $\Omega = \mathbb{N}$. Find the values of the constant C for which the following are probability measures:
(a) $f(x) = C2^{-x}$
(b) $f(x) = \frac{C2^{-x}}{x}$
(c) $f(x) = Cx^{-2}$
(d) $f(x) = \frac{C2^{x}}{x!}$

4. Recall the Poisson measure on $(\Omega, 2^{\Omega})$, where $\Omega = \mathbb{N} \cup \{0\}$. What is the probability assigned to the set of odd numbers? Prime numbers?

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 7: Borel Sets and Lebesgue Measure

Lecturer: Dr. Krishna Jagannathan *Scribes: Ravi Kolla, Aseem Sharma, Vishakh Hegde*

In this lecture, we discuss the case where the sample space is uncountable. This case is more involved than the case of a countable sample space, mainly because it is often not possible to assign probabilities to all subsets of Ω. Instead, we are forced to work with a smaller σ-algebra. We consider assigning a "uniform probability measure" on the unit interval.

7.1 Uncountable sample spaces

Consider the experiment of picking a real number at random from $\Omega = [0, 1]$, such that every number is "equally likely" to be picked. It is quite apparent that a simple strategy of assigning probabilities to singleton subsets of the sample space gets into difficulties quite quickly. Indeed,

(i) If we assign some positive probability to each elementary outcome, then the probability of an event with infinitely many elements, such as $A = \{1, \frac{1}{2}, \frac{1}{3}, \cdots\}$, would become unbounded.

(ii) If we assign zero probability to each elementary outcome, this alone would not be sufficient to determine the probability of a uncountable subset of Ω, such as $\left[\frac{1}{2}, \frac{2}{3}\right]$. This is because probability measures are not additive over uncountable disjoint unions (of singletons in this case).

Thus, we need a different approach to assign probabilities when the sample space is uncountable, such as $\Omega = [0, 1]$. In particular, we need to assign probabilities directly to specific subsets of Ω. Intuitively, we would like our 'uniform measure' μ on $[0, 1]$ to possess the following two properties.

(i) $\mu((a, b)) = \mu((a, b]) = \mu([a, b)) = \mu([a, b])$

(ii) Translational Invariance. That is, if $A \in [0, 1]$, then for any $x \in \Omega$, $\mu(A \oplus x) = \mu(A)$ where, the set $A \oplus x$ is defined as

$$A \oplus x = \{a + x | a \in A, a + x \leq 1\} \cup \{a + x - 1 | a \in A, a + x > 1\}$$

However, the following impossibility result asserts that there is no way to consistently define a uniform measure on all subsets of $[0, 1]$.

Theorem 7.1 (Impossibility Result) *There does not exist a definition of a measure $\mu(A)$ for all subsets of $[0, 1]$ satisfying (i) and (ii).*

Proof: Refer proposition 1.2.6 in [1].

Therefore, we must compromise, and consider a smaller σ-algebra that contains certain "nice" subsets of the sample space $[0, 1]$. These "nice" subsets are the intervals, and the resulting σ-algebra is called the Borel σ-algebra. Before defining Borel sets, we introduce the concept of generating σ-algebras from a given collection of subsets.

7.2 Generated σ-algebra and Borel sets

The σ-algebra generated by a collection of subsets of the sample space is the smallest σ-algebra that contains the collection. More formally, we have the following theorem.

Theorem 7.2 *Let $\mathcal{C}$ be an arbitrary collection of subsets of Ω, then there exists a smallest σ-algebra, denoted by $\sigma(\mathcal{C})$, that contains all elements of $\mathcal{C}$. That is, if $\mathcal{H}$ is any σ-algebra such that $\mathcal{C} \subseteq \mathcal{H}$, then $\sigma(\mathcal{C}) \subseteq \mathcal{H}$. $\sigma(\mathcal{C})$ is called the σ-algebra generated by $\mathcal{C}$.*

Proof: Let $\{\mathcal{F}_i,\ i \in \mathcal{I}\}$ denote the collection of all σ-algebras that contain $\mathcal{C}$. Clearly, the collection $\{\mathcal{F}_i,\ i \in \mathcal{I}\}$ is non-empty, since it contains at least the power set, 2^Ω. Consider the intersection $\bigcap_{i\in\mathcal{I}} \mathcal{F}_i$. Since the intersection of σ-algebras results in a σ-algebra (homework problem!) and the intersection contains $\mathcal{C}$, it follows that $\bigcap_{i\in\mathcal{I}} \mathcal{F}_i$ is a σ-algebra that contains $\mathcal{C}$. Finally, if $\mathcal{C} \subseteq \mathcal{H}$, then $\mathcal{H}$ is one of $\mathcal{F}_i$'s for some $i \in \mathcal{I}$. Hence $\bigcap_{i\in\mathcal{I}} \mathcal{F}_i$ is the smallest σ-algebra generated by $\mathcal{C}$. ■

Intuitively, we can think of $\mathcal{C}$ as being the collection of subsets of Ω which are of interest to us. Then, $\sigma(\mathcal{C})$ is the smallest σ-algebra containing all the 'interesting' subsets.

We are now ready to define Borel sets.

Definition 7.3

(a) Consider $\Omega = (0,1]$. Let $\mathcal{C}_0$ be the collection of all open intervals in $(0,1]$. Then $\sigma(\mathcal{C}_0)$, the σ - algebra generated by $\mathcal{C}_0$, is called the Borel σ - algebra. It is denoted by $\mathcal{B}((0,1])$.

(b) An element of $\mathcal{B}((0,1])$ is called a Borel-measurable set, or simply a Borel set.

Thus, every open interval in $(0,1]$ is a Borel set. We next prove that every singleton set in $(0,1]$ is a Borel set.

Lemma 7.4 *Every singleton set $\{b\},\ 0 < b \leq 1$, is a Borel set, i.e., $\{b\} \in \mathcal{B}((0,1])$.*

Proof: Consider the collection of sets set $\left\{\left(b - \frac{1}{n}, b + \frac{1}{n}\right), n \geq 1\right\}$. By the definition of Borel sets,

$$\left(b - \frac{1}{n}, b + \frac{1}{n}\right) \in \mathcal{B}((0,1]).$$

Using the properties of σ-algebra,

$$\begin{aligned}
&\left(b - \frac{1}{n}, b + \frac{1}{n}\right)^c \in \mathcal{B}((0,1]) \\
\Longrightarrow &\bigcup_{n=1}^{\infty} \left(b - \frac{1}{n}, b + \frac{1}{n}\right)^c \in \mathcal{B}((0,1]) \\
\Longrightarrow &\left(\bigcap_{n=1}^{\infty} \left(b - \frac{1}{n}, b + \frac{1}{n}\right)\right)^c \in \mathcal{B}((0,1]) \\
\Longrightarrow &\bigcap_{n=1}^{\infty} \left(b - \frac{1}{n}, b + \frac{1}{n}\right) \in \mathcal{B}((0,1]). \qquad (7.1)
\end{aligned}$$

Next, we claim that

$$\{b\} = \bigcap_{n=1}^{\infty} \left(b - \frac{1}{n}, b + \frac{1}{n} \right). \tag{7.2}$$

i.e., b is the only element in $\bigcap_{n=1}^{\infty} \left(b - \frac{1}{n}, b + \frac{1}{n}\right)$. We prove this by contradiction. Let h be an element in $\bigcap_{n=1}^{\infty} \left(b - \frac{1}{n}, b + \frac{1}{n}\right)$ other than b. For every such h, there exists a large enough n_0 such that $h \notin \left(b - \frac{1}{n_0}, b + \frac{1}{n_0}\right)$. This implies $h \notin \bigcap_{n=1}^{\infty} \left(b - \frac{1}{n}, b + \frac{1}{n}\right)$. Using (7.1) and (7.2), thus, proves that $\{b\} \in \mathcal{B}((0,1])$. ■

As an immediate consequence to this lemma, we see that every half open interval, $(a, b]$, is a Borel set. This follows from the fact that

$$(a, b] = (a, b) \cup \{b\},$$

and the fact that a countable union of Borel sets is a Borel set. For the same reason, every closed interval, $[a, b]$, is a Borel set.

Note: Arbitrary union of open sets is always an open set, but infinite intersections of open sets need not be open.

Further reading for the enthusiastic: (try Wikipedia for a start)

- Non-Borel sets
- Non-measurable sets (Vitali set)
- Banach-Tarski paradox (a bizzare phenomenon about cutting up the surface of a sphere. See `https://www.youtube.com/watch?v=Tk4ubu7BlSk`
- The cardinality of the Borel σ-algebra (on the unit interval) is the same as the cardinality of the reals. Thus, the Borel σ-algebra is a much 'smaller' collection than the power set $2^{[0,1]}$. See `https://math.dartmouth.edu/archive/m103f08/public_html/borel-sets-soln.pdf`

7.3 Caratheodory's Extension Theorem

In this section, we discuss a formal procedure to define a probability measure on a general measurable space $(\Omega, \mathcal{F})$. Specifying the probability measure for all the elements of $\mathcal{F}$ directly is difficult, so we start with a smaller collection $\mathcal{F}_0$ of 'interesting' subsets of Ω, which need not be a σ-algebra. We should take $\mathcal{F}_0$ to be rich enough, so that the σ-algebra it generates is same as $\mathcal{F}$. Then we define a function $\mathbb{P}_0 : \mathcal{F}_0 \to [0,1]$, such that it corresponds to the probabilities we would like to assign to the interesting subsets in $\mathcal{F}_0$. Under certain conditions, this function $\mathbb{P}_0$ can be extended to a legitimate probability measure on $(\Omega, \mathcal{F})$ by using the following fundamental theorem from measure theory.

Theorem 7.5 (Caratheodory's extension theorem) *Let $\mathcal{F}_0$ be an algebra of subsets of 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 $(\Omega, \mathcal{F})$. That is, there exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that $\mathbb{P}(A) = \mathbb{P}_0(A)$ for all $A \in \mathcal{F}_0$.

Proof: Refer Appendix A of [2]. ■

We use this theorem to define a uniform measure on $(0, 1]$, which is also called the Lebesgue measure.

7.4 The Lebesgue measure

Consider $\Omega = (0, 1]$. Let $\mathcal{F}_0$ consist of the empty set and all sets that are finite unions of the intervals of the form $(a, b]$. A typical element of this set is of the form

$$F = (a_1, b_1] \cup (a_2, b_2] \cup \ldots \cup (a_n, b_n]$$

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

Lemma 7.6

a) $\mathcal{F}_0$ *is an algebra*

b) $\mathcal{F}_0$ *is not a* σ*-algebra*

c) $\sigma(\mathcal{F}_0) = \mathcal{B}$

Proof:

a) By definition, $\Phi \in \mathcal{F}_0$. Also, $\Phi^C = (0, 1] \in \mathcal{F}_0$. The complement of $(a_1, b_1] \cup (a_2, b_2]$ is $(0, a_1] \cup (b_1, a_2] \cup (b_2, 1]$, which also belongs to $\mathcal{F}_0$. Furthermore, the union of finitely many sets each of which are finite unions of the intervals of the form $(a, b]$, is also a set which is the union of finite number of intervals, and thus belongs to $\mathcal{F}_0$.

b) To see this, note that $\left(0, \frac{n}{n+1}\right] \in \mathcal{F}_0$ for every n, but $\bigcup_{n=1}^{\infty} \left(0, \frac{n}{n+1}\right] = (0, 1) \notin \mathcal{F}_0$.

c) First, the null set is clearly a Borel set. Next, we have already seen that every interval of the form $(a, b]$ is a Borel set. Hence, every element of $\mathcal{F}_0$ (other than the null set), which is a finite union of such intervals, is also a Borel set. Therefore, $\mathcal{F}_0 \subseteq \mathcal{B}$. This implies $\sigma(\mathcal{F}_0) \subseteq \mathcal{B}$.

Next we show that $\mathcal{B} \subseteq \sigma(\mathcal{F}_0)$. For any interval of the form (a, b) in $\mathcal{C}_0$, we can write $(a, b) = \bigcup_{n=1}^{\infty} \left(\left(a, b - \frac{1}{n}\right] \cap \Omega\right)$. Since every interval of the form $\left(a, b - \frac{1}{n}\right] \in \mathcal{F}_0$, a countable number of unions of such intervals belongs to $\sigma(\mathcal{F}_0)$. Therefore, $(a, b) \in \sigma(\mathcal{F}_0)$ and consequently, $\mathcal{C}_0 \subseteq \sigma(\mathcal{F}_0)$. This gives $\sigma(\mathcal{C}_0) \subseteq \sigma(\mathcal{F}_0)$. Using the fact that $\sigma(\mathcal{C}_0) = \mathcal{B}$ proves the required result.

■

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

$$F = (a_1, b_1] \cup (a_2, b_2] \cup \ldots \cup (a_n, b_n],$$

we define a function $\mathbb{P}_0 : \mathcal{F}_0 \to [0, 1]$ such that

$$\mathbb{P}_0(\Phi) = 0 \text{ and } \mathbb{P}_0(F) = \sum_{i=1}^{n} (b_i - a_i).$$

Note that $\mathbb{P}_0(\Omega) = \mathbb{P}_0((0,1]) = 1$. Also, if $(a_1, b_1], (a_2, b_2], \ldots, (a_n, b_n]$ are disjoint sets, then

$$\begin{aligned}\mathbb{P}_0\left(\bigcup_{i=1}^{n}((a_i, b_i])\right) &= \sum_{i=1}^{n}\mathbb{P}_0((a_i, b_i]) \\ &= \sum_{i=1}^{n}(b_i - a_i)\end{aligned}$$

implying finite additivity of $\mathbb{P}_0$. It turns out that $\mathbb{P}_0$ is countably additive on $\mathcal{F}_0$ as well i.e., if $(a_1, b_1], (a_2, b_2], \ldots$ are disjoint sets such that $\bigcup_{i=1}^{\infty}((a_i, b_i]) \in \mathcal{F}_0$, then $\mathbb{P}_0\left(\bigcup_{i=1}^{\infty}((a_i, b_i])\right) = \sum_{i=1}^{\infty}\mathbb{P}_0((a_i, b_i]) = \sum_{i=1}^{\infty}(b_i - a_i)$. The proof is non-trivial and beyond the scope of this course (see [Williams] for a proof). Thus, in view of Theorem 7.5, there exists a unique probability measure $\mathbb{P}$ on $((0,1], \mathcal{B})$ which is the same as $\mathbb{P}_0$ on $\mathcal{F}_0$. This unique probability measure on $(0,1]$ is called the **Lebesgue** or **uniform** measure.

The Lebesgue measure formalizes the notion of length. This suggests that the Lebesgue measure of a singleton should be zero. This can be shown as follows. Let $b \in (0,1]$. Using (7.2), we write

$$\mathbb{P}(\{b\}) = \mathbb{P}\left(\bigcap_{n=1}^{\infty}\left(b - \frac{1}{n}, b\right] \cap \Omega\right)$$

Let $A_n = \left(b - \frac{1}{n}, b\right]$. For each n, the lebesgue measure of A_n is

$$\mathbb{P}(A_n) = \frac{1}{n} \tag{7.3}$$

Since A_n is a decreasing sequence of nested sets,

$$\begin{aligned}\mathbb{P}(\{b\}) =& \mathbb{P}\left(\bigcap_{n=1}^{\infty}A_n\right) \\ =& \lim_{n\to\infty}\mathbb{P}(A_n) \\ =& \lim_{n\to\infty}\frac{1}{n} \\ =& 0\end{aligned}$$

where the second equality follows from the continuity of probability measures.

Since any countable set is a countable union of singletons, the probability of a countable set is zero. For example, under the uniform measure on $(0,1]$, the probability of the set of rationals is zero, since the rational numbers in $(0,1]$ form a countable set.

For $\Omega = (0,1]$, the Lebesgue measure is also a probability measure. For other intervals (for example $\Omega = (0,2]$), it will only be a finite measure, which can be normalized as appropriate to obtain a uniform probability measure.

Definition 7.7 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An event A is said to occur almost surely (a.s) if $\mathbb{P}(A) = 1$.*

Caution: $\mathbb{P}(A) = 1$ does not mean $A = \Omega$.

Lebesgue Measure of the Cantor set: Consider the cantor set K. It is created by repeatedly removing the open middle thirds of a set of line segments. Consider its complement. It contains countable number of disjoint intervals. Hence we have:

$$\mathbb{P}(K^c) = \frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots = \frac{\frac{1}{3}}{1 - \frac{2}{3}} = 1.$$

Therefore $\mathbb{P}(K) = 0$. It is very interesting to note that though the Cantor set is equicardinal with $(0, 1]$, its Lebesgue measure is equal to 0 while the Lebesgue measure of $(0, 1]$ is equal to 1.
We now extend the definition of Lebesgue measure on $[0, 1]$ to the real line, $\mathbb{R}$. We first look at the definition of a Borel set on $\mathbb{R}$. This can be done in several ways, as shown below.

Definition 7.8 *Borel sets on $\mathbb{R}$:*

- *Let $\mathcal{C}$ be a collection of open intervals in $\mathbb{R}$. Then $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{C})$ is the Borel set on $\mathbb{R}$.*
- *Let $\mathcal{D}$ be a collection of semi-infinite intervals $\{(-\infty, x]; x \in \mathbb{R}\}$, then $\sigma(\mathcal{D}) = \mathcal{B}(\mathbb{R})$.*
- *$A \subseteq \mathbb{R}$ is said to be a Borel set on $\mathbb{R}$, if $A \cap (n, n+1]$ is a Borel set on $(n, n+1]$ $\forall n \in \mathbb{Z}$.*

Exercise: Verify that the three statements are equivalent definitions of Borel sets on $\mathbb{R}$.

Definition 7.9 *Lebesgue measure of $A \subseteq \mathbb{R}$:*

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

Theorem 7.10 *$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ is an infinite measure space.*

Proof: We need to prove following:

- $\lambda(\mathbb{R}) = \infty$
- $\lambda(\Phi) = 0$
- The countable additivity property

We see that

$$\mathbb{P}_n(\mathbb{R} \cap (n, n+1]) = 1, \forall n \in \mathbb{I}$$

Hence we have

$$\lambda(\mathbb{R}) = \sum_{n=-\infty}^{\infty} 1 = \infty$$

Now consider $\Phi \cap (n, n+1]$. This is a null set for all n. Hence we have,

$$\mathbb{P}_n(\Phi \cap (n, n+1]) = 0, \forall n \in \mathbb{I}$$

which implies,

$$\lambda(\Phi) = \sum_{n=-\infty}^{\infty} \mathbb{P}_n(\Phi \cap (n, n+1]) = 0$$

We now need to prove the countable additivity property. For this we consider $A_i \in \mathcal{B}(\mathbb{R})$ such that the sequence $A_1, A_2, \ldots, A_n, \ldots$ are arbitrary pairwise disjoint sets in $\mathcal{B}(\mathbb{R})$. Therefore we obtain,

$$\begin{aligned}\lambda(\bigcup_{i=1}^{\infty} A_i) &= \sum_{n=-\infty}^{\infty} \mathbb{P}_n(\bigcup_{i=1}^{\infty} A_i \cap (n, n+1]) \\ &= \sum_{n=-\infty}^{\infty} \sum_{i=1}^{\infty} \mathbb{P}_n(A_i \cap (n, n+1]) \\ &= \sum_{i=1}^{\infty} \sum_{n=-\infty}^{\infty} \mathbb{P}_n(A_i \cap (n, n+1])\end{aligned}$$

The second equality above comes from the fact that the probability measure has countable additivity property. The last equality above comes from the fact that the summations can be interchanged (from Fubini's theorem). We also have the following:

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

We now immediately see that

$$\lambda(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \lambda(A_i)$$

Hence proved. ■

7.5 Exercises

1. Let $\mathcal{F}$ be a σ-algebra corresponding to a sample space Ω. Let H be a subset of Ω that does not belong to $\mathcal{F}$. Consider the collection $\mathcal{G}$ of all sets of the form $(H \cap A) \cup (H^c \cap B)$, where A and $B \in \mathcal{F}$.
   (a) Show that $H \cap A \in \mathcal{G}$.
   (b) Show that $\mathcal{G}$ is a σ-algebra.
2. Show that $\mathcal{C} = \sigma(\mathcal{C})$ iff $\mathcal{C}$ is a σ-algebra.
3. Let $\mathcal{C}$ and $\mathcal{D}$ be two collections of subsets of Ω such that $\mathcal{C} \subseteq \mathcal{D}$. Prove that $\sigma(\mathcal{C}) \subseteq \sigma(\mathcal{D})$.
4. Prove that the following subsets of $(0, 1]$ are Borel-measurable.
   (a) any countable set
   (b) the set of irrational numbers
   (c) the Cantor set (Hint: rather than defining it in terms of ternary expansions, it's easier to use the equivalent definition of the Cantor set that involves sequentially removing the "middle-third" open intervals; see Wikipedia for example).
   (d) The set of numbers in $(0, 1]$ whose decimal expansion does not contain 7.
5. Let $\mathcal{B}$ denote the Borel σ-algebra as defined in class. Let $\mathcal{C}_c$ denote the set of all closed intervals contained in $(0, 1]$. Show that $\sigma(\mathcal{C}_c) = \mathcal{B}$. In other words, we could have very well defined the Borel σ-algebra as being generated by closed intervals, rather than open intervals.

6. Let $\Omega = [0, 1]$, and let $\mathcal{F}_3$ consist of all countable subsets of Ω, and all subsets of Ω having a countable complement. It can be shown that $\mathcal{F}_3$ is a σ-algebra (Refer Lecture 4, Exercises, 6(d)). Let us define $\mathbb{P}(A) = 0$ if A countable, and $\mathbb{P}(A) = 1$ if A has a countable complement. Is $(\Omega, \mathcal{F}_3, \mathbb{P})$ a legitimate probability space?

7. We have seen in 4(c) that the Cantor set is Borel-measurable. Show that the Cantor set has zero Lebesgue measure. Thus, although the Cantor set can be put into a bijection with $[0, 1]$, it has zero Lebesgue measure!

References

[1] Rosenthal, J. S. (2006). A first look at rigorous probability theory (Vol. 2). Singapore: World Scientific.

[2] Williams, D. (1991). Probability with martingales. Cambridge university press.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 8: The Infinite Coin Toss Model

Lecturer: Dr. Krishna Jagannathan *Scribe: Subrahmanya Swamy P*

In this lecture, we will discuss the random experiment where each trial consists of tossing a coin infinite times. We will describe the sample space, an appropriate σ-algebra, and a probability measure that intuitively corresponds to fair coin tosses. If we denote Heads/Tails with 0/1, the sample space of this experiment turns out to be $\Omega = \{0,1\}^\infty$, and each elementary outcome is some infinite binary string. As we have seen before, this is an uncountable sample space, so defining a useful σ-algebra on Ω takes some effort.

8.1 A σ-algebra on $\Omega = \{0,1\}^\infty$

Let $\mathcal{F}_n$ be the collection of subsets of Ω whose occurrences can be decided by looking at the result of the first n tosses. More formally, the elements of $\mathcal{F}_n$ can be described as follows: $A \in \mathcal{F}_n$ if and only if there exists some $A^{(n)} \subseteq \{0,1\}^n$ such that $A = \{\omega \in \Omega | (\omega_1, \omega_2, ..., \omega_n) \in A^{(n)}\}$.

Examples:

1 Let A_1 be the set of all elements of Ω such that there are exactly 2 heads during the first 4 coin tosses. Clearly, $A_1 \in \mathcal{F}_4$.

2 Let A_2 be the set of all elements of Ω such that the third toss is a Head. Then, $A_2 \in \mathcal{F}_3$.

Also note that the following relation holds:

$$\mathcal{F}_n \subseteq \mathcal{F}_{n+1} \ \ \forall n \in \mathbb{N}. \tag{8.1}$$

Although $\mathcal{F}_n$ is a σ-algebra, it has the drawback that it allows us to describe only those subsets which can be decided in n tosses. For example, the singleton set containing all Heads is not an element of $\mathcal{F}_n$ for any n.

In order to overcome this drawback, we define $\mathcal{F}_0 = \bigcup_{i \in \mathbb{N}} \mathcal{F}_i$. In words, $\mathcal{F}_0$ is the collection of all subsets of Ω that can be decided in *finitely many* coin tosses, since an element of $\mathcal{F}_0$ must be an element of $\mathcal{F}_i$ for some $i \in \mathbb{N}$.

Proposition 8.1 *We claim the following:*

(i) $\mathcal{F}_0$ *is an algebra.*

(ii) $\mathcal{F}_0$ *is not a* σ*-algebra.*

Proof:

(i) This is just definition chasing! (*Left as an exercise*).

(ii) Consider the following example: Let $E = \{\omega \in \Omega \mid \text{every odd toss results in Heads}\}$. Clearly, $E \notin \mathcal{F}_0$ since we cannot decide the occurrence of E in finitely many tosses. On the other hand, E can be expressed as a countable intersection of elements in $\mathcal{F}_0$:

$$E = \bigcap_{i=1}^{\infty} A_{2i-1},$$

where $A_i \in \mathcal{F}_0$ is the set of all binary strings with Heads in the ith toss.

■

Next, consider the smallest σ-algebra containing all the elements of $\mathcal{F}_0$, i.e., define

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

8.2 A Probability Measure on $(\Omega = \{0,1\}^{\infty}, \mathcal{F})$

Now, we shall define a uniform probability measure on $\mathcal{F}$ that corresponds to a 'fair' coin toss model. We shall first define a finitely additive function $\mathbb{P}_0$ on $\mathcal{F}_0$ that also satisfies $\mathbb{P}_0(\Omega) = 1$. Then, we shall subsequently extend $\mathbb{P}_0$ to a probability measure $\mathbb{P}$ on $\mathcal{F}$.

If $A \in \mathcal{F}_0$, then by the definition of $\mathcal{F}_0$, $\exists n$ such that $A \in \mathcal{F}_n$. By the definition of $\mathcal{F}_n$, we know that for every $A \in \mathcal{F}_n$, there exists a corresponding $A^{(n)} \subseteq \{0,1\}^n$. We will use this $A^{(n)}$ in the definition of $\mathbb{P}_0$. We define $\mathbb{P}_0 : \mathcal{F}_0 \to [0,1]$ as follows:

$$\mathbb{P}_0(A) = \frac{|A^{(n)}|}{2^n}.$$

Having defined $\mathbb{P}_0$ this way, we need to verify that this definition is consistent. In particular, we note that if $A \in \mathcal{F}_n$, $A \in \mathcal{F}_{n+1}$, which is trivially true because $\mathcal{F}_n's$ are nested increasing. We therefore need to prove that when we apply the definition $\mathbb{P}_0(A)$ for different choices of n, we obtain the same value. We leave it to the reader to supply a formal proof for the consistency of $\mathbb{P}_0$. However, we illustrate this consistency using the examples provided in Section 8.1.

(i) The occurrence of the event A_2 can be decided in the first 3 tosses. So, $A_2 \in \mathcal{F}_3$. The elements in $A^{(3)} \subseteq \{0,1\}^3$ corresponding to the event A_2 are $\{(0,0,0),(0,1,0),(1,0,0),(1,1,0)\}$. So $|A^{(3)}| = 4$. So, $\mathbb{P}_0(A_2) = \frac{4}{2^3} = \frac{1}{2}$.
The event A_2 can also be looked as an event in $\mathcal{F}_4$ since, $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$. The elements in the corresponding $A^{(4)}$ will be $\{(0,0,0,0),(0,1,0,0),(1,0,0,0),(1,1,0,0),(0,0,0,1),(0,1,0,1),(1,0,0,1),(1,1,0,1)\}$. So $|A^{(4)}| = 8$. So, $\mathbb{P}_0(A_2) = \frac{8}{2^4} = \frac{1}{2}$.

(ii) A_1 can be decided by looking at the outcome of the first four tosses. So, $A_1 \in \mathcal{F}_4$. It is easy to see that the number of elements in $A^{(4)} \subseteq \{0,1\}^4$ corresponding to the event A_1 that has exactly two heads is $\binom{4}{2}$. Hence, $\mathbb{P}_0(A_1) = \frac{\binom{4}{2}}{2^4}$. Next, can you compute $\mathbb{P}_0(A_1)$ by considering A_1 as an element of, say $\mathcal{F}_5$?

From the above examples, we can observe that

(a) The definition of $\mathbb{P}_0$ is consistent over different choices on n namely $n = 3$ and $n = 4$ for a given set A_2.

(b) The definition of $\mathbb{P}_0$ is also consistent with the intuition of a fair coin toss model with probability of heads being $\frac{1}{2}$.

It can be easily verified that $\mathbb{P}_0(\Omega) = 1$ and $\mathbb{P}_0$ is finitely additive. It also turns out that $\mathbb{P}_0$ is countably additive on $\mathcal{F}_0$ (the proof of this fact is non-trivial and is omitted here). This allows us to invoke the Caratheodory extension theorem and extend $\mathbb{P}_0$ to $\mathbb{P}$, a legitimate probability measure on $(\Omega, \mathcal{F})$ which agrees with $\mathbb{P}_0$ on $\mathcal{F}_0$. In other words, there exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$.

As an example, let us consider the event E that is defined above (i.e. the set of strings in which all the odd tosses are heads). As $E \notin \mathcal{F}_0$, $\mathbb{P}_0$ is not defined for the event E. However, it is clear that $E \in \mathcal{F}$, so that $\mathbb{P}$ is defined for E. Let us calculate the probability of the event E. Recall that

$$E = \bigcap_{i=1}^{\infty} A_{2i-1}, \text{ where } A_i \quad = \quad \{\underline{\omega} \in \Omega \mid \omega_i = 0\}.$$

Let us define the event $E_m = \bigcap_{i=1}^{m} A_{2i-1}$. In other words, E_m is set of outcomes in which the first $2m$ tosses have the property of all odd tosses being heads. We can easily verify that $\mathbb{P}(E_m) = \mathbb{P}_0(E_m) = \frac{1}{2^m}$. Note that $\{E_m, m \geq 1\}$ is a sequence of nested decreasing events i.e., $E_m \supseteq E_{m+1}$, $\forall m \geq 1$. It can be easily verified that E can be expressed in terms of these decreasing nested events as $E = \bigcap_{m=1}^{\infty} E_m$.

Thus,

$$\begin{aligned}
\mathbb{P}(E) &= \mathbb{P}\left(\bigcap_{m=1}^{\infty} E_m\right) \\
&\overset{(a)}{=} \lim_{m\to\infty} \mathbb{P}\left(E_m\right) \\
&= \lim_{m\to\infty} \frac{1}{2^m} \\
&= 0,
\end{aligned}$$

where the equality (a) follows from the continuity of probability measures.

8.3 Exercises

1. Show that $\mathcal{F}_n$ (defined in equation 8.1) is a σ-algebra $\forall n \in \mathbb{N}$.

2. Recall the infinite coin toss model with $\Omega = \{0, 1\}^{\infty}$; where '0' denotes heads and '1' denotes tails. Define $\mathcal{F}_n$ as the collection of subsets of Ω whose occurrence can be decided by looking at the results of the first n tosses. *Exercise:*

(a) Show that $\mathcal{F}_n$ is a σ-algebra.

It turns out that the σ-algebra $\mathcal{F}_n$ for any fixed n is too small; after all, it can only serve to model the first n tosses. Let us define

$$\mathcal{F}_0 = \bigcup_{i=1}^{\infty} \mathcal{F}_n. \tag{8.2}$$

(b) Give a verbal description of the collection $\mathcal{F}_0$.

(c) Show that $\mathcal{F}_0$ is an algebra on Ω.

(d) Consider the subset $\underline{A} \subset \Omega$ consisting of sequences in which Tails occurs infinitely many times. Does $\underline{A} \in \mathcal{F}_0$?

(e) Is A^c countable?

(f) Let B be the set of all infinite sequences for which $\omega_n = 0$ for every odd n; i.e., every odd numbered toss is Heads. Show that B can be written as a countable intersection of subsets in $\mathcal{F}_0$, but $B \notin \mathcal{F}_0$. Therefore $\mathcal{F}_0$ is not a σ-algebra.

Define $\mathcal{F} = \sigma(\mathcal{F}_0)$, the σ-algebra generated by $\mathcal{F}_0$.

(g) Show that every singleton $\{\omega\}$ is $\mathcal{F}$ measurable. Show that the uniform measure on $(\Omega, \mathcal{F})$ defined in class assigns zero probability measure to singletons.

(h) Let A_i be the set of all outcomes such that the i^{th} toss is Tails. Note that $A_i \in \mathcal{F}_0$. Show that $\underline{A}$ in part (e) can be written as

$$\underline{A} = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i \tag{8.3}$$

Hence show that $\underline{A}$ is $\mathcal{F}$ measurable.What is $\mathbb{P}\{\underline{A}\}$ under the uniform measure?

(i) Let $T \subset \Omega$ be the set of all coin toss sequences in which the fraction of Tails is exactly 1/2: More precisely,

$$T = \{\omega \in \Omega \ : \ \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \omega_i}{n} = \frac{1}{2}\} \tag{8.4}$$

The set T is called the strong-law truth set, for reasons that will become clear later. Does $T \in \mathcal{F}_0$?

(j) Show that T can be expressed as

$$T = \bigcap_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{\omega \in \Omega \mid \frac{\sum_{i=1}^{n} \omega_i}{n} - \frac{1}{2} \mid < \frac{1}{k}\} \tag{8.5}$$

Argue that the subset inside the nested union and intersection above belongs to $\mathcal{F}_0$: Hence show that T is $\mathcal{F}$-measurable. Hint: Don't get intimidated by the multiple unions and intersections! Write-out the limit in the definition of T as the set of all $\omega \in \Omega$ such that for all $k \geq 1$; there exists an m for which for all $n > m$; we have

$$\mid \frac{\sum_{i=1}^{n} \omega_i}{n} - \frac{1}{2} \mid < \frac{1}{k} \tag{8.6}$$

EE5110: Probability Foundations **July-November 2015**

Lecture 9: Conditional Probability and Independence

Lecturer: Dr. Krishna Jagannathan *Scribe: Vishakh Hegde*

9.1 Conditional Probability

Definition 9.1 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. Then the* conditional probability of A given B *is defined as,*

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

Caution: We cannot condition on sets of zero probability measure. For example, if $\Omega = [0,1]$ endowed with the Borel σ-algebra and a uniform probability measure, we cannot condition on the set of rationals.

Theorem 9.2 *Let $B \in \mathcal{F}$ and $\mathbb{P}(B) > 0$. Then, $\mathbb{P}(\cdot\,|B) : \mathcal{F} \longmapsto [0,1]$ is a probability measure on $(\Omega, \mathcal{F})$.*

Proof: We need to show that the three properties of probability measure holds true, namely:

- $\mathbb{P}(\Omega|B) = 1$.
- $\mathbb{P}(\phi|B) = 0$.
- Countable additivity property.

We have,

$$\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$$

$$\mathbb{P}(\phi|B) = \frac{\mathbb{P}(\phi \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\phi)}{\mathbb{P}(B)} = 0.$$

We are now left with proving countable additivity property. Let $A_1, A_2, \ldots$ be disjoint. We need to show that,

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

Consider,

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

Since A_i are disjoint, $A_i \cap B$ are also disjoint. Therefore we can write the following:

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

■

9.1.1 Properties of Conditional Probability

1. *The Law of Total Probability:* Let $A \in \mathcal{F}$ and let $\{B_i,\ i = 1, 2, \ldots\}$ be events that partition Ω (by partition we mean $\bigcup_{i \in \mathbb{N}} B_i = \Omega$ and $B_i \cap B_j = \phi,\ \forall i \neq j$), with $\mathbb{P}(B_i) > 0,\ \forall i$. Then,

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

Proof: We know that $\{B_i,\ i = 1, 2, \ldots\}$ partitions Ω. Hence $\{A \cap B_i,\ i = 1, 2, \ldots\}$ partitions A. Therefore, by the countable additivity property, we have

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

and $\mathbb{P}(A \cap B_i) = \mathbb{P}(A|B_i)\,\mathbb{P}(B_i),\ \forall i$. Therefore,

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

■

Note: In particular, if B is such that $0 < \mathbb{P}(B) < 1$, then,

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

2. *Bayes' Rule:* Let $A \in \mathcal{F}$, with $\mathbb{P}(A) > 0$ and let $\{B_i,\ i = 1, 2, \ldots\}$ be a partition of Ω such that $\mathbb{P}(B_i) > 0\ \forall i$. Then, we have,

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

Proof:

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

■

3. For any sequence of events $\{A_i\}$, we have the following relation:

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \mathbb{P}(A_1) \prod_{i=2}^{\infty} \mathbb{P}(A_i|A_1 \cap A_2 \cap \ldots \cap A_{i-1}).$$

as long as all the conditional probabilities are well defined.
Proof: We know that the following holds for finite set of events:

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbb{P}(A_1) \prod_{i=2}^{n} \mathbb{P}(A_i|A_1 \cap A_2 \cap \ldots \cap A_{i-1}).$$

Now taking limits, we have:

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \lim_{n\to\infty} \mathbb{P}(A_1) \prod_{i=2}^{n} \mathbb{P}(A_i|A_1 \cap A_2 \cap \ldots \cap A_{i-1}).$$

Now using continuity of probability, we get the required relation,

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \mathbb{P}(A_1) \prod_{i=2}^{\infty} \mathbb{P}(A_i|A_1 \cap A_2 \cap \ldots \cap A_{i-1}).$$

■

9.2 Independence

Definition 9.3 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two events A and B are said to be* independent *(under the probability measure $\mathbb{P}$) if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.*

Note: If $\mathbb{P}(B) > 0$ and, A and B are independent, then we have,

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

Example: Can disjoint sets be independent at all? Let $A, B \in \mathcal{F}$ be two disjoint sets. Therefore, we have $\mathbb{P}(A \cap B) = \mathbb{P}(\phi)$. This means that $\mathbb{P}(A \cap B) = 0$. For independence, we need to have $\mathbb{P}(A)\,\mathbb{P}(B) = \mathbb{P}(A \cap B) = 0$. This can happen when $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$. Therefore, in general, two disjoint events are independent if and only if at least one of them has zero probability.

Definition 9.4 *$A_1, A_2, \ldots, A_n$ are independent if for all non-empty $I_0 \subseteq \{1, 2, \ldots, n\}$, we have,*

$$\mathbb{P}\left(\bigcap_{i \in I_0} A_i\right) = \prod_{i \in I_0} \mathbb{P}(A_i).$$

Next, we define independence of an arbitrary collection of events.

Definition 9.5 *$\{A_i, i \in I\}$ are said to be independent if for every non-empty finite subset I_0 of I, we have*

$$\mathbb{P}\left(\bigcap_{i \in I_0} A_i\right) = \prod_{i \in I_0} \mathbb{P}(A_i).$$

9.2.1 Independence of σ-algebras

Definition 9.6 *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two sub-σ-algebras of $\mathcal{F}$. We say that $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent σ-algebras if for all $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$, A_1 and A_2 are independent events.*

Example: A simple example we can construct is the following: Let $A, B \in \mathcal{F}$, then $\mathcal{F}_1 = \{\phi, \Omega, A, A^c\}$ and $\mathcal{F}_2 = \{\phi, \Omega, B, B^c\}$ are independent iff A and B are independent.

We now define independence on a collection of sub-σ algebras.

Definition 9.7 *Let $\{\mathcal{F}_i, i \in I\}$ (where I is an index set) be a collection of sub σ algebras of $\mathcal{F}$. Then, $\{\mathcal{F}_i, i \in I\}$ are said to be independent if for every choice of $A_i \in \mathcal{F}_i$, we have $\{A_i, i \in I\}$ are independent.*

Example (from [Lecture 2, MIT OCW]): Consider the infinite coin toss model discussed previously.

- Let A_i be the event that the i^{th} coin toss resulted in heads (say). If $i \neq j$, the events A_i and A_j are independent.
- The following infinite family of events are independent: $\{A_i | i \in \mathbb{N}\}$. This example captures the intuitive idea of independent coin tosses.

- 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 toss at odd times (respectively, at even times) n. More formally, let H_i be the event that the i^{th} toss resulted in heads. Let $\mathcal{C} = \{H_i \mid i \text{ is odd}\}$ and let $\mathcal{F}_1 = \sigma(\mathcal{C})$, so that $\mathcal{F}_1$ is the smallest σ-algebra 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-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ turn out to be independent. Intuitively, this implies that any event whose occurrence is determined completely by the outcomes of the tosses at odd times, is independent of any event whose occurrence is determined completely by the outcomes of the tosses at even times.
- Let $\mathcal{F}_n$ be the collection of all events whose occurrence can be decided by looking at the coin tosses $2n$ and $2n+1$. We know that $\mathcal{F}_n$ is a σ-algebra with finitely many events $\forall n \in \mathbb{N}$. It turns out that $\{\mathcal{F}_n, n \in \mathbb{N}\}$ are independent.

9.3 Exercises

1. (a) Let C, $C \in \mathcal{F}$, where $\mathcal{F}$ is a sigma algebra on Ω. Show that $\mathcal{F}_1 = \{\phi, \Omega, C, C^c\}$ and $\mathcal{F}_2 = \{\phi, \Omega, D, D^c\}$ are independent iff C and D are independent.
   (b) Let $\Omega = \{1, 2, 3, ...p\}$ where p is a prime, $\mathcal{F}$ be the collection of all subsets of Ω, and $\mathbb{P}(A) = \frac{|A|}{p}$ (where $|A|$ denotes cardinality of A) for all $A \in \mathcal{F}$. Show that, if A and B are independent events, then at least one of A and B is either ϕ or Ω.
2. In a box, there are four red balls, six red cubes, six blue balls and an unknown number of blue cubes. When an object from the box is selected at random, the shape and colour of the object are independent. Determine the number of blue cubes.
3. A man is known to speak the truth 3 out of 4 times. He throws a die and reports that it is a six. Find the probability that it is actually a six.
4. *[Exercise: Q29, Bertsekas & Tsitsiklis]* Let A and B be events such that $P(A|B) > P(A)$. Show that $P(B|A) > P(B)$ and $P(A|B^c) < P(A)$.
5. *[MIT OCW Assignment problem]* A coin is tossed independently n times. The probability of heads at each toss is p. At each time k ($k = 2, 3, ..., n$) we get a reward at time $k+1$ if k^{th} toss was a head and the previous toss was a tail. Let A_k be the event that a reward is obtained at time k.
   a) Are events A_k and A_{k+1} independent?
   b) Are events A_k and A_{k+2} independent?
6. *[Assignment problem, University of Cambridge]* A drawer contains two coins. One is an unbiased coin, which when tossed, is equally likely to turn up heads or tails. The other is a biased coin, which will turn up heads with probability p and tails with probability $1-p$. One coin is selected (uniformly) at random from the drawer. Two experiments are performed:
   a) The selected coin is tossed n times. Given that the coin turns up heads k times and tails $n-k$ times, what is the probability that the coin is biased?
   b) The selected coin is toss repeatedly until it turns up heads k times. Given that the coin is tossed n times in total, what is the probability that the coin is biased?
7. *[MIT OCW Assignment problem]* Fred is giving out samples of dog food. He makes calls door to door, but he leaves a sample (one can) only on those calls for which the door is answered and a dog is in residence. On any call the probability of the door being answered is 3/4, and the probability that any household has a dog is 2/3. Assume that the events "door answered" and "a dog lives here" are independent and also that the outcomes of all calls are independent.

a) Determine the probability that Fred gives away his first sample on his third call.

b) Given that he has given away exactly four samples on his first eight calls, determine the conditional probability that Fred will give away his fifth sample on his eleventh call.

c) Determine the probability that he gives away his second sample on his fifth call.

d) Given that he did not give away his second sample on his second call, determine the conditional probability that he will leave his second sample on his fifth call.

e) We will say that Fred needs a new supply immediately after the call on which he gives away his last can. If he starts out with two cans, determine the probability that he completes at least five calls before he needs a new supply.

8. *[MIT OCW Assignment problem]* Let A, B, A_1, A_2, ... be events. Suppose that for each k, we have $A_k \subseteq A_{k+1}$, and that A_k is independent of B, $\forall k \geq 1$. If $A = \cup_{k \in \mathbb{N}} A_k$, then show that B is independent of A.

9. *[Assignment problem University of Cambridge]* Consider pairwise disjoint events B_1, B_2, B_3 and C, with $P(B_1) = P(B_2) = P(B_3) = p$ and $P(C) = q$, where $3p + q \leq 1$. Suppose $p = -q + \sqrt{q}$, then prove that the events $B_1 \cup C$, $B_1 \cup C$ and $B_1 \cup C$ are pairwise independent. Also, prove or disprove that there exist $p > 0$ and $q > 0$ such that these three events are independent.

References

[1] MIT OCW - 6.436J / 15.085J Fundamentals of Probability, Fall 2008, Lecture 2.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 10: The Borel-Cantelli Lemmas

Lecturer: Dr. Krishna Jagannathan *Scribe: Aseem Sharma*

The Borel-Cantelli lemmas are a set of results that establish if certain events occur infinitely often or only finitely often. We present here the two most well-known versions of the Borel-Cantelli lemmas.

Lemma 10.1 (First Borel-Cantelli lemma) *Let $\{A_n\}$ be a sequence of events such that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. Then, almost surely, only finitely many A_n's will occur.*

Lemma 10.2 (Second Borel-Cantelli lemma) *Let $\{A_n\}$ be a sequence of independent events such that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. Then, almost surely, infinitely many A_n's will occur.*

It should be noted that only the second lemma stipulates independence. The event "A_n occurs infinitely often (A_n *i.o.*)" is the set of all $\omega \in \Omega$ that belong to infinitely many A_n's. It is defined as

$$\{A_n \text{ } i.o.\} \triangleq \bigcap_{n=1}^{\infty} \underbrace{\bigcup_{m=n}^{\infty} A_m}_{B_n}. \tag{10.1}$$

Here, B_n is the event that atleast one of $A_n, A_{n+1}, A_{n+2}, \ldots$ occur. Hence, $\{A_n \text{ } i.o.\}$ is the event that for every $n \in \mathbb{N}$, there exists atleast one $m \in \{n, n+1, \ldots, \infty\}$ such that A_m occurs. Taking complement of both sides in (10.1), we get the expression for the event that A_n occurs finitely often (A_n *f.o.*)

$$\{A_n \text{ } f.o.\} = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} {A_m}^c.$$

In order to prove the Borel-Cantelli lemmas, we require the following lemma.

Lemma 10.3 *If $\sum_{i=1}^{\infty} p_i = \infty$, then $\lim_{n\to\infty} \prod_{i=1}^{n} (1-p_i) = 0$.*

Proof: Since $\ln(1-p_i) \leq -p_i$,

$$\begin{aligned}\prod_{i=1}^{n} (1-p_i) &= \prod_{i=1}^{n} e^{\ln(1-p_i)} \\ &\leq \prod_{i=1}^{n} e^{-p_i} \\ &= e^{-\sum_{i=1}^{n} p_i}.\end{aligned}$$

Taking limit on both the sides gives

$$\begin{aligned}\lim_{n\to\infty} \prod_{i=1}^{n} (1-p_i) &\leq \lim_{n\to\infty} e^{-\sum_{i=1}^{n} p_i} \\ &= 0.\end{aligned}$$

■

We now proceed towards proving the Borel-Cantelli lemmas.

Proof:

1. First, note that the assumption $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ implies $\lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbb{P}(A_m) = 0$. Next, since $B_{n+1} \subset B_n$, we can use continuity of probability to write

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m\right) &= \mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right) \\
&= \lim_{n\to\infty} \mathbb{P}(B_n) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m=n}^{\infty} A_m\right) \\
&\leq \lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbb{P}(A_m) \\
&= 0.
\end{aligned}$$

   We have used the union bound in writing the '$\leq$' above. Since $\mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} A_i\right) \geq 0$, we conclude that $\mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} A_i\right) = 0$. This implies that, A_n occurs finitely often with probability 1.

2. The event that A_n occurs finitely often (A_n *f.o.*) is given by

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

   Now,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{n=1}^{\infty}\bigcap_{i=n}^{\infty} A_i^c\right) &\leq \sum_{n=1}^{\infty} \mathbb{P}\left(\bigcap_{i=n}^{\infty} A_i^c\right) \quad \text{(Using union bound)} \\
&= \sum_{n=1}^{\infty} \lim_{m\to\infty} \mathbb{P}\left(\bigcap_{i=n}^{m} A_i^c\right) \quad \text{(By continuity of probability)} \\
&= \sum_{n=1}^{\infty} \prod_{i=n}^{m} \mathbb{P}(A_i^c) \quad \text{(By independence)} \\
&= 0 \quad \text{(By lemma 10.3)} \qquad (10.2)
\end{aligned}$$

   Since $\mathbb{P}\left(\bigcup_{n=1}^{\infty}\bigcap_{i=n}^{\infty} A_i^c\right) \geq 0$, we conclude that $\mathbb{P}\left(\bigcup_{n=1}^{\infty}\bigcap_{i=n}^{\infty} A_i^c\right) = 0$. This implies that, A_n occurs infinitely often with probability 1.

■

We now illustrate the usefulness of the Borel-Cantelli lemmas using an example. Consider an experiment in which a coin is tossed independently many times. Let $\mathbb{P}(H_n)$ be the probability of obtaining head at the n^{th} toss (and similarly for T_n).

1. Suppose $\mathbb{P}(H_n) = \frac{1}{n},\ n \geq 1$. Then $\sum_{n=1}^{\infty} \mathbb{P}(H_n) = \infty$. By the second Borel-Cantelli lemma, it follows that almost surely, infinitely many heads will occur. This might appear surprising at first sight, since as n becomes large, the probability of getting heads becomes vanishingly small. However, the decay rate $1/n$ is not 'fast enough.' In particular, for any n we choose (no matter how large), there occurs a head beyond n almost surely!

2. Suppose now that $\mathbb{P}(H_n) = \frac{1}{n^2}$. Then $\sum_{n=1}^{\infty} \mathbb{P}(H_n) < \infty$, and hence by the first Borel-Cantelli lemma, almost surely, only finitely many heads will occur. In this case, the occurence of heads is decreasing fast enough that after a finite n, there will almost surely be no heads. Note that independence is not required in this case.

Exercises

1. Suppose that a monkey sits in front of a computer and starts hammering keys randomly on the keyboard. Show that the famous Shakespeare monologue starting All the worlds a stage will eventually appear (with probability 1) in its entirety on the screen, although our monkey is not particularly known for its good taste in literature. You can make reason- able additional assumptions to form a probability model; for example, you can assume that the monkey picks characters uniformly at random on the keyboard, and that the successive key strokes are independent.

2. *[MIT OCW problem set]* Let $A_n, n \geq 1$ be a sequence of events such that $\mathbb{P}(A_n) \to 0$ as $n \to \infty$, and

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n^c \cap A_{n+1}) < \infty$$

Show that almost surely, only finitely many of the A_ns will occur.

3. **Online dating**: On a certain day, Alice decides that she will start looking for a potential life partner on an online dating portal. She decides that everyday, she will pick a guy uniformly at random from among the male members of the dating portal, and go out on a date with him. What Alice does not know, is that her neighbor Bob is interested in dating her. Being of a shy disposition, Bob decides that he will not ask Alice out himself. Instead, he decides that he will go out on a date with Alice only on the days that Alice happens to pick him from the dating portal, of which he is already a member. For the first two parts, assume that 50 new male members and 40 new female members join the dating portal everyday.

(a) What is the probability that Alice and Bob would have a date on the nth day? Do you think Bob and Alice would eventually stop meeting? Justify your answer, clearly stating any additional assumptions.

(b) Now suppose that Bob also picks a girl uniformly at random everyday, from among the female members of the portal, and that Alice behaves exactly as before. Assume also that Bob and Alice will meet on a given day if and only if they both happen to pick each other. In this case, do you think Bob and Alice would eventually stop meeting?

(c) For this part, suppose that Alice and Bob behave as in part (a), i.e.,Alice picks a guy uniformly at random, but Bob is only interested in dating Alice. However, the number of male members in the portal increases by 1 percent everyday. Do you think Bob and Alice would eventually stop meeting?

4. Let $\{S_n : n \geq 0\}$ be a simple random walk which moves to the right with probability p at each step, and suppose that $S_0 = 0$. Write $X_n = S_n - S_{n-1}$.

(a) Show that $\{S_n = 0 \ i.o\}$ is not a tail event of the sequence $\{X_n\}$.

(b) Show that $\mathbb{P}(S_n = 0 \ i.o) = 0$ if $p \neq \frac{1}{2}$

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 11: Random Variables

Lecturer: Dr. Krishna Jagannathan *Scribe: Sudharsan, Gopal, Arjun B, Debayani*

The study of random variables is motivated by the fact that in many scenarios, one might not be interested in the precise elementary outcome of a random experiment, but rather in some numerical function of the outcome. For example, in an experiment involving ten coin tosses, the experimenter may only want to know the total number of heads, rather than the precise sequence of heads and tails.

The term random variable is a misnomer, because a random variable is neither random, nor is it a variable. A random variable X is a function from the sample space Ω to real field $\mathbb{R}$. The term 'random' actually signifies the underlying randomness in picking an element ω from the sample space Ω. Once the elementary outcome ω is fixed, the random variable takes a fixed real value, $X(\omega)$. It is important to remember that the probability measure is associated with subsets (events), whereas a random variable is associated with each elementary outcome ω.

Just as not all subsets of the sample space are not necessarily considered events, not all functions from Ω to $\mathbb{R}$ are considered random variables. In particular, a random variable is an *$\mathcal{F}$-measurable* function, as we define below.

Definition 11.1 *Measurable function:*
Let $(\Omega, \mathcal{F})$ be a measurable space. A function $f : \Omega \to \mathbb{R}$ is said to be an $\mathcal{F}$-measurable function if the pre-image of every Borel set is an $\mathcal{F}$-measurable subset of Ω.

In the above definition, the pre-image of a Borel set B under the function f is given by

$$f^{-1}(B) \triangleq \{\omega \in \Omega \mid f(\omega) \in B\}. \tag{11.1}$$

Thus, according to the above definition, $f : \Omega \to \mathbb{R}$ is an $\mathcal{F}$-measurable function if $f^{-1}(B)$ is an $\mathcal{F}$-measurable subset of Ω for *every* Borel set B.

Definition 11.2 *Random Variable:*
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A random variable X is an $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}$.

In other words, for every Borel set B, its pre-image under a random variable X is an *event*. In Figure 11.1, X is a random variable that maps every element ω in the sample space Ω to the real line $\mathbb{R}$. B is a Borel set, i.e., $B \in \mathcal{B}(\mathbb{R})$. The inverse image of B is an event $E \in \mathcal{F}$.

Since the set $\{\omega \in \Omega | X(\omega) \in B\}$ is an event for every Borel set B, it has an associated probability measure. This brings us to the concept of the *probability law* of the random variable X.

Definition 11.3 *Probability law of a random variable X:*
The probability law $\mathbb{P}_X$ of a random variable X is a function $\mathbb{P}_X : \mathcal{B}(\mathbb{R}) \to [0, 1]$, which is defined as $\mathbb{P}_X(B) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \in B\})$.

Thus, the probability law can be seen as the composition of $\mathbb{P}(\cdot)$ with the inverse image $X^{-1}(\cdot)$, i.e., $\mathbb{P}_X(\cdot) = \mathbb{P} \circ X^{-1}(\cdot)$. Indeed, the probability law of a random variable completely specifies the statistical properties of

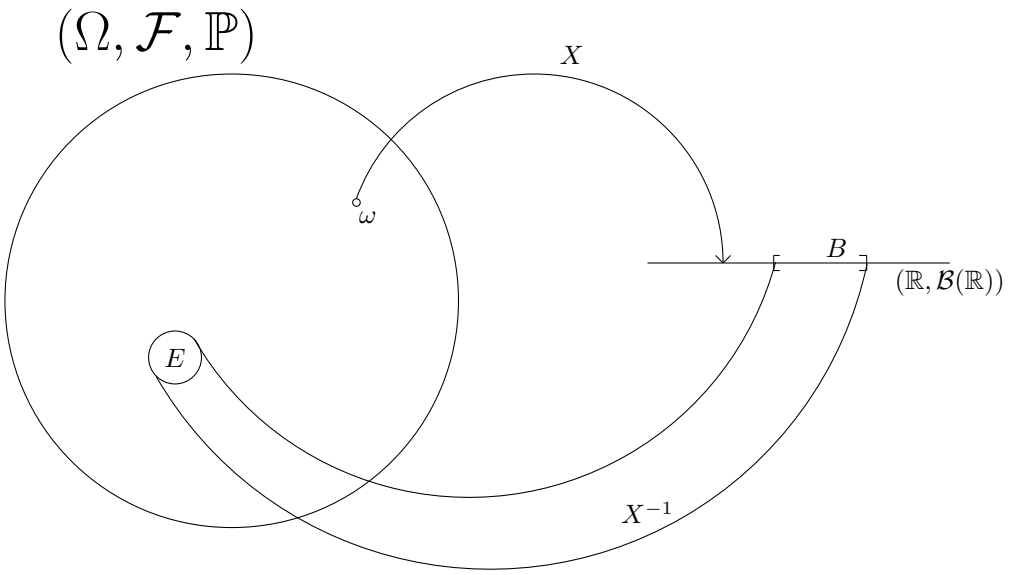

Figure 11.1: A random variable $X : \Omega \to \mathbb{R}$. The pre-image of a Borel set B is an event E.

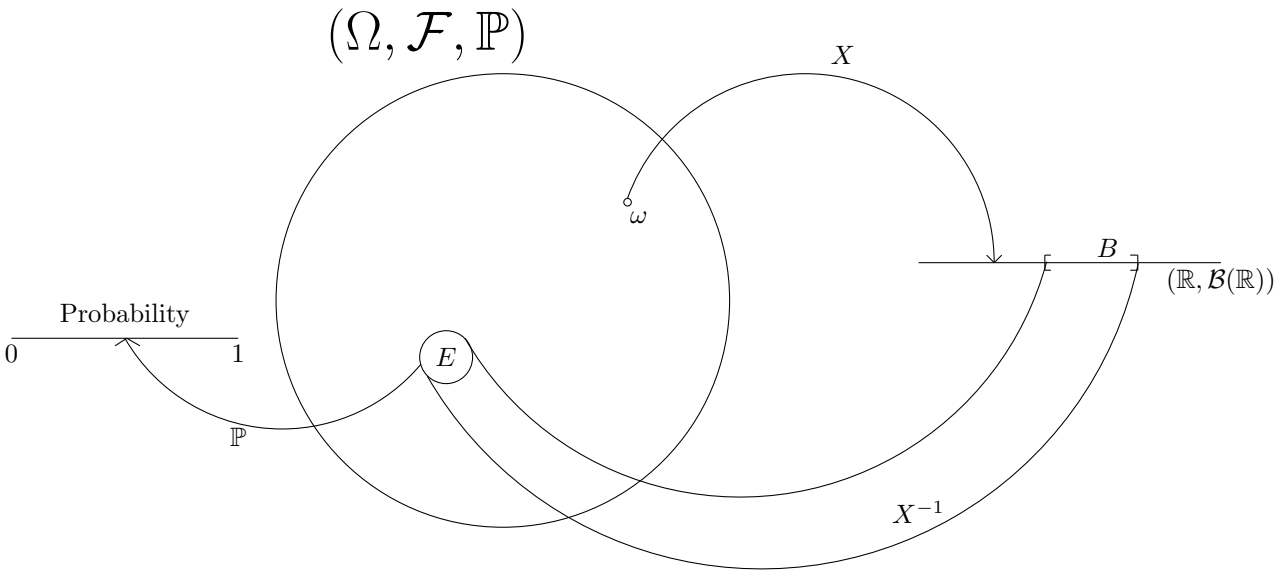

Figure 11.2: The probability law $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$ specifies the probability that the random variable X takes a value in some particular Borel set.

the random variable, as it specifies the probability of the random variable taking values in any given Borel set.

In Figure 11.2, $\mathbb{P}$ is the mapping from event E to the probability space. $\mathbb{P}_X$ is the mapping from B to the probability space such that $\mathbb{P}_X$ is a composition of $\mathbb{P}$ with X^{-1}.

Theorem 11.4 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let X be a real-valued random variable. Then, the probability law $\mathbb{P}_X$ of X is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*

Next, we make a short digression to introduce a mathematical structure known as a *π-system* (read as pi-system).

Definition 11.5 *Given a set Ω, a π-system on Ω is a non-empty collection of subsets of Ω that is stable under finite intersections. That is, $\mathcal{P}$ is a π-system on Ω, if $A, B \in \mathcal{P}$ implies $A \cap B \in \mathcal{P}$.*

One of the most commonly used π-systems on $\mathbb{R}$ is the class of all closed semi-infinite intervals defined as

$$\pi(\mathbb{R}) \triangleq \{(-\infty, x] : x \in \mathbb{R}\}. \tag{11.2}$$

Lemma 11.6 *The σ-algebra generated by $\pi(\mathbb{R})$ is the Borel σ-algebra, i.e.,*

$$\mathcal{B}(\mathbb{R}) = \sigma(\pi(\mathbb{R})).$$

Now, we turn our attention to a key result from measure theory, which states that if two finite measures agree on a π-system, then they also agree on the σ-algebra generated by that π-system.

Lemma 11.7 ***Uniqueness of extension, π-systems:-*** *Let Ω be a given set, and let $\mathcal{P}$ be a π-system over Ω. Also, let $\Sigma = \sigma(\mathcal{P})$ be the σ-algebra generated by the π-system $\mathcal{P}$. Suppose μ_1 and μ_2 are measures defined on the measurable space (Ω, Σ) such that $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ and $\mu_1 = \mu_2$ on $\mathcal{P}$. Then,*

$$\mu_1 = \mu_2 \text{ on } \Sigma.$$

Proof: See Section A1.4 of [1]. ■

In particular, for probability measures, we have the following corollary:

Corollary 11.8 *If two probability measures agree on a π-system, then they agree on the σ-algebra generated by that π-system.*

In particular, if two probability measures agree on $\pi(\mathbb{R})$, then they must agree on $\mathcal{B}(\mathbb{R})$. This result is of importance to us since working with σ-algebras is difficult, whereas working with π-systems is easy!

11.1 Cumulative Distribution Function (CDF) of a Random Variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a Probability Space and let $X : \Omega \longrightarrow \mathbb{R}$ be a random variable. Consider the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ induced on the real line by X. Recall that $\mathcal{B}(\mathbb{R}) = \sigma(\pi(\mathbb{R}))$ is the Borel σ-algebra whose

generating class is the collection of semi-infinite intervals (or equivalently, the open intervals). Therefore, for any $x \in \mathbb{R}$,

$$(-\infty, x] \in \mathcal{B}(\mathbb{R}) \Rightarrow X^{-1}\left((-\infty, x]\right) \in \mathcal{F}.$$

It is therefore legitimate to look at the probability law $\mathbb{P}_X$ of these semi-infinite intervals. This is, by definition, the Cumulative Distribution Function (CDF) of X, and is denoted by $F_X(.)$.

Definition 11.9 *The CDF of a random variable X is defined as follows:*

$$F_X(x) \triangleq \mathbb{P}_X\left((-\infty, x]\right) = \mathbb{P}\left(\{\omega | X(\omega) \leq x\}\right), \quad x \in \mathbb{R}. \tag{11.3}$$

Since the notation $\mathbb{P}(\{\omega | X(\omega) \leq x\})$ is a bit tedious, we will use $\mathbb{P}(X \leq x)$ although it is an abuse of notation. Remarkably, it turns out that it is enough to specify the CDF in order to completely characterize the probability law of the random variable! The following theorem asserts this:

Theorem 11.10 *The Probability Law $\mathbb{P}_X$ of a random variable X is uniquely specified by its CDF $F_X(.)$.*

Proof: This is a consequence of the uniqueness result, Lemma 11.7. Another approach is to use the Carathèodory's extension theorem. Here, we present only an overview of the proof.
Let $\mathcal{F}_0$ denote the collection of finite unions of sets of the form $(a, b]$, where $a < b$ and $a, b \in \mathbb{R}$. Define a set function $\mathbb{P}^0 : \mathcal{F}_0 \longrightarrow [0, 1]$ as $\mathbb{P}^0((a, b]) = F_X(a) - F_X(b)$. Having verified countable additivity of $\mathbb{P}^0$ on $\mathcal{F}_0$, we can invoke Carathèodory's Theorem, thereby obtaining a measure $\mathbb{P}_X$ which uniquely extends $\mathbb{P}^0$ on $\mathcal{B}(\mathbb{R})$.

11.2 Properties of CDF

Theorem 11.11 *Let X be a random variable with CDF $F_X(.)$. Then $F_X(.)$ posses the following properties:*

1. *If $x \leq y$, then $F_X(x) \leq F_X(y)$ i.e. the CDF is monotonic non-decreasing in its argument.*
2. *$\lim_{x \longrightarrow \infty} F_X(x) = 1$ and $\lim_{x \longrightarrow -\infty} F_X(x) = 0$.*
3. *$F_X(.)$ is right-continuous i.e. $\forall\ x \in \mathbb{R}$, $\lim_{\epsilon \downarrow 0} F_X(x + \epsilon) = F_X(x)$.*

Proof:

1. Since, for $x \leq y$, $\{\omega | X(\omega) \leq x\} \subseteq \{\omega | X(\omega) \leq y\}$, from monotonicity of the probability measure, it follows that $F_X(x) \leq F_X(y)$.
2. We have

$$\begin{aligned}
\lim_{x \longrightarrow -\infty} F_X(x) &\overset{(a)}{=} \lim_{x \longrightarrow -\infty} \mathbb{P}(X \leq x), \\
&\overset{(b)}{=} \lim_{n \longrightarrow \infty} \mathbb{P}(X \leq x_n), \\
&\overset{(c)}{=} \mathbb{P}\left(\bigcap_{n \in \mathbb{N}} \{\omega : X(\omega) \leq x_n\}\right), \\
&= \mathbb{P}(\emptyset) = 0,
\end{aligned}$$

where (a) follows from the definition of a CDF, (b) follows by considering a sequence $\{x_n\}_{n\in\mathbb{N}}$ that decreases monotonically to $-\infty$, and (c) is a consequence of continuity of probability measures.

Following a very similar derivation, and considering a sequence $\{x_n\}_{n\in\mathbb{N}}$ that monotonically increases to $+\infty$, we get:

$$\begin{aligned}
\lim_{x\longrightarrow\infty} F_X(x) &= \lim_{x\longrightarrow\infty} \mathbb{P}(X \leq x), \\
&= \lim_{n\longrightarrow\infty} \mathbb{P}(X \leq x_n), \\
&= \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} \{\omega : X(\omega) \leq x_n\}\right), \\
&= \mathbb{P}(\Omega), \\
&= 1.
\end{aligned}$$

3. Consider a sequence $\{\epsilon_n\}_{n\in\mathbb{N}}$ decreasing to zero. Therefore, for each $x \in \mathbb{R}$,

$$\begin{aligned}
\lim_{\epsilon\downarrow 0} F_X(x+\epsilon) &\overset{(a)}{=} \lim_{\epsilon\downarrow 0} \mathbb{P}(X \leq x+\epsilon), \\
&= \lim_{n\longrightarrow\infty} \mathbb{P}(X \leq x+\epsilon_n), \\
&\overset{(b)}{=} \mathbb{P}\left(\bigcap_{n\in\mathbb{N}} \{\omega : X(\omega) \leq x+\epsilon_n\}\right), \\
&= \mathbb{P}(X \leq x), \\
&= F_X(x),
\end{aligned}$$

where (a) follows from the definition of CDF, and (b) follows from continuity of probability measures.

Note that, in general, a CDF need not be continuous. But right-continuity must necessarily be satisfied by any CDF. It turns out that not only are the above three properties satisfied by all CDFs, but any function that satisfies these properties is necessarily a CDF of some random variable!

Theorem 11.12 *Let F be a function satisfying the three properties of a CDF as in theorem* (11.11)*. Consider the Probability Space $\Omega = ([0,1), \mathcal{B}([0,1)), \lambda)$. Then, there exists a random variable $X : \Omega \longrightarrow \mathbb{R}$ whose CDF is F.*

A constructive proof can be found in [3].

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 11: Random Variables: Types and CDF

Lecturer: Dr. Krishna Jagannathan *Scribe: Sudharsan, Gopal, Arjun B, Debayani*

In this lecture, we will focus on the types of random variables. Random variables are categorized into various types, depending on the nature of the measure $\mathbb{P}_X$ induced on the real line (or to be more precise, on the Borel σ-algebra). Indeed, there are three fundamentally different types of measures possible on the real line. According to an important theorem in measure theory, called the Lebesgue decomposition theorem (see Theorem 12.1.1 of [2]), any probability measure on $\mathbb{R}$ can be uniquely decomposed into a sum of these three types of measures. The three fundamental types of measure are

- Discrete,
- Continuous, and
- Singular.

In other words, there are three 'pure type' random variables, namely discrete random variables, continuous random variables, and singular random variables. It is also possible to 'mix and match' these three types to get four kinds of mixed random variables, altogether resulting in *seven* types of random variables.

Of the three fundamental types of random variables, only the discrete and continuous random variables are important for practical applications in the field of engineering and statistics. Singular random variables are largely of academic interest. Therefore, we will spend most of our effort in studying discrete and continuous random variables, although we will define and give an example of a singular random variable.

11.1 Discrete Random Variables

Definition 11.1 *Discrete Random Variable:*
A random variable X is said to be discrete if it takes values in a countable subset of $\mathbb{R}$ with probability 1.

Thus, there is a countable set $E = \{x_1, x_2, \dots\}$, such that $\mathbb{P}_X(E) = 1$. Note that the definition does not necessarily demand that the range of the random variable is countable. In particular, for a discrete random variable, there might exist some zero probability subset of the sample space, which can potentially map to an uncountable subset of $\mathbb{R}$. (Can you think of such an example?)

Definition 11.2 *Probability Mass Function (PMF):*
If X is a discrete random variable, the function $p_X : \mathbb{R} \to [0, 1]$ defined by $p_X(x)$=$\mathbb{P}(X = x)$ for every x is called the probability mass function of X.

Although the PMF is defined for all $x \in \mathbb{R}$, it is clear from the definition that the PMF is non-zero only on the set E. Also, since $\mathbb{P}_X(E) = 1$, we must have (by countable additivity)

$$\sum_{i=1}^{\infty} \mathbb{P}(X = x_i) = 1.$$

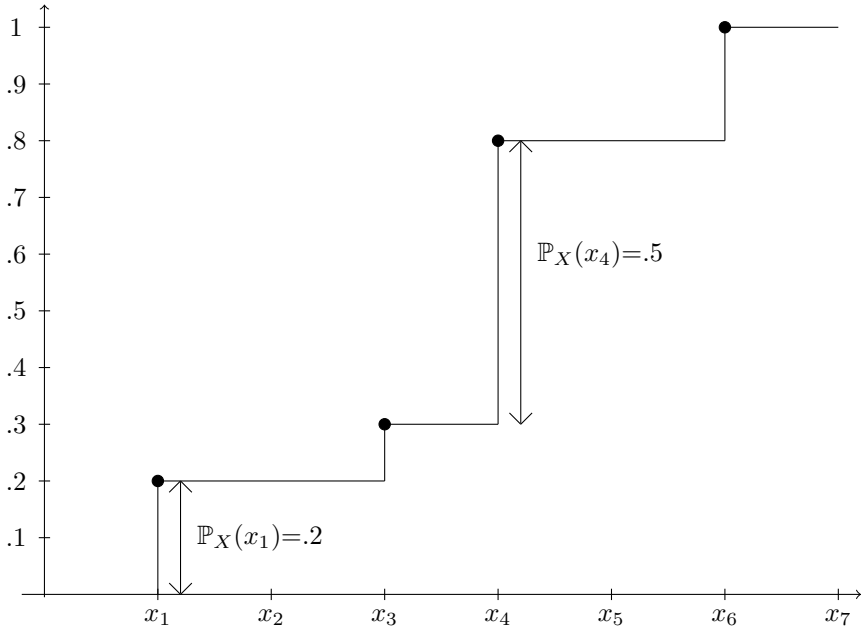

Figure 11.1: CDF of a discrete random variable

Interestingly, for a discrete random variable X, the PMF is enough to get a complete characterization of the probability law $\mathbb{P}_X$. Indeed, for any Borel set B, we can write

$$\mathbb{P}_X(B) = \sum_{i:\ x_i \in B} \mathbb{P}(X = x_i).$$

The CDF of a discrete random variable is given by

$$F_X(x) = \sum_{i: x_i \leq x} \mathbb{P}(X = x_i).$$

Figure 15.3 represents the Cumulative Distribution Function of a discrete random variable. One can observe that the CDF plotted in Figure 15.3 satisfies all the properties discussed earlier.

Next, we give some examples of some frequently encountered discrete random variables.

1. Indicator random variable: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A \in \mathcal{F}$ be any event. Define

$$I_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

   It can be verified that I_A is indeed a random variable (since A and A^c are $\mathcal{F}$-measurable), and it is clearly discrete, since it takes only two values.

2. Bernoulli random variable: Let $p \in [0, 1]$, and define $p_X(0) = p$, and $p_X(1) = 1 - p$. This random variable can be used to model a single coin toss, where 0 denotes a head and 1 denotes a tail, and the probability of heads is p. The case $p = 1/2$ corresponds to a fair coin toss.

3. Discrete uniform random variable: Parameters are a and b where $a < b$.
   $p_X(m) = 1/(b - a + 1)$, $m = a$, $a + 1$,..b, and $p_X(m)$=0 otherwise.

4. Binomial random variable: $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, where $n \in \mathbb{N}$ and $p \in [0,1]$. In the coin toss example, a binomial random variable can be used to model the number of heads observed in n independent tosses, where p is the probability of head appearing during each trial.
5. Geometric random variable: $p_X(k) = p(1-p)^{k-1},\ k = 1, 2, \ldots$ and $0 < p \leq 1$. A geometric random variable with parameter p represents the number of (independent) tosses of a coin until heads is observed for the first time, where p represents the probability of heads during each toss.
6. Poisson: Fix the parameter $\lambda > 0$, and define $p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$, where $k = 0, 1, \ldots$

Note that except for the indicator random variable, we have described only the PMFs of the random variables, rather than the explicit mapping from Ω.

11.2 Continuous Random Variables

11.2.1 Definitions

Let us begin with the definition of absolute continuity which will allow us to define continuous random variables formally. Let μ and ν be measures on $(\Omega, \mathcal{F})$.

Definition 11.3 *We say ν is* absolutely continuous *with respect to μ if for every $N \in \mathcal{F}$ such that $\mu(N) = 0$, we have $\nu(N) = 0$.*

Now, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable.

Definition 11.4 *X is said to be a continuous random variable if the law $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure λ.*

Here, both $\mathbb{P}_X$ and λ are measures on $(\mathbb{R}, \mathcal{B})$. The above definition says that X is a continuous random variable if for any Borel set N set of Lebesgue measure zero, we have $\mathbb{P}_X(N) = \mathbb{P}(\omega | X(\omega) \in N) = 0$.

In particular, it is *not* the case that a random variable is continuous if it takes values in an uncountable set.

Next, we invoke without proof a special case of the *Radon-Nikodym Theorem* [3], which deals with absolutely continuous measures.

Theorem 11.5 *Suppose $\mathbb{P}_X$ is absolutely continuous with respect to λ, the Lebesgue measure, then there exists a non-negative, measurable function $f_X : \mathbb{R} \to [0, \infty)$, such that for any $B \in \mathcal{B}(\mathbb{R})$, we have*

$$\mathbb{P}_X(B) = \int_B f_X \, d\lambda. \tag{11.1}$$

The integral in the above theorem is not the usual Riemann integral, as B may be any Borel measurable set, such as the Cantor set, for example. We will get a precise understanding of the integral in (15.1) when we study abstract integration later in the course. For the time being, we can just think of the set B as an interval $[a, b]$, so (15.1) essentially says that the probability of X taking values in the interval $[a, b]$ can be written as $\int_a^b f_X dx$ for some non-negative measurable function f_X. Here, when we say f_X is measurable, we mean the pre-images of Borel sets are also Borel sets. In measure theoretic parlance, f_X is called the Radon-Nikodym derivative of $\mathbb{P}_X$ with respect to the Lebesgue measure λ.

In particular, taking $B = (-\infty, x]$, we can write the cumulative distribution function (CDF) as

$$F_X(x) \triangleq \mathbb{P}_X((-\infty, x]) = \int_{-\infty}^{x} f_X(y)\, dy. \tag{11.2}$$

Thus, we can understand f_X as the *probability density function (PDF)* of X, which is nothing but the Radon-Nikodym derivative of $\mathbb{P}_X$ with respect to the Lebesgue measure λ. Also,

$$\mathbb{P}_X(\mathbb{R}) = 1 = \int_{-\infty}^{\infty} f_X(y)\, dy.$$

Unlike the probability mass function in the case of a discrete random variable, the PDF has *no* interpretation as a probability; only integrals of the PDF can be interpreted as a probability.

The function f_X is unique only up to a set of Lebesgue measure zero, as we will understand later. We also remark that many authors (including [4]) define a random variable as being continuous if the CDF satisfies (15.2). This definition can be shown to be equivalent to the one we have given above.

11.2.2 Examples

The following are some common examples of continuous random variables:

1. Uniform: It is a scaled Lebesgue measure on a closed interval $[a, b]$.

   (a) PDF- $f_X(x) = \begin{cases} 0 & \text{for} \quad x < a \\ \frac{1}{b-a} & \text{for} \quad a \le x \le b \\ 0 & \text{for} \quad x > b \end{cases}$

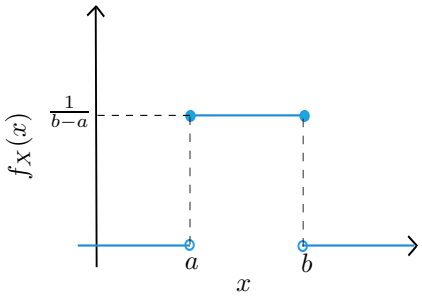

Figure 11.2: The PDF of a uniform random variable

   (b) CDF: $F_X(x) = \begin{cases} 0 & \text{for} \quad x < a \\ \frac{x-a}{b-a} & \text{for} \quad a \le x \le b \\ 1 & \text{for} \quad x > b \end{cases}$

2. Exponential: It is a non-negative random variable, characterized by a single parameter $\lambda > 0$.

   (a) PDF: $f_X(x) = \lambda e^{-\lambda x} \quad \text{for} \quad x \ge 0$

   (b) CDF: $F_X(x) = 1 - e^{-\lambda x} \quad \text{for} \quad x \ge 0$

   (c) The exponential random variable posses an interesting property called the 'memoryless' property. We first give the definition of the memoryless property, and then show that the exponential random variable has this property.

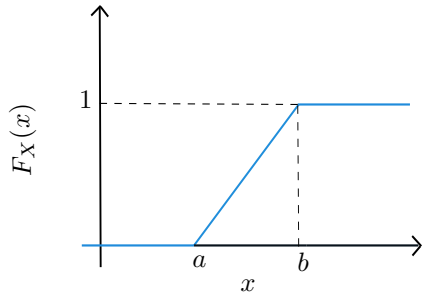

Figure 11.3: The CDF of a uniform random variable

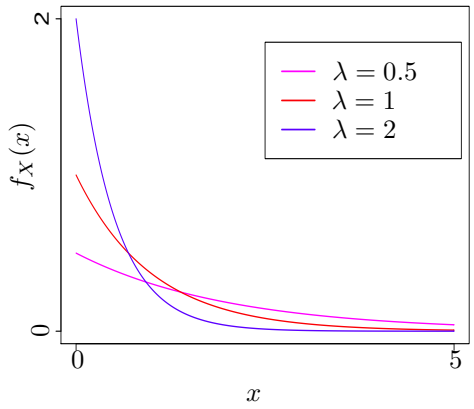

Figure 11.4: The PDF of an exponential random variable, for various values of the parameter λ

Definition 11.6 *A non-negative random variable X is said to be memoryless if* $\mathbb{P}(X > s+t|X > t) = \mathbb{P}(X > s) \quad \forall s, t \geq 0$.

For an exponential random variable,

$$\begin{aligned}
\mathbb{P}(X > s+t|X > t) &= \frac{\mathbb{P}((X > s+t)\,\&\,(X > t))}{\mathbb{P}(X > t)} \\
&= \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > t)} \\
&= \frac{e^{-(s+t)\lambda}}{e^{-t\lambda}} \\
&= e^{-s\lambda} \\
&= \mathbb{P}(X > s).
\end{aligned}$$

Therefore, the exponential random variable is memoryless. For example, if the failure time of a light bulb is distributed exponentially, then the further time to failure, given that the bulb has not failed until time t, has the same distribution as the unconditional failure time of a new light bulb! Interestingly, it can also be shown that the exponential random variable is the *only* continuous random variable which possesses the memoryless property.

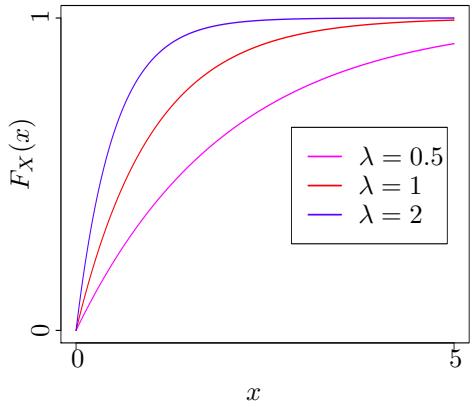

Figure 11.5: The CDF of an exponential random variable, for various values of the parameter λ

3. Gaussian (or Normal): This is a two parameter distribution, and as we shall interpret later, these parameters are the mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$. It has wide applications in engineering and statistics, owing to a 'stable-attractor" property of Gaussian random variables. We will study these properties later.

   (a) PDF: The probability density function of a Gaussian random variable is given by $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ for $x \in \mathbb{R}$.
   The above distribution is denoted $N(\mu, \sigma^2)$. In particular, when $\mu = 0$ and $\sigma^2 = 1$, we get the *standard Gaussian* PDF: $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

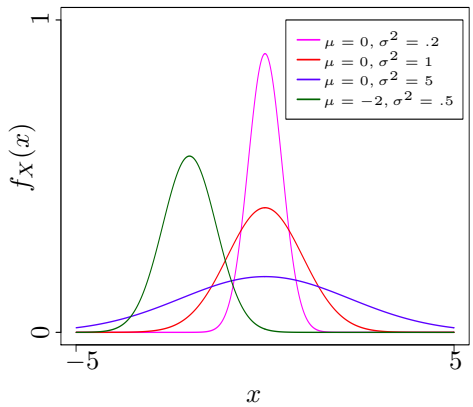

Figure 11.6: The PDF of a normal random variable, for various parameters

(b) CDF: There is no closed-form expression for the CDF of a Gaussian distribution (although the notion of a 'closed-form' is itself rather arbitrary, and over-rated!). For convenience, we call the CDF of the standard Gaussian the "error-function" $\text{Erf}(x) \triangleq \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy$.

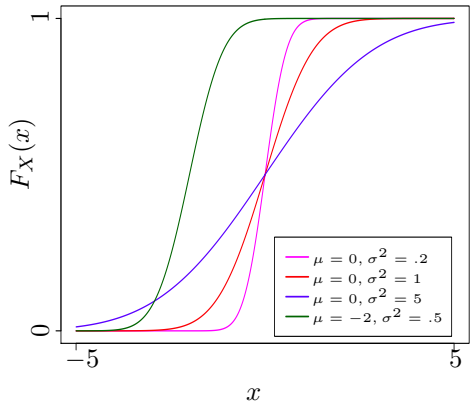

Figure 11.7: The CDF of a normal random variable, for various parameters

4. Cauchy: This is a two-parameter distribution parametrised by $x_0 \in \mathbb{R}$, the centering parameter, and $\gamma > 0$, the scale parameter. It is qualitatively very different from the previous distributions, because it is "heavy-tailed," i.e., its complementary CDF $1 - F_X(x)$ decays slower than any exponential. Heavy-tailed random variables tend to take very large values with non-negligible probability, and are used to model high variability and burstiness in engineering applications.

   (a) PDF: $f_X(x) = \frac{1}{\pi} \frac{\gamma}{(x-x_0)^2+\gamma^2}$.

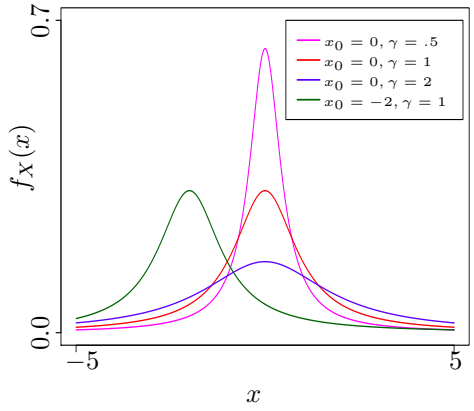

Figure 11.8: The PDF of a Cauchy random variable, for different parameters

11.3 Singular Random Variable

Singular random variables are rather bizzare, and in some sense, they occupy the 'middle-ground' between discrete and continuous random variables. In particular, singular random variables take values with probability one on an uncountable set of Lebesgue measure zero!

Definition 11.7 *A random variable X is said to be singular if, for every $x \in \mathbb{R}$, we have $\mathbb{P}_X(\{x\}) = 0$, and there exists a zero Lebesgue measure set $F \in \mathcal{B}(\mathbb{R})$, such that $\mathbb{P}_X(F) = 1$.*

Although it is not stated explicitly in the definition, it is clear that F must be an *uncountable* set of Lebesgue measure zero. (Why?)

Example A random variable having the Cantor distribution as its CDF is an example of a Singular random variable. The range of this random variable is the Cantor Set, C, which is a Borel set with Lebesgue measure zero. Further, if $x \in C$, then x has a ternary expansion of the following form

$$x = \sum_{i=1}^{\infty} \frac{x_i}{3^i}, \quad \text{where} \quad x_i \in \{0, 2\}\,. \tag{11.3}$$

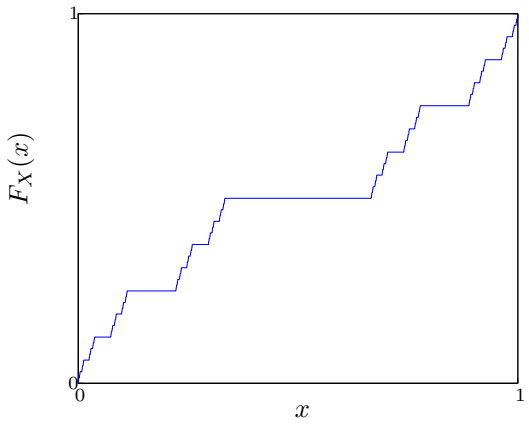

Figure 11.9: The Cantor Function

To look at a concrete example, consider an infinite sequence of independent tosses of a fair coin. When the outcome is a head, we record $x_i = 2$, otherwise, we record $x_i = 0$. Using these values of x_i we form a number x using (15.3). This results in a random variable X. This random variable satisfies the two properties that make it a Singular Random variable, namely $\mathbb{P}_X(C) = 1$, and $\mathbb{P}_X(\{x\}) = 0$, $\forall\ x \in [0, 1]$. The cumulative distribution function of this random variable, shown in Figure 15.9, is the *Cantor function* (which is sometimes referred to as the Devil's staircase). The Cantor function is continuous everywhere, since all singletons have zero probability under this distribution. Also, the derivative is zero wherever it exists, and the derivative does not exist at points in the Cantor set. The CDF only increases at these Cantor points, but does so without a well defined derivative, or any jump discontinuities for that matter!

11.4 Exercises:

1. (a) Prove Theorem 15.4.

(b) Verify that $\pi(\mathbb{R})$, defined in the lecture on Random Variables is indeed a π-system over $\mathbb{R}$.

(c) Prove Lemma 15.6.

(d) Plot the CDF of the indicator random variable.

2. For a random variable X, prove that $\mathbb{P}_X(\{y\}) = F_X(y) - \lim_{x \uparrow y} F_X(x)$. Hence show that F_X is continuous at y if and only if $\mathbb{P}_X(\{y\}) = 0$.

3. Among the functions given below, find the functions that are valid CDFs and find their respective densities. For those that are not valid CDFs, explain what fails.

(a)
$$F(x) = \begin{cases} 1 - e^{-x^2} & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{11.4}$$

(b)
$$F(x) = \begin{cases} e^{-\frac{1}{x}} & x > 0 \\ 0 & x \leq 0. \end{cases} \tag{11.5}$$

(c)
$$F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{3} & 0 < x \leq \frac{1}{2} \\ 1 & x > \frac{1}{2}. \end{cases} \tag{11.6}$$

4. **Negative Binomial Random Variable.** Consider a sequence of independent Bernoulli trials $\{X_i\}_{i \in \mathbb{N}}$ with parameter of success $p \in (0, 1]$. The number of successes in first n trials is given by

$$Y_n = \textstyle\sum_{i=1}^{n} X_i.$$

Y_n is distributed as Binomial with parameters n and p.
Consider the random variable defined by

$$V_k = \min\{n \in \mathbb{N}_+ : Y_n = k\}.$$

Note that V_1 is distributed as Geometric with parameter p.

(a) Give a verbal description of the random variable V_k.

(b) Show that the probability mass function of the random variable V_k is given by

$$\mathbb{P}(V_k = n) = (\tfrac{n-1}{k-1}) p^k (1-p)^{(n-k)}$$

where $n \in \{k, k+1, ...\}$. This is known as Negative Binomial Distribution with parameters k and p.

(c) Argue that Binomial and Negative Binomial Distributions are inverse to each other in the sense that

$$Y_n \geq k \Leftrightarrow V_k \leq n.$$

5. **Radioactive decay.** Assume that a radioactive sample emits a random number of α particles in any given hour, and that the number of α particles emitted in an hour is Poisson distributed with parameter λ. Suppose that a faulty Geiger-Muller counter is used to count these particle emissions. In particular, the faulty counter fails to register an emission with probability p, independently of other emissions.

(a) What is the probability that the faulty counter will register exactly k emissions in an hour?

(b) Given that the faulty counter registered k emissions in an hour, what is the PMF of the actual number of emissions that happened from the source during that hour?

6. Buses arrive at ten minute intervals starting at noon. A man arrives at the bus stop at a random time X minutes after noon, where X has the CDF:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{60} & 0 \leq x \leq 60 \\ 1 & x > 60. \end{cases} \tag{11.7}$$

What is the probability that he waits less than five minutes for a bus?

7. Find the values of a and b such that the following function is a valid CDF:

$$F(x) = \begin{cases} 1 - ae^{-x/b} & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{11.8}$$

Also, find the values of a and b such that the function above corresponds to the CDF of some

(a) Continuous Random Variable

(b) Discrete Random Variable

(c) Mixed type Random Variable

8. Let X be a continuous random variable. Show that X is memoryless iff X is an exponential random variable.

References

[1] David Williams, "Probability with Martingales", *Cambridge University Press*, Fourteenth Printing, 2011.

[2] Paul Halmos, "Measure Theory", *Springer-Verlog*, Second Edition, 1978.

[3] David Gamarnick and John Tsitsiklis, "Introduction to Probability", *MIT OCW*, , 2008.

[4] Geoffrey Grimmett and David Stirzaker, "Probability and Random Processes", *Oxford University Press*, Third Edition, 2001.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 12: Multiple Random Variables and Independence

Instructor: Dr. Krishna Jagannathan Scribes: Debayani Ghosh, Gopal Krishna Kamath M, Ravi Kolla

12.1 Multiple Random Variables

In this lecture, we consider multiple random variables defined on the same probability space. To begin with, let us consider two random variables X and Y, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. It is important to understand that the realizations of X and Y are governed by the same underlying randomness, namely $\omega \in \Omega$. For example, the underlying sample space could be something as complex as the weather on a particular day; the random variable X could denote the temperature on that day, and another random variable Y, the humidity level. Since the same underlying outcome governs both X and Y, it is reasonable to expect X and Y to posses a certain degree of interdependence. In the above example, a high temperature on a given day usually says something about the humidity.

In Figure (12.1), the top picture shows two random variables X and Y, each mapping Ω to $\mathbb{R}$. These two random variables are measurable functions from the same probability space to the real line. The bottom picture in Figure (12.1) shows $(X(\cdot), Y(\cdot))$ mapping Ω to $\mathbb{R}^2$. Indeed, the bottom picture is more meaningful, since it captures the interdependence between X and Y.

Now, an important question arises: is the function $(X(\cdot), Y(\cdot)) : \Omega \to \mathbb{R}^2$ measurable, given that X and Y are measurable functions? In order to pose this question properly and answer it, we first need to define the Borel σ-algebra on $\mathbb{R}^2$. The Borel σ-algebra on $\mathbb{R}^2$ is the σ-algebra generated by the class $\pi\left(\mathbb{R}^2\right) \triangleq \{(-\infty, x] \times (-\infty, y] \,|\, x, y \in \mathbb{R}\}$. That is,

$$\mathcal{B}(\mathbb{R}^2) = \sigma\left(\pi\left(\mathbb{R}^2\right)\right).$$

The following theorem asserts that whenever X and Y are random variables, the function $(X, Y) : \Omega \to \mathbb{R}^2$ is $\mathcal{F}$-measurable, in the sense that the pre-images of Borel sets on $\mathbb{R}^2$ are necessarily events.

Theorem 12.1 *Let X and Y be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, $(X(\cdot), Y(\cdot)) : \Omega \to \mathbb{R}^2$ is $\mathcal{F}$-measurable, i.e., the pre-images of Borel sets on $\mathbb{R}^2$ under $(X(\cdot), Y(\cdot))$ are events.*

Proof: Let $\mathcal{G}$ be the collection of all subsets of $\mathbb{R}^2$ whose pre-images under $(X(\cdot), Y(\cdot))$ are events. To prove the theorem, it is enough to prove that $\mathcal{B}\left(\mathbb{R}^2\right) \subseteq \mathcal{G}$.

Claim 1: $\mathcal{G}$ is a σ- algebra of subsets of $\mathbb{R}^2$.

Next, note that $\{\omega | X(\omega) \leq x\}$, $\{\omega | Y(\omega) \leq y\} \in \mathcal{F}$, $\forall x, y \in \mathbb{R}$, since X and Y are random variables. Thus, $\{\omega | X(\omega) \leq x\} \cap \{\omega | Y(\omega) \leq y\} \in \mathcal{F}$, $\forall x, y \in \mathbb{R}$, since $\mathcal{F}$ is a σ- algebra.

So, $\{\omega | X(\omega) \leq x, Y(\omega) \leq y\} \in \mathcal{F}$, $\forall x, y \in \mathbb{R} \Rightarrow (-\infty, x] \times (-\infty, y] \in \mathcal{G}$, $\forall x, y \in \mathbb{R}$ (from the definition of $\mathcal{G}$) $\Rightarrow \pi\left(\mathbb{R}^2\right) \subseteq \mathcal{G} \Rightarrow \sigma\left(\pi\left(\mathbb{R}^2\right)\right) \subseteq \sigma(\mathcal{G}) \Rightarrow \mathcal{B}\left(\mathbb{R}^2\right) \subseteq \mathcal{G}$. ■

Since the pre-images of Borel sets on $\mathbb{R}^2$ are events, we can assign probabilities to them. This leads us to the definition of the joint probability law.

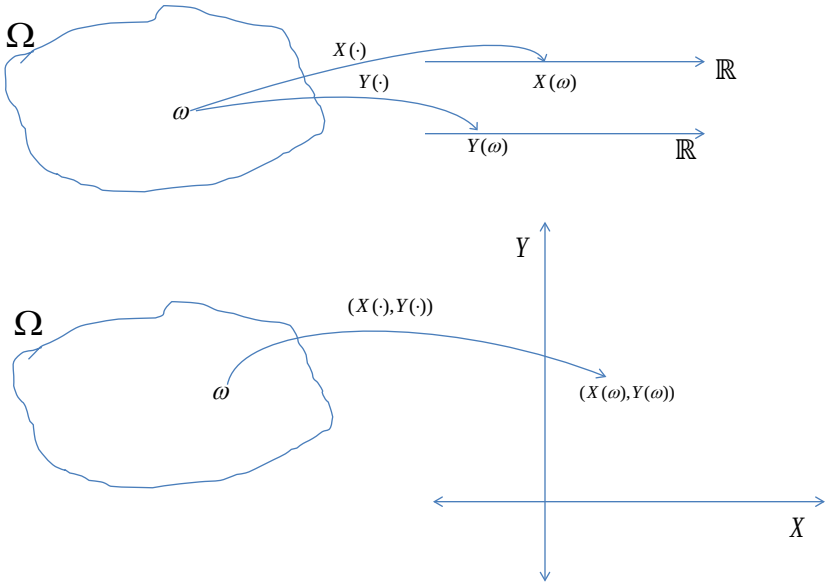

Figure 12.1: Illustration of Multiple Random Variables

Definition 12.2 *The joint probability law of the random variables X and Y is defined as:*

$$\mathbb{P}_{X,Y}(B) = \mathbb{P}\left(\{\omega \in \Omega | \left(X(\omega), Y(\omega)\right) \in B\}\right), \quad B \in \mathcal{B}(\mathbb{R}^2),$$

where $\mathcal{B}(\mathbb{R}^2)$ is the Borel σ-algebra on $\mathbb{R}^2$.

In particular, when $B = (-\infty, x] \times (-\infty, y]$, we have

$$\mathbb{P}_{X,Y}\left((-\infty, x] \times (-\infty, y]\right) = \mathbb{P}\left(\{\omega | X(\omega) \leq x, Y(\omega) \leq y\}\right). \tag{12.1}$$

The LHS in (12.1) is well defined, and hence the RHS in (12.1) is well defined and is called as the joint CDF of X and Y.

12.2 Joint CDF

Definition 12.3 *Let X and Y be two random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The joint CDF of X and Y is defined as follows:*

$$F_{X,Y}(x, y) = \mathbb{P}\left(\{\omega | X(\omega) \leq x, Y(\omega) \leq y\}\right), \quad \forall x, y \in \mathbb{R}.$$

In short hand, we write $F_{X,Y}(x, y) = \mathbb{P}\left(X \leq x, Y \leq y\right)$.

12.2.1 Properties of joint CDF:

1. $\lim_{\substack{x\to\infty\\y\to\infty}} F_{X,Y}(x,y) = 1, \qquad \lim_{\substack{x\to-\infty\\y\to-\infty}} F_{X,Y}(x,y) = 0.$
   Proof: Let $\{x_n\}$ and $\{y_n\}$ be two unbounded, monotone-increasing sequences. We have

$$\begin{aligned}\lim_{\substack{x\to\infty\\y\to\infty}} F_{X,Y}(x,y) &= \lim_{\substack{x\to\infty\\y\to\infty}} \mathbb{P}(X\le x, Y\le y),\\ &= \lim_{n\to\infty} \mathbb{P}(X\le x_n, Y\le y_n),\\ &\overset{(a)}{=} \mathbb{P}\left(\bigcup_{n=1}^{\infty}\{\omega : X(\omega)\le x_n, Y(\omega)\le y_n\}\right),\\ &= \mathbb{P}(\Omega),\\ &= 1,\end{aligned}$$

   where (a) is due to continuity of probability measures (Lecture #5, Property 6). Proof of the other part follows on the similar lines and is left as an exercise to the reader.
   Note that the order of the two limits does not matter here. ■

2. *Monotonicity:* For any $x_1 \le x_2$, $y_1 \le y_2, F_{X,Y}(x_1,y_1) \le F_{X,Y}(x_2,y_2)$.
   Proof: Let $x_1 \le x_2$ and $y_1 \le y_2$. Clearly, events $\{X\le x_1, Y\le y_1\} \subseteq \{X\le x_2, Y\le y_2\}$. Then, $\mathbb{P}(X\le x_1, Y\le y_1) \le \mathbb{P}(X\le x_2, Y\le y_2) \Rightarrow F_{X,Y}(x_1,y_1) \le F_{X,Y}(x_2,y_2)$. ■

3. $F_{X,Y}$ is continuous from above, i.e., $\lim_{\substack{u\to0^+\\v\to0^+}} F_{X,Y}(x+u,y+v) = F_{X,Y}(x,y), \forall x,y\in\mathbb{R}$.
   Exercise: Prove this.

4. $\lim_{y\to\infty} F_{X,Y}(x,y) = F_X(x)$
   Proof: Let $\{y_n\}$ be an unbounded, monotone-increasing sequence. Then, $\lim_{y\to\infty} F_{X,Y}(x,y) = \lim_{n\to\infty} F_{X,Y}(x,y_n)$. Hence,

$$\begin{aligned}\lim_{n\to\infty} F_{X,Y}(x,y_n) &= \lim_{n\to\infty} \mathbb{P}(X\le x, y\le y_n),\\ &\overset{(a)}{=} \mathbb{P}\left(\bigcup_{n=1}^{\infty}\{\omega : X(\omega)\le x, Y(\omega)\le y_n\}\right),\\ &= \mathbb{P}(\omega : X(\omega)\le x),\\ &= F_X(x),\end{aligned}$$

   where (a) is due to continuity of probability measure.(Lecture #5, Property 6). ■
   Using the above property, we can calculate the marginal CDFs from joint CDF. However, the joint CDF cannot be obtained from the marginals alone, since the marginals do not capture the inter-dependence of X and Y.

12.3 The σ-algebra generated by a random variable

Before we proceed to define the independence of random variables, it is useful to understand the notion of the *σ-algebra generated by a random variable*. We first state an elementary result that holds for any arbitrary function.

Proposition 12.4 *Let Ω and S be two non-empty sets and let $f : \Omega \longrightarrow S$ be a function. If $\mathcal{H}$ is a σ-algebra of subsets of S, then $\mathcal{G} \triangleq \{A \mid A = f^{-1}(B)\ , B \in \mathcal{H}\}$ is a σ-algebra of subsets of Ω.*

In words, Proposition (12.4) says that the collection of pre-images of all the sets belonging to some σ-algebra on the range of a function, is a σ-algebra on the domain of that function.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a Probability Space and $X : \Omega \longrightarrow \mathbb{R}$ be a random variable. X in turn induces the probability triple $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ on the real line.

Definition 12.5 *The σ-algebra generated by the random variable X is defined as*

$$\sigma(X) \triangleq \{E \subseteq \Omega | E = X^{-1}(B), \forall B \in \mathcal{B}(\mathbb{R})\}. \tag{12.2}$$

Proposition (12.4) asserts that $\sigma(X)$ defined above is indeed a σ-algebra on Ω.

Proposition 12.6 *$\sigma(X) \subseteq \mathcal{F}$, i.e., the σ-algebra generated by X is a sub-σ-algebra of $\mathcal{F}$.*

Figure (12.3) shows a pictorial representation of the σ-algebra generated by X. Each Borel set B maps back to an event E. A collection of all such preimages of Borel sets constitutes the σ-algebra generated by X. Thus, $\sigma(X)$ is a σ-algebra that consists precisely of those events whose occurrence or otherwise is completely determined by looking at the realised value $X(\omega)$. To get a more concrete idea of this concept, let us look at the following examples:

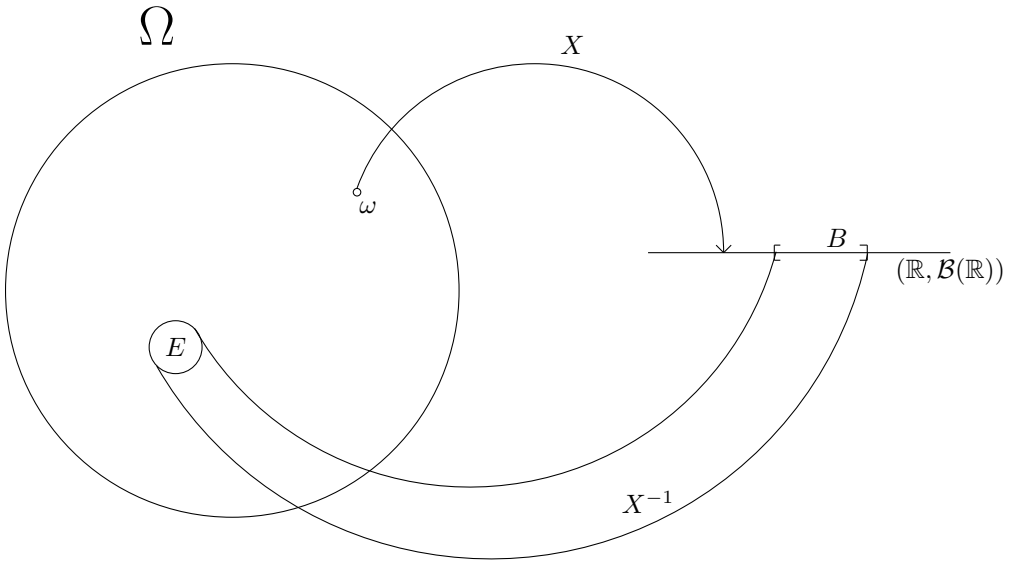

Figure 12.2: The collection of the pre-images of all Borel Sets is the σ-algebra generate by the random variable X, denoted $\sigma(X)$.

Example 1:- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A \in \mathcal{F}$ be some event. Consider the Indicator random variable of event A, $\mathbb{I}_A$. It is easy to see that $\sigma(\mathbb{I}_A) = \{\emptyset, A, A^C, \Omega\}$. Also, $\sigma(\mathbb{I}_A) \subset \mathcal{F}$.

Example 2:- Let $([0,1], \mathcal{B}([0,1]), \lambda)$ be the probability space in consideration, and consider a random variable $X(\omega) = \omega$, $\forall \omega \in \Omega$. It can be seen that $\sigma(X) = \mathcal{F}$.

Remark: 12.7 *As seen from the above two examples, $\sigma(X)$ could either be "small" (as seen in example 1 above) or as "large" as the σ-algebra $\mathcal{F}$ itself (as seen in example 2 above).*

Now, we introduce the important notion of independence of random variables.

12.4 Independence of Random Variables

Definition 12.8 *Random variables X and Y are said to be independent if $\sigma(X)$ and $\sigma(Y)$ are independent σ-algebras.*

In other words, X and Y are independent if, for any two borel sets B_1 and B_2 on $\mathbb{R}$, the events $\{\omega : X(\omega) \in B_1\}$ and $\{\omega : Y(\omega) \in B_2\}$ are independent i.e.,

$$\mathbb{P}\left(\{\omega : X(\omega) \in B_1\} \bigcap \{\omega : Y(\omega) \in B_2\}\right) = \mathbb{P}\left(\{\omega : X(\omega) \in B_1\}\right) \mathbb{P}\left(\{\omega : Y(\omega) \in B_2\}\right), \forall B_1, B_2 \in \mathcal{B}(\mathbb{R}).$$

The following theorem gives a useful characterization of independence of random variables, in terms of the *joint CDF being equal to the product of the marginals.*

Theorem 12.9 *X and Y are independent if and only if*

$$F_{X,Y}(x,y) = F_X(x) F_Y(y).$$

Proof: First, we prove the necessary part, which is straightforward. Let X and Y are independent. Consider $B_1 \in \mathcal{B}(\mathbb{R})$, $B_2 \in \mathcal{B}(\mathbb{R})$ then the events $\{\omega | X(\omega) \in B_1\}$ and $\{\omega | Y(\omega) \in B_2\}$ are independent (due to definition (12.8)) $\Rightarrow \mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}(X \in B_1)\mathbb{P}(Y \in B_2) \Rightarrow \mathbb{P}_{X,Y}(B_1 \times B_2) = \mathbb{P}(X \in B_1)\mathbb{P}(Y \in B_2)$. But, this is true for all borel sets in $\mathbb{R}$. In particular, choose $B_1 = (-\infty, x]$ and $B_2 = (-\infty, y]$ then we get $F_{X,Y}(x,y) = F_X(x)F_Y(y)$, $\forall x, y \in \mathbb{R}$ which completes the proof of the necessary part.

The sufficiency part is more involved; refer [1][Section 4.2]. ■

Definition 12.10 *$X_1, X_2, \cdots X_n$ random variables are said to be independent if σ-algebras $\sigma(X_1), \sigma(X_2), \cdots, \sigma(X_n)$ are independent i.e., for any $B_i \in \mathcal{B}(\mathbb{R})$, $1 \leq i \leq n$, we have*

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2, \cdots X_n \in B_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i).$$

Theorem 12.11 *$X_1, X_2, \cdots X_n$ are independent if and only if*

$$F_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

Proof: Refer [1][Section 4.2]. ■

Finally, we define independence for an arbitrary family of random variables.

Definition 12.12 *An arbitrary family of random variables, $\{X_i, \in I\}$, is said to be independent if the $\sigma-$ algebras $\{\sigma(X_i), i \in I\}$ are independent (Lecture #9, Section 9.2, Definition 9.7).*

12.5 Exercises

1. Prove **Claim 1** under Theorem 12.1.
2. For random variables X and Y defined on same probability space, with joint CDF $F_{X,Y}(x, y)$, prove that $\lim_{\substack{x\to-\infty \\ y\to-\infty}} F_{X,Y}(x, y) = 0$.
3. Prove Propositions (12.4) and (12.6).
4. *[Quiz II 2014]* Suppose X and Y are independent random variables, and $f(X)$ and $g(Y)$ are functions of X and Y respectively. Will the random variables $f(X)$ and $g(Y)$ be independent? Justify your answer.

References

[1] David Williams, "Probability with Martingales", *Cambridge University Press*, Fourteenth Printing, 2011.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 13: Conditional Distributions and Joint Continuity

Lecturer: Dr. Krishna Jagannathan *Scribe: Subrahmanya Swamy P*

13.1 Conditional Probability for Discrete Random Variables

If X and Y are discrete random variables, then the range of the map $(X(\cdot), Y(\cdot))$ is a countable subset of $\mathbb{R}^2$. This is because the Cartesian product of two countable sets is countable (Why?). Hence, $(X(.), Y(.))$ is a discrete random variable on $\mathbb{R}^2$. We will see later that we do not have similar result when X and Y are continuous random variables i.e., if X and Y are marginally continuous random variables, they need not be jointly continuous.

The joint pmf of discrete random variables X and Y is defined as:

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

The joint pmf uniquely specifies the joint law. In particular, for any $B \in \mathcal{B}(\mathbb{R}^2)$,

$$\mathbb{P}_{X,Y}(B) = \sum_{x,y \in B} p_{X,Y}(x,y). \tag{13.1}$$

An example of two discrete random variables is shown in Figure (13.1).

13.1.1 Conditional pmf

Now, we define the conditional pmf for discrete random variables.

Definition 13.1 *Let X and Y be discrete random variables defined on $(\Omega, \mathbb{F}, \mathbb{P})$. Conditional probability of X given Y is defined as:*

$$p_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)} \ \text{ where } p_Y(y) > 0.$$

The following theorem characterizes independence of discrete random variables in terms of the conditional pmf.

Theorem 13.2 *The following statements are equivalent for discrete random variables X and Y:*

(a) X, Y are independent.

(b) For all $x, y \in \mathbb{R}$, $\{X = x\}$ and $\{Y = y\}$ are independent.

(c) For all $x, y \in \mathbb{R}$, $\mathbb{P}_{X,Y}(x,y) = \mathbb{P}_X(x)\mathbb{P}_Y(y)$.

(d) For all $x, y \in \mathbb{R}$ such that $p_Y(y) > 0, p_{X|Y}(x|y) = p_X(x)$.

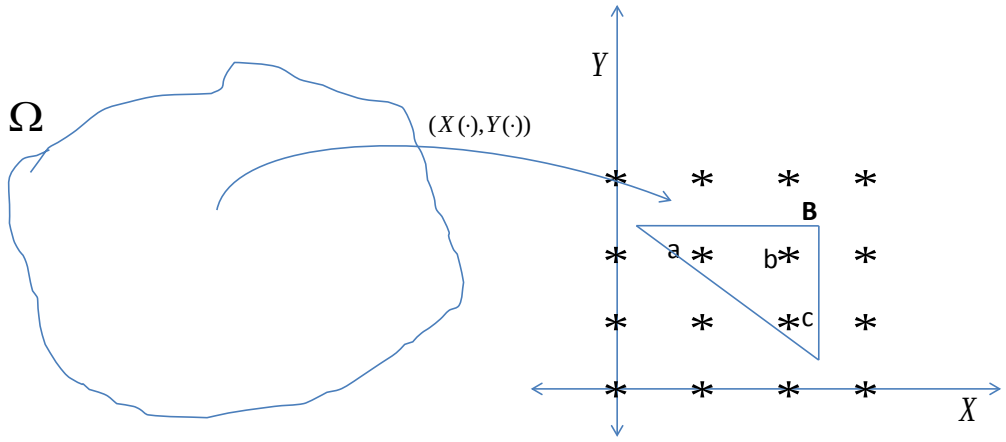

Figure 13.1: An example of two discrete random variables. The probability measure assigned to any Borel set B on $\mathbb{R}2$ can be obtained by summing the joint pmf over the set B; see (13.1).

Proof: $(b) \Leftrightarrow (c)$ and $(c) \Leftrightarrow (d)$ are directly follow from the definitions. Now, we prove equivalence of (a) and (c).
$(a) \Rightarrow (c)$:
X and Y are independent $\Rightarrow \mathbb{P}(X \in B_1, Y \in B_2) = \mathbb{P}(X \in B_1)\,\mathbb{P}(Y \in B_2)$. Take $B_1 = \{x\}$ and $B_2 = \{y\}$ then the result follows.
$(c) \Rightarrow (a)$:

$$\mathbb{P}(X \in B_1, Y \in B_2) = \sum_{x \in B_1, y \in B_2} p_{X,Y}(x,y) = \sum_{x \in B_1} \sum_{y \in B_2} p_X(x) p_Y(y) = \sum_{x \in B_1} p_X(x) \sum_{y \in B_2} p_Y(y)$$
$$= \mathbb{P}(X \in B_1)\,\mathbb{P}(Y \in B_2)\,.$$

■

13.2 Jointly Continuous Distributions

Definition 13.3 *Two Random variables X and Y are said to be* jointly continuous*, if the joint probability law $\mathbb{P}_{X,Y}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$. That is, for every Borel set $N \subset \mathbb{R}^2$ of Lebesgue measure zero, we have $\mathbb{P}(\{(X,Y) \in N\}) = 0$.*

The Radon-Nikodym theorem for this situation would assert the following

Theorem 13.4 *X, Y are jointly continuous random variables if and only if there exists a measurable function*

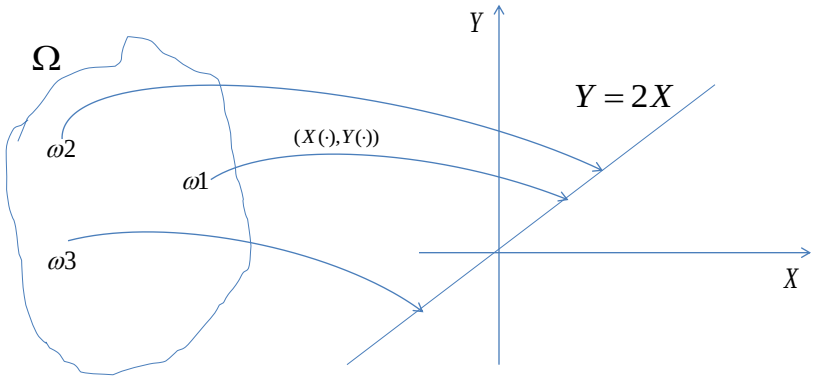

Figure 13.2: $Y = 2X$

$f_{X,Y} : \mathbb{R}^2 \to [0, \infty)$ *such that for any Borel set* B *on* $\mathbb{R}^2$,

$$\mathbb{P}_{X,Y}(B) = \int_B f_{X,Y} d\lambda,$$

where λ *is Lebesgue measure on* $\mathbb{R}^2$.

In particular, taking $B = (-\infty, x] \times (-\infty, y]$, we have

$$F_{X,Y}(x, y) = \mathbb{P}(\{X \leq x, Y \leq y\}) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) dv du, \tag{13.2}$$

where $F_{X,Y}(X, Y)$ and $f_{X,Y}(x, y)$ are the joint cdf and joint pdf respectively. The joint pdf is thus a complete characterization of the joint law, for jointly continuous random variables.

Caution: If X is continuous and Y is continuous, (X, Y) need not be jointly continuous. This can be seen from the following example.

Example: Let $X \sim \mathcal{N}(0, 1)$ and $Y = 2X$. i.e. $Y \sim \mathcal{N}(0, 4)$. In this case, though X is continuous and Y is continuous, (X, Y) are not jointly continuous.
This can be understood from the Figure 13.2. Each $\omega \in \Omega$ is mapped on to the straight line $Y = 2X$ on $\mathbb{R}^2$. The Lebesgue measure of the line (set), $L = \{(x, y) \in \mathbb{R}^2 : y = 2x\}$ is zero, but the corresponding probability, $\mathbb{P}_{X,Y}(L) = 1$, since every $\omega \in \Omega$ is mapped to this straight line. Thus, from the definition of jointly continuous random variables, X and Y are not jointly continuous.

On the other hand, if X and Y are jointly continuous, their marginals are necessarily continuous. To see this, note that

$$\mathbb{P}(\{X \leq x, Y \leq y\}) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)dvdu$$

$$\Rightarrow \mathbb{P}(X \leq x) = \int_{-\infty}^{x}\left(\int_{-\infty}^{\infty} f_{X,Y}(u,v)dv\right)du = \int_{-\infty}^{x} f_X(u)du. \tag{13.3}$$

In (13.3), it is clear that the inner integral in the parentheses produces a non-negative measurable function of u. Thus, (13.3) asserts that the marginal CDF of X can be written as the integral of a non-negative measurable function, which can be identified as the marginal pdf f_X. Thus, X is continuous and f_X given by the inner integral in (13.3) is the pdf of X. A similar argument holds for the marginal pdf of Y.

13.3 Independence of Jointly Continous Random Variables

For any two random variables X and Y, they are said to be independent iff,

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \forall x,y \in \mathbb{R}$$

Applying this definition, for the particular case of jointly continuous random variables,

$$\begin{aligned}\int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)dvdu &= \left(\int_{-\infty}^{x} f_X(u)du\right)\left(\int_{-\infty}^{y} f_Y(v)dv\right)\\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(u)f_Y(v)dvdu.\end{aligned}$$

Since the above equality holds for all $x, y \in \mathbb{R}$, the integrands must be equal almost every where, i.e.,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \forall\ x,y \in \mathbb{R}$$

except possibly on a subset of $\mathbb{R}^2$ of Lebesgue measure zero. Indeed, the above condition can be seen to be both necessary and sufficient for the independence of two jointly continuous random variables.

13.4 Conditional pdf for jointly continuous random variables

We would like to define the conditional cdf, $F_{X|Y}(x|y) \approx \mathbb{P}(X \leq x|Y = y)$. But the event, $\{Y = y\}$ has zero probability $\forall y$, when Y is continuous! To overcome this technical difficulty,we proceed by conditioning on a Y taking values in a small interval $(y, y+\epsilon)$, and then take the limit $\epsilon \downarrow 0$. More concretely, let us consider the following derivation, which motivates the definition of the conditional pdf for the jointly continuous case.

Informal Motivation
We can approximately define the conditional CDF of X, given that Y takes a value "close to y" as

$$\begin{aligned} F_{X|Y}(x|y) &= \mathbb{P}\left(X \le x \mid y \le Y \le y+\epsilon\right) \quad \text{(for small } \epsilon) \\ &= \frac{\mathbb{P}\left(\{X \le x\} \cap \{y \le Y \le y+\epsilon\}\right)}{\mathbb{P}\left(\{y \le Y \le y+\epsilon\}\right)} \\ &= \frac{F_{X,Y}(x, y+\epsilon) - F_{X,Y}(x,y)}{F_Y(y+\epsilon) - F_Y(y)} \\ &= \frac{\frac{F_{X,Y}(x,y+\epsilon) - F_{X,Y}(x,y)}{\epsilon}}{\frac{F_Y(y+\epsilon)-F_Y(y)}{\epsilon}}. \end{aligned}$$

As $\epsilon \to 0$, the RHS looks like, $\frac{\text{partial derivative of } F_{X,Y} \text{ w.r.t } y}{\text{derivate of } F_Y \text{ w.r.t } y}$.
This motivates the following definition for the conditional cdf and conditional pdf.

Definition 13.5 *a) The Conditional cdf of X given Y is defined as follows:*

$$F_{X|Y}(x|y) = \int_{-\infty}^{x} \frac{f_{X,Y}(u,y)}{f_Y(y)} du.$$

b) The Conditional pdf of X given Y is defined as follows:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{for any } y \text{ such that } f_Y(y) > 0.$$

c) The Conditional probability of an event $A \in \mathcal{B}(\mathbb{R})$ given $Y = y$ is defined by

$$\begin{aligned} \mathbb{P}(X \in A \mid Y = y) &= \int_A f_{X|Y}(v|y) dv \\ &= \int_{-\infty}^{\infty} \mathbb{I}_A(v) f_{X|Y}(v|y) dv, \end{aligned}$$

where $\mathbb{I}_A(x)$ is indicator function for the event $\{x \in A\}$.

Example:
Let X, Y be jointly continuous with $f_{X,Y}(x,y) = 1$ in the region shown in Figure 13.3. Find all the marginals and conditional distributions.

One can easily verify that $f_{X,Y}(x,y)$ is a valid joint pdf by integrating it over the region.

$$\text{i.e. } \int_0^2 \int_0^1 f_{X,Y}(x,y) dx dy = 1.$$

The marginal pdf of Y can be calculated as follows. In the Figure 13.3, the equation of the line is $y = -2x+2$. So, for a given y, $f_{X,Y}(x,y)$ is non zero only in the range $x \in (0, 1-\frac{y}{2})$. This can be seen from the Figure 13.3.

Thus,

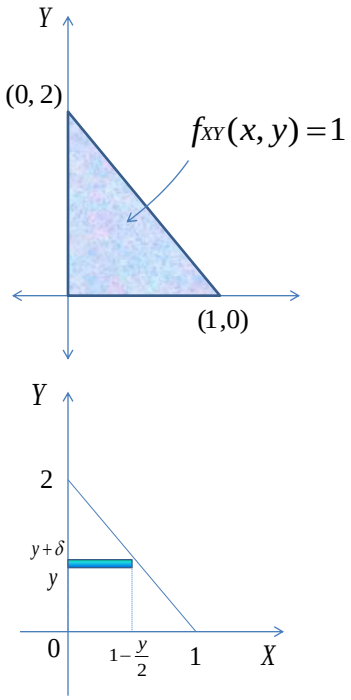

Figure 13.3:

$$f_Y(y) = \int_0^{1-\frac{y}{2}} 1dx = 1 - \tfrac{y}{2} \quad 0 \le y \le 2 \ .$$

Similarly, the marginal pdf of X can be calculated as follows. From the line equation, $y = -2x + 2$, we can find that for a given x, $f_{X,Y}(x, y)$ is non zero only in the range $y \in (0, 2 - 2x)$. Thus,

$$f_X(x) = \int_0^{2-2x} 1dy = 2 - 2x \quad 0 \le x \le 1.$$

The marginals of Y and X have been plotted in Figure 13.4.

Now, lets find the conditional pdf, $f_{X|Y}(x|y)$. Here, we are computing a function of x for a given(fixed) y.

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{1}{1-\frac{y}{2}} \\ &= \frac{2}{2-y} \qquad x \in \left(0, 1 - \frac{y}{2}\right). \end{aligned}$$

Though x does not appear in the expression, it appears in the constraint. So this is a conditionally uniform distribution in x i.e., given $\{Y = y\}$, X is uniformly distributed in $x \in \left(0, 1 - \frac{y}{2}\right)$ as $f_{X|Y}(x|y)$ is constant w.r.t x for the specified range.
Similarly, we can find,

$$f_{Y|X}(y|x) = \frac{1}{2-2x} \qquad y \in (0, 2 - 2x) \, .$$

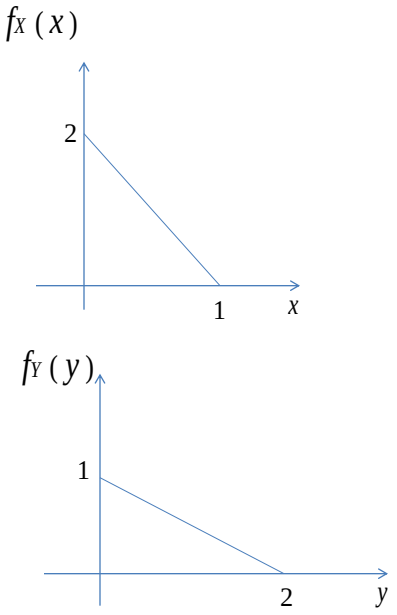

Figure 13.4: Marginal pdf of X and Y

13.5 Exercises

1. Two persons X and Y live in cities A and B but work in cities B and A respectively. Every morning they start for work at a uniformly random time between 9 am and 10 am independent of each other. Both of them travel at the same constant speed and it takes 20 minutes to reach the other city. What is the probability that X and Y meet each other on their way to work ?

2. Data is taken on the height and shoe size of a sample of MIT students. *Height* (X) is coded by 3 values: 1 (short), 2 (average), 3 (tall) and *Shoe size* (Y) is coded by 3 values 1 (small), 2 (average), 3 (large). The joint counts are given in the following table:

	X=1	X=2	X =3
Y=1	234	225	84
Y=2	180	453	161
Y=3	39	192	157

   (a) Find the joint and marginal pmf of X and Y.

   (b) Are X and Y independent? Discuss in detail.

3. John is vacationing in Monte Carlo. Each evening, the amount of money he takes to the casino is a random variable X with the pdf

$$f_X(x) = \begin{cases} Cx & 0 < x \leq 100 \\ 0 & \text{elsewhere.} \end{cases}$$

   At the end of each night, the amount Y he returns with is uniformly distributed between zero and twice the amount he came to casino with.

(a) Find the value of C.

(b) For a fixed $\alpha, 0 \leq \alpha \leq 100$, what is the conditional pdf of Y given $X = \alpha$?

(c) If John goes to the casino with α dollars, what is the probability he returns with more than α dollars ?

(d) Determine the joint pdf, $f_{X,Y}(x, y)$, of X and Y as well as the marginal pdf, $f_Y(y)$, of Y.

4. A rod is broken at two points that are chosen uniformly and independently at random. What is the probability that the three resulting pieces form a triangle?

5. *[https : //engineering.purdue.edu/ ipollak/ece302/.../problems/problems_4]* Melvin Fooch, a student of probability theory, has found that the hours he spends working (W) and sleeping (S) in preparation for a final exam are random variables described by:

$$f_{W,S}(w, s) = \begin{cases} K, & 10 \leq w + s \leq 20 \text{ and } w \geq 0, s \geq 0 \\ 0, & elsewhere. \end{cases}$$

What poor Melvin does not know, and even his best friends will not tell him, is that working only furthers his confusion and that his grade, G, can be described by $G = 2.5(S - W) + 50$.

(a) The instructor has decided to pass Melvin if, on the exam, he achieves $G \geq 75$. What is the probability that this will occur?

(b) Suppose Melvin got a grade greater than or equal to 75 on the exam. Determine the conditional probability that he spent less than one hour working in preparation for this exam.

(c) Are the random variables W and S independent? Justify.

6. *[https : //engineering.purdue.edu/ ipollak/ece302/.../problems/problems_4]* Stations A and B are connected by two parallel message channels. A message from A to B is sent over both channels at the same time. Continuous random variables X and Y represent the message delays (in hours) over parallel channels I and II, respectively. These two random variables are independent, and both are uniformly distributed from 0 to 1 hours. A message is considered received as soon as it arrives on any one channel, and it is considered verified as soon as it has arrived over both channels.

(a) Determine the probability that a message is received within 15 minutes after it is sent.

(b) Determine the probability that the message is received but not verified within 15 minutes after it is sent.

(c) If the attendant at B goes home 15 minutes after the message is received, what is the probability that he is present when the message should be verified?

7. *[https : //engineering.purdue.edu/ ipollak/ece302/.../problems/problems_4]* Random variables B and C are jointly uniform over a $2l \times 2l$ square centered at the origin, i.e., B and C have the following joint probability density function:

$$f_{B,C}(b, c) = \begin{cases} \frac{1}{4l^2}, & -l \leq b \leq l; -l \leq c \leq l \\ 0, & elsewhere. \end{cases}$$

It is given that $l \geq 1$. Find the probability that the quadratic equation $x^2 + 2Bx + C = 0$ has real roots (Answer will be an expression involving l). What is the limit of this probability as $l \to \infty$?

8. (a) *[Dimitri P.Bertsekas]* Consider four independent rolls of a 6-sided die. Let X be the number of 1's and let Y be the number of 2's obtained. What is the joint PMF of X and Y?

(b) *[MIT OCW problem set]* Let X_1, X_2, X_3 be independent random variables, uniformly distributed on $[0, 1]$. Let Y be the median of X_1 , X_2, X_3 (that is the middle of the three values). Find the conditional CDF of X_1, given that $Y = 0.5$. Under this conditional distribution, is X_1 continuous? Discrete?

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 14: Introduction to Transformation of Random Variables

Lecturer: Dr. Krishna Jagannathan *Scribe: Arjun Nadh*

Suppose we are able to observe a random variable or a collection of random variables. In many practical situations, we may be more interested in some *function* of the observed random variable(s). For example, in communication systems, the logarithm of the noise power is often more useful to an engineer than the noise realisation itself.

Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $f : \mathbb{R} \to \mathbb{R}$ be a function. We are interested in characterising the properties of $f(X)$. Since random variable X is itself a function, $f(X)$ is a composed function that maps Ω to $\mathbb{R}$. First, we have to ask if $f(X)$ is indeed a legitimate random variable. Consider the composed function $f \circ X(\cdot)$, depicted in Figure . If f is an arbitrary function, $f(X)$ may not be a random variable. However, if $f : \mathbb{R} \to \mathbb{R}$ is a Borel-measurable function (i.e., pre-images of Borel sets under f are also Borel sets), then it is clear that the pre-images of Borel sets under the composed function $f \circ X(\cdot)$ are events (why?), and it follows that $f(X)$ is indeed a random variable. Similarly, for a Borel-measurable function $f : \mathbb{R}^n \to \mathbb{R}$, and random variables $X_1, X_2, X_3, \ldots, X_n$, it can be argued that $f(X_1, \ldots, X_n)$ is a random variable.

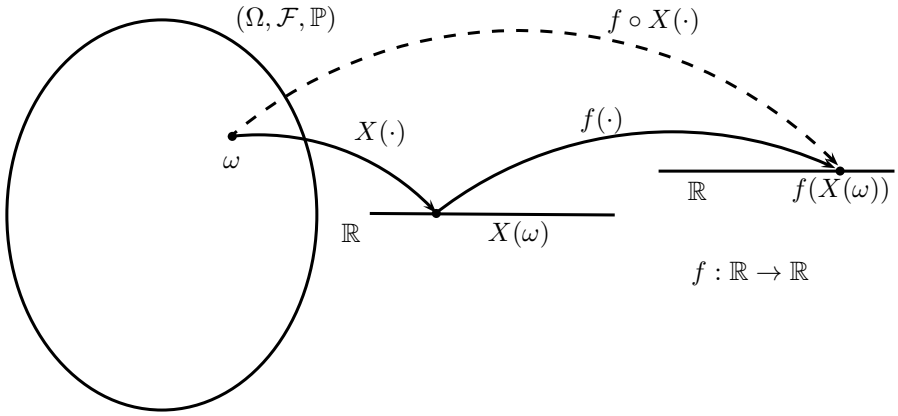

Figure 14.1: Transformation of random variable

Now that we have established conditions under which a function of a random variable is a random variable, we ask after the probability law of $f(X)$, given the probability law $\mathbb{P}_X$ of X. Equivalently, given the CDF of X, we want to find the CDF of $f(X)$. We begin by considering some elementary functions such as maximum, minimum, and summations, and then proceed to more general transformations.

14.1 Maximum and Minimum

Let $X_1, X_2, X_3, \ldots, X_n$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with joint CDF $F_{X_1, X_2, \ldots, X_n}$. Define

$$Y_n = \min(X_1, X_2, X_3, \ldots, X_n).$$

and

$$Z_n = \max(X_1, X_2, X_3, \ldots, X_n).$$

Here we are interested in finding the CDF of Y_n and Z_n.

First let us check that Z_n is indeed a random variable. Note that $\{Z_n \leq x\}$ is equivalent to saying that each of $X_1, X_2, X_3, \ldots, X_n$ is less than or equal to x. Thus, we have,

$$\{Z_n \leq x\} = \{X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x\}.$$

Now, in order to see that $\{\omega : Z(\omega) \leq z\}$ is an event, note that $\{\omega : Z(\omega) \leq z\} = \bigcap_{i=1}^{n} \{\omega : X_i(\omega \leq z)\}$. This is a finite intersection of events, since the X_is are random variables. Therefore, Z_n is a legitimate random variable.

Next, for the minimum, note that the if $\{Y_n > x\}$ is equivalent to saying that each of $X_1, X_2, X_3, \ldots, X_n$ is greater than x. Thus,

$$\{Y_n > x\} = \{X_1 > x, X_2 > x, \ldots, X_n > x\}.$$

We can prove that Y_n is also a random variable, by using arguments similar to those used for proving that Z_n is a random variable.

We now proceed to compute the CDF of random variables, Z_n and Y_n.

$$\begin{aligned}\mathbb{P}(\{Z_n \leq x\}) &= \mathbb{P}(\{X_1 \leq x\} \cap \{X_2 \leq x\} \cdots \cap \{X_n \leq x\}),\\ &= F_{X_1, X_2, \ldots, X_n}(x, x, \ldots, x).\end{aligned}$$

Similarly for Y_n,

$$\begin{aligned}\mathbb{P}(\{Y_n > x\}) &= \mathbb{P}(\{X_1 > x\} \cap \{X_2 > x\} \cdots \cap \{X_n > x\}\}),\\ \bar{F}_{Y_n}(x) &= 1 - F_{Y_n}(x),\\ &= \bar{F}_{X_1, X_2 \ldots . X_n}(x, x, \ldots, x),\end{aligned}$$

where $\bar{F}_{X_1 X_2, \ldots, X_n}(\cdot)$ denotes the joint complementary CDF.

In particular if $X_1, X_2, \ldots, X_n$ are independent

$$\begin{aligned}F_{Z_n}(x) &= F_{X_1}(x) F_{X_2}(x) \ldots F_{X_n}(x).\\ \bar{F}_{Y_n}(x) &= \bar{F}_{X_1}(x) \bar{F}_{X_2}(x) \ldots \bar{F}_{X_n}(x).\end{aligned}$$

Further if they are i.i.d (independent and identically distributed), then

$$\begin{aligned}F_{Z_n}(x) &= [F_X(x)]^n.\\ \bar{F}_{Y_n}(x) &= [\bar{F}_X(x)]^n.\end{aligned}$$

Example 1:- Consider U_1, U_2 to be i.i.d, Unif$[0, 1]$,
Let Y $= \min(U_1, U_2)$ and $Z = \max(U_1, U_2)$.
Let $F_{U_1}(z)$ and $F_{U_2}(z)$ be CDF's of random variables U_1 and U_2 respectively. Since they are identically distributed

$$F_{U_1}(z) = F_{U_2}(z) = [F_U(z)],$$

where

$$F_U(z) = \begin{cases} 0 & z < 0, \\ z & z \in [0,1], \\ 1 & z > 1. \end{cases}$$

Since U_1 and U_2 are also independent

$$F_Z(z) = F_{U_1}(z)F_{U_2}(z) = [F_U(z)]^2.$$

$$[F_U(z)]^2 = \begin{cases} 0 & z < 0, \\ z^2 & z \in [0,1], \\ 1 & z > 1. \end{cases}$$

Its pdf is given by

$$f_z(z) = \begin{cases} 2z & z \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

Similarly for $F_Y(y)$ we can write

$$\bar{F}_Y(y) = \bar{F}_{U_1}(z)\bar{F}_{U_2}(z) = [\bar{F}_U(y)]^2,$$

where $\bar{F}_Y(y)$ denotes the complementary CDF of Y.

$$F_Y(y) = 1 - [\bar{F}_U(y)]^2.$$

$$F_Y(y) = \begin{cases} 0 & y < 0, \\ 1 - (1-y)^2 & y \in [0,1], \\ 1 & y > 1. \end{cases}$$

The pdf is given by

$$f_y(y) = \begin{cases} 0 & y < 0, \\ 2(1-y) & y \in [0,1], \\ 1 & y > 1. \end{cases}$$

Example 2:-

Let $X_1, X_2, X_3, \ldots, X_n$ be independent random variables which are exponentially distributed with the parameters λ_1, λ_2, $\lambda_3, \ldots, \lambda_n > 0$. $F_{X_i}(x) = 1 - e^{-\lambda_i x}$ for $x > 0$.

Let

$$Y_n = \min(X_1, X_2, ...X_n).$$

Then the complementary CDF of Y_n :

$$\begin{aligned}\bar{F}_{Y_n}(y) &= \prod_{i=1}^{n} \bar{F}_{X_i}(y), \\ &= \prod_{i=1}^{n} e^{-\lambda_i y}, \\ &= e^{(-\sum_{i=1}^{n} \lambda_i) y}.\end{aligned}$$

We can see that Y_n is an exponential random variable with parameter $\lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_n$. Thus, the minimum of independent exponential random variables is another exponential random variable!

14.2 Exercises

1. **Light bulbs with Amnesia:** Suppose that n light bulbs in a room are switched on at the same instant. The life time of each bulb is exponentially distributed with parameter $\mu = 1$, and are independent.
   (a) Starting from the time they are switched on, find the distribution of the time when the first bulb fuses out.
   (b) Find the CDF and the density of the time when the room goes completely dark.
   (c) Would your answers to the above parts change if the bulbs were not switched on at the same time, but instead, turned on at arbitrary times? Assume however that all bulbs were turned on before the first one fused out.
   (d) Suppose you walk into the room and find m bulbs glowing. Starting from the instant of your walking in, what is the distribution of the time it takes until you see a bulb blow out?
2. Let X and Y be independent exponentially distributed random variables with parameters λ and μ respectively.
   (a) Show that $Z = \min(X, Y)$ is independent of the event $\{X < Y\}$, and interpret this result verbally? [Definition: A random variable X is said to be independent of an event A if X and $\mathbb{I}_A$ are independent random variables, where $\mathbb{I}_A$ is the Indicator random variable of the event A.]
   (b) Find $\mathbb{P}(X = Z)$.

EE5110 : Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 15: Sums of Random Variables

Lecturer: Dr. Krishna Jagannathan *Scribes: R.Ravi Kiran*

15.1 Sum of Two Random Variables

In this section, we will study the distribution of the sum of two random variables. Before we discuss their distributions, we will first need to establish that the sum of two random variables is indeed a random variable.

Theorem 15.1 *Let X and Y be random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and define $Z(\omega) = X(\omega) + Y(\omega),\ \forall \omega \in \Omega$. Then, Z is a random variable.*

Proof: To prove that Z is a random variable, we need to show that $\{\omega \in \Omega : Z(\omega) > z\} \in \mathcal{F},\ \forall z \in \mathbb{R}$.

Now, $\forall z \in \mathbb{R}$, $Z(\omega) > z$ if and only if there exists a rational q such that $X(\omega) > q$ and $Y(\omega) > z - q$. This follows from the fact that the set of rationals is dense in $\mathbb{R}$. Thus,

$$\begin{aligned} \{\omega \in \Omega : Z(\omega) > z\} &= \bigcup_{q \in \mathbb{Q}} \{\omega \in \Omega : X(\omega) > q, Y(\omega) > z - q\} \\ &= \bigcup_{q \in \mathbb{Q}} \left(\{\omega \in \Omega : X(\omega) > q\} \cap \{\omega \in \Omega : Y(\omega) > z - q\}\right). \end{aligned} \tag{15.1}$$

We know that $\forall q \in \mathbb{Q}$, $\{\omega \in \Omega : X(\omega) > q\} \cap \{\omega \in \Omega : Y(\omega) > z - q\} \in \mathcal{F}$ because X and Y are random variables. Since the set of rationals is countable, we have a countable union of sets from $\mathcal{F}$, which should also be in $\mathcal{F}$ as it is a σ-algebra. Thus, $\{\omega \in \Omega : Z(\omega) > z\} \in \mathcal{F}$, proving that the sum, $Z = X + Y$ is a random variable. ■

We will now start with random variables in the discrete domain. Assume that X and Y are discrete random variables with a known joint pmf $p_{X,Y}(\cdot)$. Let the random variable Z be defined as $Z = X + Y$. We will now characterize the pmf of Z, $p_Z(\cdot)$:

$$\begin{aligned} p_Z(z) &= \mathbb{P}(Z = z) \\ &= \sum_{x+y=z} p_{X,Y}(x, y) \\ &= \sum_{x} \mathbb{P}(X = x, Y = z - x) \\ &= \sum_{x} p_{X,Y}(x, z - x) \end{aligned} \tag{15.2}$$

In particular, if X and Y are independent, the pmf of Z simplifies to

$$p_Z(z) = \sum_{x} p_X(x) p_Y(z - x), \tag{15.3}$$

which is simply the discrete convolution of the two pmfs.

Let us now look at an example.

Example 15.2 *Let X and Y be independent, random variables with distributions given by $Pois(\lambda)$ and $Pois(\mu)$ respectively. Define $Z = X + Y$. Then, the pmf of Z, can be computed, by invoking (15.3) :*

$$\begin{aligned} p_Z(z) &= \sum_{x=0}^{z} \frac{e^{-\lambda}\lambda^x}{x!}\frac{e^{-\mu}\mu^{z-x}}{(z-x)!} \\ &= \frac{e^{-(\lambda+\mu)}}{z!}\sum_{x=0}^{z}\binom{z}{x}\lambda^x\mu^{z-x} \\ &= \frac{e^{-(\lambda+\mu)}(\lambda+\mu)^z}{z!} \end{aligned}$$

The above computation establishes that the sum of two independent Poisson distributed random variables, with mean values λ and μ, also has Poisson distribution of mean $\lambda + \mu$.

We can easily extend the same derivation to the case of a finite sum of independent Poisson distributed random variables.

Next, we consider the case of two jointly continuous random variables. Assume that X and Y are jointly continuous random variables, with joint pdf given by $f_{X,Y}(x, y)$. Let $Z = X + Y$. Then,

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \le z) \\ &= \mathbb{P}(X + Y \le z) \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z-x} f_{X,Y}(x,y)dy\right)dx \qquad (15.4) \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z} f_{X,Y}(x,t-x)dt\right)dx \\ &= \int_{-\infty}^{z}\underbrace{\left(\int_{-\infty}^{\infty} f_{X,Y}(x,t-x)dx\right)}_{f_Z(t)}dt. \qquad (15.5) \end{aligned}$$

From (15.5), we can see that the pdf of Z is given by $f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)dx$.

In the special case of X and Y being independent continuous random variables, we get

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx = f_X * f_Y, \qquad (15.6)$$

which is the convolution of the two marginal pdfs.

Example 15.3 *Assume that X_1 and X_2 are independent exponential random variables with parameters μ_1 and μ_2 respectively. Let $Z = X_1 + X_2$. Using (15.6) and the fact that the support for the exponential random variable is $\mathbb{R}^+ \cup \{0\}$, we get,*

$$\begin{aligned} f_Z(z) &= f_{X_1} * f_{X_2}, \\ &= \int_0^z \mu_1 e^{-\mu_1 x}\mu_2 e^{-\mu_2(z-x)}dx, \\ &= \mu_1\mu_2 e^{-\mu_2 z}\int_0^z e^{(\mu_2-\mu_1)x}dx. \end{aligned}$$

We can see from the above integral that

$$f_Z(z) = \begin{cases} \frac{\mu_1\mu_2}{\mu_2-\mu_1}\left(e^{-\mu_1 z} - e^{-\mu_2 z}\right) & \text{if } \mu_1 \neq \mu_2, \\ \mu^2 z e^{-\mu z} & \mu_1 = \mu_2 = \mu. \end{cases}$$

In fact, the process can be extended to the case of a sum of a finite number n of random variables of distribution exp(μ), and we can observe that the pdf of the sum, Z_n, is given by Erlang (n, μ), i.e,

$$f_{Z_n}(z) = \frac{\mu^n z^{n-1} e^{-\mu z}}{(n-1)!}. \tag{15.7}$$

The above example describes the process of computing the pdf of a sum of continuous random variables.

The methods described above can be easily extended to deal with finite sums of random variables too.

15.2 Sum of a random number of random variables

In this section, we consider a sum of independent random variables, where the number of terms in the summation is itself random. Let N be a positive integer valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with known pmf $\mathbb{P}(N = n)$. Let $X_1, X_2, ...$, be independent random variables on the same probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, with distributions, $F_{X_1}(.), F_{X_2}(.), ...$, respectively. Further, we will assume that N is independent of $\{X_i, i \geq 1\}$.

Define, $S_N = \sum_{i=1}^{N} X_i$. That is, $S_N(\omega) = \sum_{i=1}^{N(\omega)} X_i(\omega), \forall \omega \in \Omega$. The cdf of S_N can be computed as follows :

$$\begin{aligned} F_{S_N}(x) &= \mathbb{P}(S_N \leq x), \\ &= \sum_{k=1}^{\infty} \mathbb{P}(S_N \leq x | N = k)\mathbb{P}(N = k), \\ &= \sum_{k=1}^{\infty} \mathbb{P}(S_k \leq x)\mathbb{P}(N = k), \end{aligned} \tag{15.8}$$

where (15.8) follows from the independence of N and the X_is.

In the above expression, we know how to compute $\mathbb{P}(S_k \leq x)$ from the previous section. Thus we have essentially computed the distribution of the random sum of random variables under the specified independence assumptions.

The following example is quite instructive.

Example 15.4 *Geometric Sum of Exponentials :*
Let $X_i, \forall i \geq 1$ be independent random variables with distribution exp(μ). Let N be a positive integer valued random variable of geometric distribution with parameter p.

Define $S_N = \sum_{i=1}^{N} X_i$. We will now determine the pdf of S_N.

We know that $\mathbb{P}(N = k) = (1-p)^{k-1}p, \forall k \geq 1$. Further we observed earlier (15.7) that the sum of k exponential distributions of mean $\frac{1}{\mu}$, $S_k = \sum_{i=1}^{k} X_i$, is a k^{th} order Erlang distribution. Thus, using this and

(15.8), we get,

$$\begin{aligned}
F_S(x) &= \mathbb{P}(S_N \leq x), \\
&= \sum_{k=1}^{\infty} \mathbb{P}(N=k) F_{S_k}(x), \\
&= \sum_{k=1}^{\infty} \left(p(1-p)^{k-1}\right)\left(1 - \sum_{n=0}^{k-1} \frac{1}{n!} e^{-\mu x}(\mu x)^n\right), \\
&= \sum_{k=1}^{\infty} p(1-p)^{k-1} - \sum_{k=1}^{\infty} p(1-p)^{k-1} e^{-\mu x}\left(\sum_{n=0}^{k-1} \frac{1}{n!}(\mu x)^n\right), \\
&= 1 - e^{-\mu x} \sum_{n=0}^{\infty} \frac{(\mu x)^n}{n!} \frac{p}{1-p} \sum_{k=n+1}^{\infty} (1-p)^k, \\
&= 1 - e^{-\mu x} \sum_{n=0}^{\infty} \frac{(\mu x(1-p))^n}{n!}, \\
&= 1 - e^{-\mu x} e^{\mu(1-p)x}, \\
&= 1 - e^{-(p\mu)x}.
\end{aligned}$$

The above derivation establishes that the geometric sum of exponentials has an exponential distribution with parameter $\mu' = p\mu$.

Consider a radioactive source emitting α particles where the time between two successive emissions is exponentially distributed with parameter λ. Whenever there is an emission, the detector detects it with probability p and misses it with probability $1-p$ independent of other detections. So it can be easily seen that the time between two successive detections is indeed a geometric sum of i.i.d exponential random variables which itself is an exponential random variable with parameter $p\lambda$ as seen in the above example.

The above study gives a detailed account of the random sum of random variables under the strict independence constraints earlier assumed. It is however possible to envision a scenario where the random number N is dependent on the observations, X_i themselves.

For instance let us assume that a gambler plays a game repeatedly and is rewarded or penalized in each round. Say the gambler stops only when he is "satisfied" (or "broke") with the overall outcome of the game. Let X_i be the amount he gains (or loses) in round i of the game. In this scenario, analysing the overall sum earned by the gambler at the end of his game is complicated by the dependence of N on the outcomes. This scenario motivates the theory of *stopping rules*, which shall be covered in a more advanced course (EE6150).

15.3 Exercise:

1. Let X_1 and X_2 be independent random variables with distributions $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$ respectively. Show that the distribution of $X_1 + X_2$ is $\mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$.

2. Consider two independent and identically distributed discrete random variables X and Y . Assume that their common PMF, denoted by $p(z)$, is symmetric around zero, i.e., $p(z) = p(-z)$, $\forall z$. Show that the PMF of $X + Y$ is also symmetric around zero and is largest at zero.

3. Suppose X and Y are independent random variables with $Z = X + Y$ such that $f_X(x) = ce^{-cx}$, $x \geq 0$ and $f_Z(z) = c^2 z e^{-cz}$, $z \geq 0$. Compute $f_Y(y)$.

4. Let X_1 and X_2 be the number of calls arriving at a switching centre from two different localities at a given instant of time. X_1 and X_2 are well modelled as independent Poisson random variables with parameters λ_1 and λ_2 respectively.

   (a) Find the PMF of the total number of calls arriving at the switching centre.

   (b) Find the conditional PMF of X_1 given the total number of calls arriving at the switching centre is n.

5. The random variables X, Y and Z are independent and uniformly distributed between zero and one. Find the PDF of $X + Y + Z$.

6. Construct an example to show that the sum of a random number of independent normal random variables is not normal.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 16: General Transformations of Random Variables

Lecturer: Dr. Krishna Jagannathan *Scribe: Ajay and Jainam*

In the previous lectures, we have seen few elementary transformations such as sums of random variables as well as maximum and minimum of random variables. Now we will look at general transformations of random variables. The motivation behind transformation of a random variable is illustrated by the following example. Consider a situation where the velocity of a particle is distributed according to a random variable V. Based on a particular realisation of the velocity, there will be a corresponding value of kinetic energy E and we are interested in the distribution of kinetic energy. Clearly, this is a scenario where we are asking for the distribution of a new random variable, which depends on the original random variable through a transformation. Such situations occur often in practical applications.

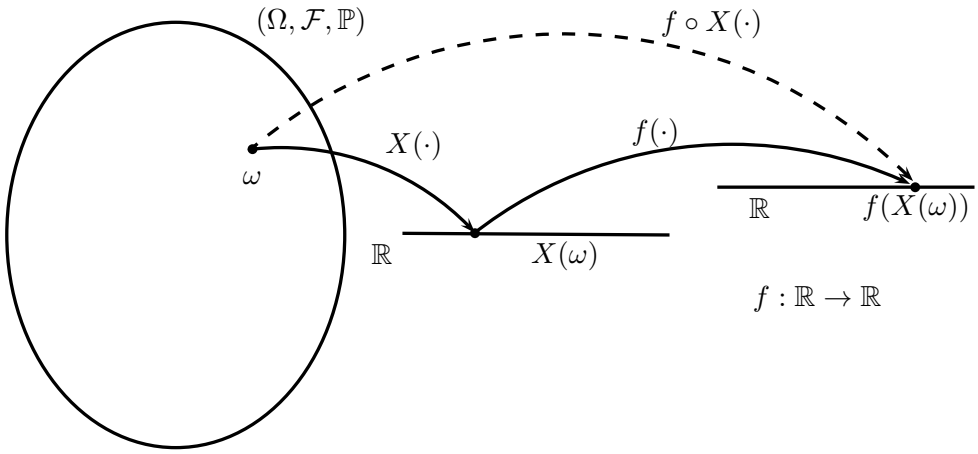

Figure 16.1: Transformation of random variable

16.1 Transformations of a Single Random Variable

Consider a random variable $X : \Omega \to \mathbb{R}$ and let $g : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Then $Y = g(X)$ is also a random variable and we wish to find the distribution of Y. Specifically, we are interested in finding the CDF $F_Y(y)$ given the CDF $F_X(x)$.

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(\{\omega | g(X(\omega)) \leq y\}).$$

Let B_y be the set of all x such that $g(x) \leq y$. Then $F_Y(y) = \mathbb{P}_X(B_y)$.

We now illustrate this with the help of an example.

Example 1: *Let X be a Gaussian random variable of mean 0 and variance 1* i. e. $X \sim \mathcal{N}(0, 1)$. *Find the distribution of $Y = X^2$.*

Solution:

$$F_Y(y) \;=\; \mathbb{P}(X^2 \le y) \;=\; \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}\, dt = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}),$$

where Φ *is the CDF of* $\mathcal{N}(0,1)$.
Now,

$$f_Y(y) = \frac{dF_Y(y)}{dy}.$$

From above,

$$F_Y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}\, dt = 2 \times \int_{0}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}\, dt.$$

Changing variables $t^2 = u$, *we get*

$$F_y(y) = 2 \times \int_0^y \frac{1}{2\sqrt{2\pi u}} e^{\frac{-u}{2}}\, du.$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{\frac{-y}{2}}, \quad \textit{for} \quad y > 0.$$

Note:

1. The random variable Y can take only non-negative values as it is square of a real valued random variable.
2. The distribution of square of the Gaussian random variable, $f_Y(y)$, is also known as Chi-squared distribution.

Thus, we see that given the distribution of a random variable X, the distribution of any function of X can be obtained by first principles. We now come up with a direct formula to find the distribution of a function of the random variable in the cases where the function is differentiable and monotonic.

Let X have a density $f_X(x)$ and g be a monotonically increasing function and let $Y = g(X)$. We then have

$$F_Y(y) = \mathbb{P}(Y \le y) \;=\; \mathbb{P}(g(X) \le y) \;=\; \int_{-\infty}^{g^{-1}(y)} f_X(x) dx.$$

Note that as g is a monotonically increasing function, $g(x) \le y \implies x \le g^{-1}(y)$.

Let $x = g^{-1}(t)$, so $g'(x)dx = dt$.

$$F_Y(y) = \int_{-\infty}^{y} f_X(g^{-1}(t)) \frac{dt}{g'(g^{-1}(t))}.$$

Differentiating, we get

$$f_Y(y) = f_X(g^{-1}(y))\frac{1}{g'(g^{-1}(y))}.$$

The second term on the right hand side of the above equation is referred to as the *Jacobian* of the transformation $g(\cdot)$.

It can be shown easily that a similar argument holds for a monotonically decreasing function g as well and we obtain

$$f_Y(y) = f_X(g^{-1}(y))\frac{-1}{g'(g^{-1}(y))}.$$

Hence, the general formula for distribution of monotonic functions of random variables is as under

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{\mid g'(g^{-1}(y)) \mid}. \tag{16.1}$$

Example 2: . *Let $X \sim \mathcal{N}(0,1)$. Find the distribution of $Y = e^X$.*

Solution: *Note that the function $g(x) = e^x$ is a differentiable, monotonically increasing function.*

As $g(x) = e^x$, we have $g^{-1}(y) = \ln(y)$ and $g'(g^{-1}(y)) = y$. Here we see that the Jacobian will be positive for all vaues of y and hence $\mid g'(g^{-1}(y)) \mid = g'(g^{-1}(y)) = y$.

Finally we have

$$f_Y(y) = \frac{f_X(\ln(y))}{y} = \frac{1}{y\sqrt{2\pi}}e^{\frac{-(\ln(y))^2}{2}} \text{ for } \quad y > 0.$$

This is the log-normal pdf.

Example 3: *Let $U \sim unif[0,1]$ i.e. U is a uniform random variable in the interval [0,1]. Find the distribution $Y = -\ln(U)$.*

Solution: *Note that $g(u) = -\ln(u)$ is a differentiable, monotonically decreasing function.*

As $g(u) = -\ln(u)$, we have $g^{-1}(y) = e^{-y}$ and $g'(g^{-1}(y)) = \frac{-1}{e^{-y}}$. Here we see that the Jacobian will be negative for all values of y and hence $\mid g'(g^{-1}(y)) \mid = -g'(g^{-1}(y)) = \frac{1}{e^{-y}}$.

Hence we have

$$f_Y(y) = \frac{f_U(e^{-y})}{\frac{1}{e^{-y}}} = \frac{1}{\frac{1}{e^{-y}}} = e^{-y} \text{ for } y \geq 0.$$

Note that Y is an exponential random variable with mean 1.

If X is a continous random variable with CDF $F_X(\cdot)$, then it can be shown that the random variable $Y = F(X)$ is uniformly distributed over $[0,1]$ (*see Exercise 2(a)*). It can be seen from this result that any continuous random variable Y can be generated from a uniform random variable $X \sim unif[0,1]$ by the transformation $Z = F_Y^{-1}(X)$ where $F_Y(\cdot)$ is the CDF of the random variable Y.

16.2 Transformation of Multiple Random Variables

Equation 16.1 can be extended to transformations of multiple random variables. Consider an n-tuple random variable $(X_1, X_2, \ldots, X_n)$ whose joint density is given by $f_{(X_1,X_2,\ldots,X_n)}(x_1, x_2, \ldots, x_n)$ and the

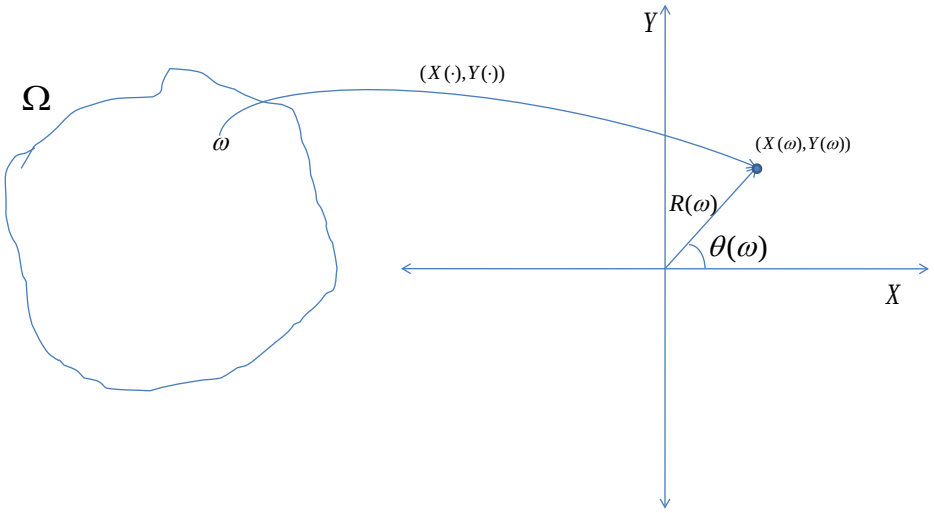

Figure 16.2: Mapping of a realization $(X(\omega), Y(\omega))$ to the polar co-ordinates $(R(\omega), \Theta(\omega))$

corresponding transformations are given by $Y_1 = g_1(X_1, X_2, \ldots, X_n), Y_2 = g_2(X_1, X_2, \ldots, X_n), \ldots, Y_n = g_n(X_1, X_2, \ldots, X_n)$. Succinctly, we denote this as a vector transformation $Y = g(X)$, where $g : \mathbb{R}^n \to \mathbb{R}^n$. We assume that the transformation g is invertible, and continuously differentiable. Under this assumption, the joint density of $f_{Y_1,Y_2,\ldots,Y_n}(Y_1, Y_2, \ldots, Y_n)$ is given by (see Section 2.2 in Lecture 10 of [1])

$$f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) = f_{(X_1,X_2,\ldots,X_n)}(g_1^{-1}(y))|J(y)|, \tag{16.2}$$

where $|J(y)|$ is the Jacobian matrix, given by

$$J(y) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \cdot & \cdot & \cdot & \frac{\partial x_n}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \cdot & \cdot & \cdot & \frac{\partial x_n}{\partial y_2} \\ \cdot & \cdot & & & & \cdot \\ & & & \ddots & & \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \frac{\partial x_1}{\partial y_n} & \frac{\partial x_2}{\partial y_n} & \cdot & \cdot & \cdot & \frac{\partial x_n}{\partial y_n} \end{vmatrix}.$$

We now illustrate this with the help of an example.

Example 4: *Let the Euclidean co-ordinates of a particle be drawn from identically distributed independent Gaussian random variables of mean* 0 *and varinace* 1 *i.e.,* $X, Y \sim \mathcal{N}(0, 1)$. *Find the distribution of the particle's polar co-ordinates,* R *and* Θ.

Solution: *The corresponding transformations are given by* $X = R\cos\Theta$ *and* $Y = R\sin\Theta$.

Let us first evaluate the Jacobian. We have $x = r\cos\theta$ *and* $y = r\sin\theta$*. So we have* $\frac{\partial x}{\partial r} = \cos\theta$*,* $\frac{\partial y}{\partial r} = \sin\theta$*,* $\frac{\partial x}{\partial \theta} = -r\sin\theta$ *and* $\frac{\partial y}{\partial \theta} = r\cos\theta$*.*

$$J = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial \theta} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & \sin\theta \\ r\sin\theta & r\cos\theta \end{vmatrix} = r\cos^2\theta + r\sin^2\theta = r(\cos^2\theta + \sin^2\theta) = r.$$

Next, we have $X, Y \sim \mathcal{N}(0,1)$ *and they are independent so*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}} \quad x, y \in \mathbb{R}.$$

From (16.2)*, we have*

$$f_{R,\Theta}(r,\theta) = \frac{1}{2\pi}e^{-\frac{(r\cos\theta)^2+(r\sin\theta)^2}{2}} \times r \text{ where } r \geq 0 \text{ and } \theta \in [0, 2\pi].$$

The marginal densities of R *and* Θ *can be obtained from the joint distribution as given below.*

$$f_R(r) = \int_0^{2\pi} f_{R,\Theta}(r,\theta)d\theta = re^{-r^2/2}, \text{ for } r \geq 0.$$

$$f_\Theta(\theta) = \int_0^{\infty} f_{R,\Theta}(r,\theta)dr = \frac{1}{2\pi}, \text{ for } \theta \in [0, 2\pi].$$

The distribution $f_R(r)$ *is called the Rayleigh distribution, which is encountered quite often in Wireless Communications to model the gain of a fading channel. Note that the random variables* R *and* Θ *are independent since the joint distribution factorizes into the product of the marginals i.e.*

$$f_{R,\Theta}(r,\theta) = f_R(r) \times f_\Theta(\theta).$$

We now illustrate how transformations of random variables help us to generate random variables with different distributions given that we can generate only uniform random variables. Specifically, consider the case where all we can generate is a uniform random variable between 0 and 1 *i. e.* $unif[0,1]$ and we wish to generate random variables having Rayleigh, exponential and Guassian distribution.

Generate U_1 and U_2 as i.i.d. $unif[0,1]$. Next, let $\Theta = 2\pi U_1$ and $Z = -\frac{\ln(U_2)}{2}$. It can be verified that $\Theta \sim Unif[0, 2\pi]$ and $Z \sim Exp(0.5)$.

Thereafter, let $R = \sqrt{Z}$. It can be shown that R is a Rayleigh distributed random variable (*see Exercise 1*).

Lastly, let $X = R\cos\Theta$ and $Y = R\sin\Theta$. It is easy to see from Example 3 that X and Y will be i.i.d. $\mathcal{N}(0,1)$.

16.3 Exercises

1. Let $X \sim exp(0.5)$. Prove that $Y = \sqrt{X}$ is a Rayleigh distributed random variable.
2. (a) Let X be a random variable with a continuous distribution F.

(i) Show that the Random Variable $Y = F(X)$ is uniformly distributed over $[0, 1]$. [Hint: Although F is the distribution of X, regard it simply as a function satisfying certain properties required to make it a CDF !]

(ii) Now, given that $Y = y$, a random variable Z is distributed as Geometric with parameter y. Find the unconditional PMF of Z. Also, given $Z = z$ for some $z \geq 1, z \in \mathbb{N}$ find the conditional PMF of Y.

(b) Let X be a continuous random variable with the pdf

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

Find the transformation Y=g(X) such that the pdf of Y will be

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} & 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

[Hint: Question 1(a) might be of use here !]

3. Suppose X and Y are independent Gaussian random variables with zero mean and variance σ^2. Show that $\frac{X}{Y}$ is cauchy.

4. (a) Particles are subject to collisions that cause them to split into two parts with each part a fraction of the parent. Suppose that this fraction is uniformly distributed between 0 and 1. Following a single particle through several splittings we obtain a fraction of the original particle $Z_n = X_1 X_2 ... X_n$ where each X_j is uniformly distributed between 0 and 1. Show that the density for the random variable Z_n is,

$$f_n(z) = \frac{1}{(n-1)!}(-log(z))^{n-1}$$

(b) Suppose X and Y are independent exponential random variables with same parameter λ. Derive the pdf of the random variable $Z = \frac{\min(X,Y)}{\max(X,Y)}$.

5. A random variable Y has the pdf $f_Y(y) = Ky^{-(b+1)}, y \geq 2$ (and zero otherwise), where $b > 0$. This random variable is obtained as the monotonically increasing transformation $Y = g(X)$ of the random variable X with pdf $e^{-x}, x \geq 0$.

(a) Determine K in terms of b.

(b) Determine the transformation $g(.)$ in terms of b.

6. (a) Two particles start from the same point on a two-dimensional plane, and move with speed V each, such that the angle between them is uniformly distributed in $[0, 2\pi]$. Find the distribution of the magnitude of the relative velocity between the two particles.

(b) A point is picked uniformly from inside a unit circle. What is the density of R, the distance of the point from the center?

7. Let X and Y be independent exponentially distributed random variables with parameter 1. Find the joint density of $U = X + Y$ and $V = \frac{X}{X+Y}$, and show that V is uniformly distributed.

References

[1] David Gamarnick and John Tsitsiklis, "Introduction to Probability", *MIT OCW*, , 2008.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 17: Integration and Expectation

Lecturer: Dr. Krishna Jagannathan *Scribe: Gopal Krishna Kamath M*

In this chapter, we introduce abstract integration, and in particular, define the integral of a measurable function, with respect to a measure. As a special case, the integral of a random variable with respect to a probability measure is known as the expectation of the random variable.

Our approach to defining the expectation of a random variable as an abstract integral serves to unify the definition. After all, you may recall from your undergraduate study of probability that the expectation of a random variable is defined via an integral if the random variable is continuous, and a summation if it is discrete. Of course, if the random variable were singular or a mixture, the elementary approach does not provide a simple definition of the expectation. On the other hand, the definition we are about to give is completely general; specifically, we do not have to provide separate definitions for different types of random variables.

In addition, the theory of abstract integration allows us to define the Lebesgue integral, which generalizes the notion of the Riemann integral from high-school calculus.

17.1 The Riemann Integral: A Review

Consider a function $f : \mathbb{R} \to \mathbb{R}$. Let $[a, b]$ be an interval in the domain of f, and $\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ be a partition of $[a, b]$. The lower and upper Riemann sums, denoted by L_n and U_n respectively, are defined as below:

$$\begin{aligned} L_n &\triangleq \sum_{i=1}^{n} \left(\inf_{x \in [x_i, x_{i+1}]} f(x) \right) \Delta x_i, \\ U_n &\triangleq \sum_{i=1}^{n} \left(\sup_{x \in [x_i, x_{i+1}]} f(x) \right) \Delta x_i. \end{aligned}$$

As n increases in a manner such that each Δx_i decreases to zero, it can be seen that L_n is monotone increasing, while U_n is monotone decreasing. So, as $n \to \infty$ it follows that L_n and U_n will both converge. The limits of L_n and U_n are called the lower and upper Riemann integrals, respectively. That is,

$$\begin{aligned} \int_a^b f(x)\, \mathrm{d}x &\triangleq \lim_{n \to \infty} L_n, \\ \overline{\int_a^b} f(x)\, \mathrm{d}x &\triangleq \lim_{n \to \infty} U_n. \end{aligned}$$

It can be shown that the values of the Lower and Upper Riemann Integrals do not depend on the choice of

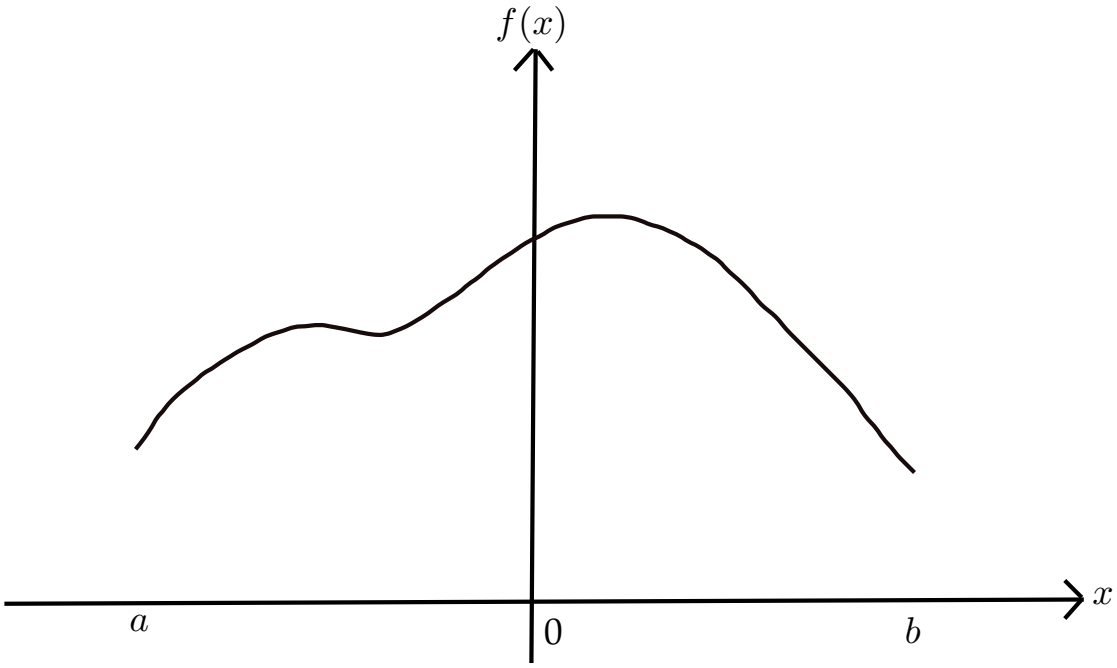

Figure 17.1: An arbitrary function f over the interval $[a, b]$.

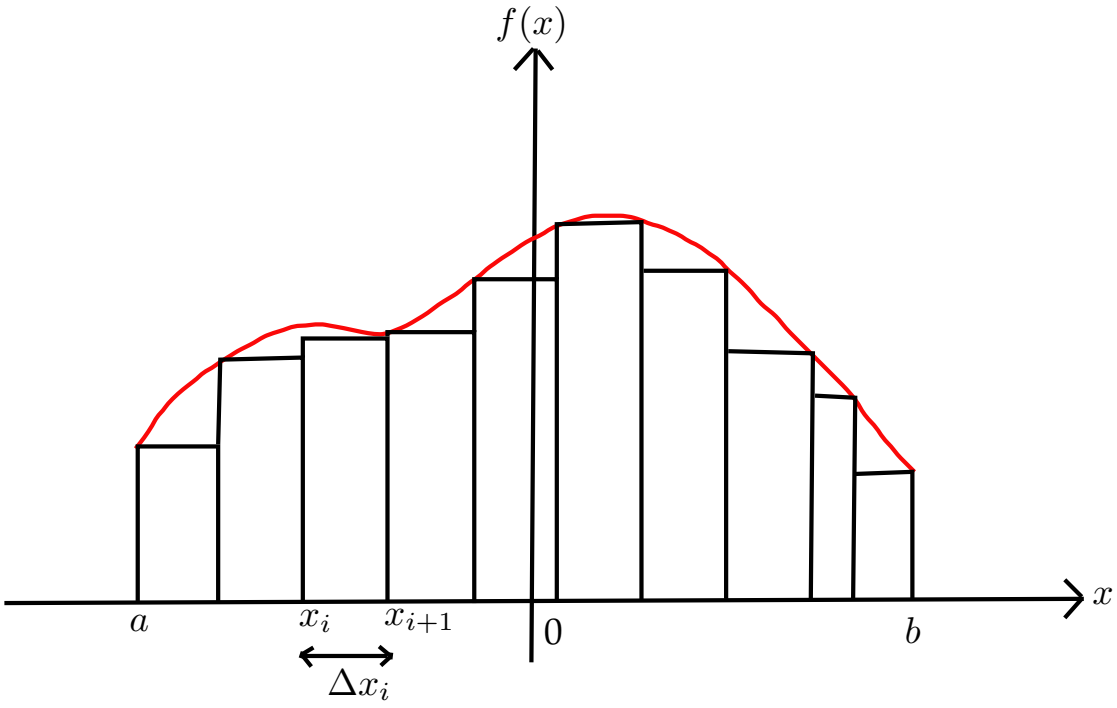

Figure 17.2: Lower Riemann approximation of f.

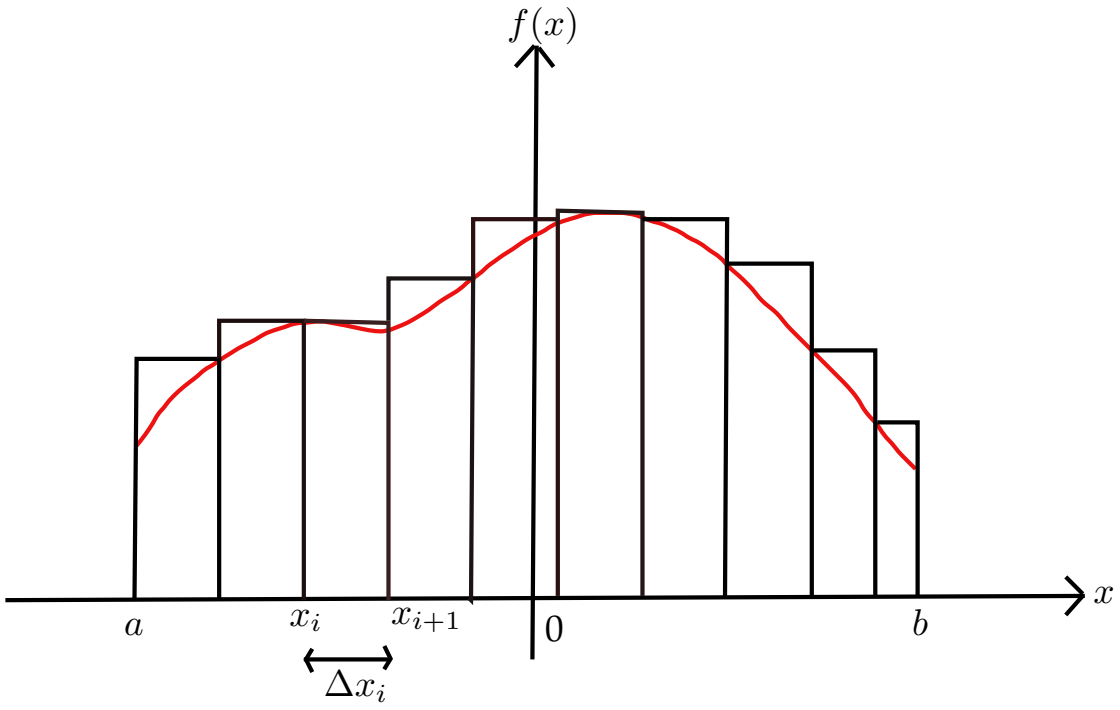

Figure 17.3: Upper Riemann approximation of f.

the partition. It is also clear that the following relation always holds

$$\underline{\int_a^b} f(x)\ \mathrm{d}x \leq \overline{\int_a^b} f(x)\ \mathrm{d}x.$$

Definition 17.1 *A function f is said to be Riemann Integrable if the values of the Lower Riemann Integral and the Upper Riemann Integral coincide. In such a case, the Riemann integral of f is that common value. That is, the Riemann Integral of f (when it exists), denoted by $\int_a^b f(x)\ dx$, is given by*

$$\int_a^b f(x)\ dx = \underline{\int_a^b} f(x)\ dx = \overline{\int_a^b} f(x)\ dx.$$

Figure (17.1) shows the graph of an arbitrary function f over the interval $[a, b]$. Figures (17.2) and (17.3) show the lower and upper Riemann approximations of f respectively, wherein f is graphed in red for reference. The lower (resp. upper) Riemann sum is the area under the lower (resp. upper) Riemann approximation in figure (17.2) (resp. (17.3)). We can imagine the Lower (resp. Upper) Riemann Sum to be approximating the area under f from below (resp. above). The intuition is that as the partitions become "finer", the Upper and Lower Riemann Sums converge to the area under f from above and below respectively. Also notice that figures (17.2) and (17.3) also represent unequal partition sizes pictorially.

Next, we turn our attention to Abstract Integrals.

17.2 Abstract Integration

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and $f : \Omega \to \mathbb{R}$ be a $\mathcal{F}$-measurable function. For any $A \in \mathcal{F}$, we would like to define

$$\int_A f \, \mathrm{d}\mu.$$

We will call the above quantity the integral of f with respect to the measure μ over the $\mathcal{F}$-measurable set A. Also, in the interest of notational simplicity, we will use the following two notations interchangeably to mean the integral of the $\mathcal{F}$-measurable function f with respect to the measure μ over the entire space

$$\int f \, \mathrm{d}\mu \equiv \int_\Omega f \, \mathrm{d}\mu.$$

Before we define the abstract integral, let us look at two very important special cases.

17.2.1 Special Cases

1. *The Lebesgue integral*: Let $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$
   Let $f : \mathbb{R} \to \mathbb{R}$ be a Borel-measurable function, and in this case the integral

   $$\int f \, \mathrm{d}\lambda$$

   is called the *Lebesgue Integral* of f over the Reals.
   The Lebesgue integral can be shown to be a generalisation of the Riemann integral. In particular, it allows us to integrate over arbitrary Borel sets, instead of just intervals. Moreover, we will see that the Lebesgue integral might exist even when the Riemann integral does not. However, if a function is Riemann integrable over an interval, then it is necessarily Lebesgue integrable, and the values of the two integrals will be equal.

2. *Expectation of a random variable:* Let $(\Omega, \mathcal{F}, \mu) = (\Omega, \mathcal{F}, \mathbb{P})$
   Let $X : \Omega \to \mathbb{R}$ be a random variable, and in this case the integral

   $$\int X \, \mathrm{d}\mathbb{P},$$

   is called the *Expectation* of the random variable X, and is denoted by $\mathbb{E}[X]$. Therefore,

   $$\mathbb{E}[X] \triangleq \int X \, \mathrm{d}\mathbb{P}.$$

Note that, so far, we have not *defined* what an abstract integral is. We have only introduced the notations and terminologies used. We now lay out the roadmap to defining the abstract integral.

17.2.2 Roadmap for defining the abstract integral

The abstract integral of an arbitrary, $\mathcal{F}$-measurable function f is defined in four steps as outlined below:

1. First, we define the integral for *simple functions*, i.e., non-negative functions that take only finitely many values.

2. Second, we define the integral for non-negative functions. This is done by approximating the function by simple functions, thus allowing us to define the integral of the non-negative function in terms of the integrals of the simple functions.

3. Third, we write the arbitrary function f as $f = f_+ - f_-$, where f_+ and f_- are non-negative functions which correspond to the positive and negative components of f. Then, we define the integral of f as
$$\int f \, \mathrm{d}\mu = \int f_+ \, \mathrm{d}\mu - \int f_- \, \mathrm{d}\mu.$$

4. And last, we define
$$\int_A f \, \mathrm{d}\mu \triangleq \int f \mathbb{I}_A \, \mathrm{d}\mu.$$

17.2.3 Abstract Integrals of Simple Functions

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to \mathbb{R}$ be a $\mathcal{F}$-measurable function.

Definition 17.2 *A function f is said to be a* simple *function if it can be written as*
$$f(\omega) = \sum_{i=1}^{n} a_i \mathbb{I}_{A_i}(\omega), \ \forall \ \omega \in \Omega, \tag{17.1}$$
where $a_i \geq 0 \ \forall \ i \in \{1, 2, \ldots, n\}$, and $A_i \in \mathcal{F} \ \forall \ i \in \{1, 2, \ldots, n\}$.

Remark: 17.3 *Note that $f(\omega)$ written in this form is not unique. This problem is circumvented using the "canonical" representation, wherein we restrict the a_i's to be distinct and the A_i's to be disjoint. It can be verified that this restriction enforces uniqueness of representation. Henceforth, whenever the term "simple function" is used, it will be taken implicitly to be in the canonical form.*

Figure (17.4) shows the canonical representation of a simple random variable X taking 4 values such that $\omega \in A_i \implies X(\omega) = a_i \ \forall \ i \in \{1, 2, 3, 4\}$. As the figure shows, disjoint events are mapped to distinct, non-negative real numbers.

Definition 17.4 *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $f \geq 0$ be a simple function with the canonical representation* (17.2)*. The abstract integral of f with respect to the measure μ is defined as*
$$\int f \ d\mu \triangleq \sum_{i=1}^{n} a_i \mu(A_i).$$

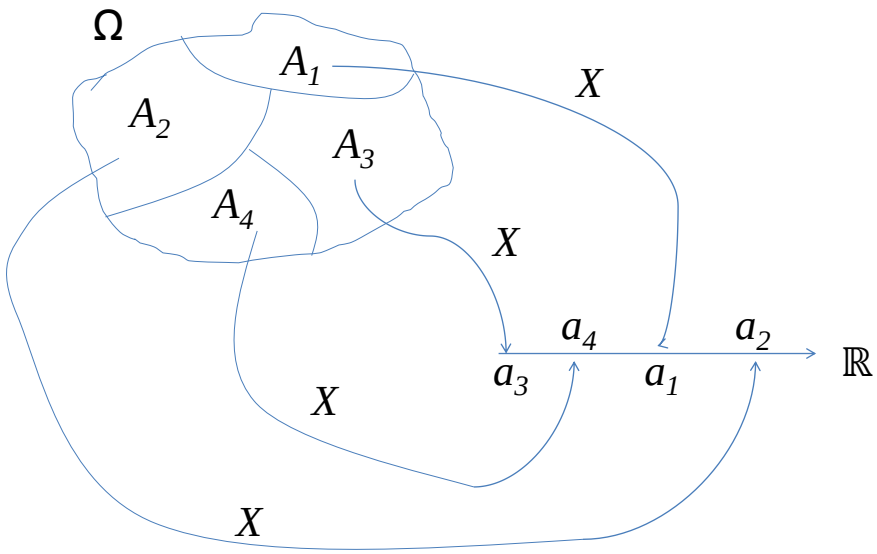

Figure 17.4: Canonical representation of a Simple Random Variable taking 4 values.

Example 1:- Consider the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, and define $f(\omega) = u(\omega) + u(\omega - 1)$ - $2u(\omega - 3)$, where $u(.)$ is the Heavyside Step function. The canonical representation of this simple function is $f(\omega) = \mathbb{I}_{[0,1]} + 2\mathbb{I}_{[1,3]}$. Therefore, the Lebesgue integral of this function is

$$\begin{aligned} \int f \, \mathrm{d}\lambda &= 1 \times \lambda([0,1]) + 2 \times \lambda([1,3]), \\ &= 1 \times 1 + 2 \times 2, \\ &= 5. \end{aligned}$$

We note that the value of the integral equals the area under the curve.
Example 2:- Consider the probability space $(\Omega = \{H, T\}^n, \mathcal{F}, \mathbb{P})$ and let $X : \Omega \to \mathbb{R}$ be a simple random variable such that $\mathbb{P}(\{H\}) = p$. This can be considered a model for n independent coin tosses. If $X(\omega)$ represents the number of heads, then the expected value of X (i.e. the integral of X with respect to $\mathbb{P}$) can be calculated as

$$\begin{aligned} \mathbb{E}[X] &= \sum_{i=1}^{n} i\mathbb{P}(X = i), \\ &= \sum_{i=1}^{n} i \binom{n}{i} p^i (1-p)^{n-i}, \\ &= np. \end{aligned}$$

Example 3:- Consider the Dirichlet function (D) defined to take on the value 1 on the rationals in the interval $[0,1]$ and the value 0 on the irrationals in the interval $[0,1]$. For this function,

$$\begin{aligned}\overline{\int_0^1} D(x)\ \mathrm{d}x &= 0,\\ \underline{\int_0^1} D(x)\ \mathrm{d}x &= 1.\end{aligned}$$

Hence, the Dirichlet function is not Riemann Integrable. On the other hand, the Dirichlet function is a simple function with the canonical representation

$$D(\omega) = \mathbb{I}_{\mathbb{Q}\cap[0,1]}.$$

Therefore, the Lebesgue Integral of the Dirichlet function is given by

$$\begin{aligned}\int D(\omega)\ \mathrm{d}\lambda &= 1 \times \lambda(\mathbb{Q}\cap[0,1]),\\ &= 0.\end{aligned}$$

This is because every partition of the horizontal axis, no matter how fine, contains both rational and irrational points. Therefore, we see that the Dirichlet function is trivially Lebesgue Integrable while not being Riemann Integrable.

17.2.4 Abstract Integral of non-negative functions

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and $f : \Omega \to \mathbb{R}_+$ be a non-negative, $\mathcal{F}$-measurable function. Denote by $S(f)$ the collection of all simple functions $q : \Omega \to \mathbb{R}_+$ such that $q(\omega) \leq f(\omega)\ \forall\ \omega \in \Omega$. That is, given a non-negative function f, we collect all the simple functions q's that approximate f from below. Having done this, we now define the abstract integral of f as follows:

Definition 17.5 *The abstract integral of f with respect to the measure μ is defined as*

$$\int f\ d\mu \triangleq \sup_{q\in S(f)} \int q\ d\mu. \tag{17.2}$$

Since q's are simple functions, calculation of their integral is known. The above equation gives a way to find the integral of any non-negative function. While being mathematically well-defined, (17.2) does not directly yield a practical method to compute the integral. We will address this issue later.

17.2.5 Abstract Integral of arbitrary functions

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and $f : \Omega \to \mathbb{R}$ be any arbitrary $\mathcal{F}$-measurable function. Then, in order to evaluate $\int f\ \mathrm{d}\mu$, we first write f as $f = f_+ - f_-$, where $f_+ \triangleq \max(f, 0) \geq 0$ and $f_- \triangleq -\min(f, 0) \geq 0$. We then define the integral of f with respect to μ as

$$\int f\ \mathrm{d}\mu \triangleq \int f_+\ \mathrm{d}\mu - \int f_-\ \mathrm{d}\mu, \tag{17.3}$$

wherein the integrals of f_+ and f_- as calculated as in the previous section. Since f_+ and f_- are nonnegative functions, both integrals on right hand side of (17.3) is well defined. The above definition is meaningful, as long as at least one of the integrals on the right hand side of (17.3) is finite. The integral of f is left undefined if the integrals of f_+ and f_- are both infinite.

17.2.6 Abstract Integral of arbitrary functions over a given set

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $f : \Omega \to \mathbb{R}$ be any arbitrary $\mathcal{F}$-measurable function, and $A \in \mathcal{F}$. Define $g \triangleq f\mathbb{I}_A$. That is, we consider the function f restricted to the set A. This is an $\mathcal{F}$-measurable function since it is a product of two $\mathcal{F}$-measurable functions. Its integral can be calculated as mentioned in the previous section. Therefore,

$$\int_A f \, \mathrm{d}\mu = \int f\mathbb{I}_A \, \mathrm{d}\mu = \int g \, \mathrm{d}\mu = \int g_+ \, \mathrm{d}\mu - \int g_- \, \mathrm{d}\mu.$$

17.3 Exercises

1. Consider the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, and define $f : \mathbb{R} \to \mathbb{R}$. Find out the Lebesgue integral of the function f for the following cases,

   (a)
   $$f(\omega) = \begin{cases} \omega, & \text{for } \omega = 0, 1, .., n \\ 0, & elsewhere. \end{cases}$$

   (b)
   $$f(\omega) = \begin{cases} 1, & \text{for } \omega = \mathbb{Q}^c \cap [0, 1] \\ 0, & elsewhere. \end{cases}$$

   (c)
   $$f(\omega) = \begin{cases} n, & \text{for } \omega = \mathbb{Q}^c \cap [0, n] \\ 0, & elsewhere. \end{cases}$$

2. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and define the random variable $X : \Omega \to \mathbb{R}$. Find $\mathbb{E}[X]$ for the following cases,

   (a) $\Omega = \{\omega_1, \omega_2, .., \omega_n\}$, with $\mathbb{P}(\omega_i) = 1/n$ for $i = 1, 2, .., n$ and $X = \mathbb{I}_A$, where $A = \{\omega_1, \omega_2, .., \omega_m\}$ with $1 \leq m \leq n$.

   (b) In part (a) if X is defined as follows,
   $$X(\omega_i) = \begin{cases} i, & \text{for } \omega_i \in A \\ 0, & elsewhere. \end{cases}$$

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 18: Properties of Abstract Integrals

Lecturer: Dr. Krishna Jagannathan *Scribe: Ravi Kumar Kolla*

In this lecture, we discuss some basic properties of abstract integrals.

18.1 Properties of Abstract Integrals

We will state the properties for a generic abstract integral, and also particularize for the special case of the expectation of a random variable.

Let $(\Omega, \mathcal{F})$ be a measurable space and f, g, h be measurable functions from Ω to $\mathbb{R}$. Let μ be a generic measure and $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$. Let X, Y be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. The first part of each property corresponds to the generic measure (μ), while the second part particularizes to the probability measure $(\mathbb{P})$.

In order to prove the properties, we follow a procedure: we begin by proving them for simple functions, and then extend it for the case of non-negative functions. Finally using these, we prove the properties for general measurable functions.

[PAI 1] $\int \mathbb{I}_A \, d\mu = \mu(A)$, for any $A \in \mathcal{F}$. In particular, for any $A \in \mathcal{F}$, we have $\mathbb{E}[\mathbb{I}_A] = \mathbb{P}(A)$.

Proof: $f(\omega) = I_A(\omega)$ is a simple function and is in canonical form. So, proof directly follows from the definition of integral of simple functions (Definition 17.4 from Lecture #17).

■

[PAI 2] If $g \geq 0$, then $\int g \, d\mu \geq 0$. If $X \geq 0$, then $\mathbb{E}(X) \geq 0$.

Proof: Let g be a simple function and $g \geq 0$. A simple function is of the form equation (17.1) from Lecture #17 with all a_i's non-negative. So, $\int g \, d\mu \geq 0$.

Now, we prove the property for non-negative functions. Let g be a non-negative function. Let $\mathcal{S}(g)$ contains all simple functions, $q(\omega)$ such that $q(\omega) \leq g(\omega), \ \forall \omega \in \Omega$.
So, $\int q \, d\mu \geq 0$, $\forall q \in \mathcal{S}(g)$, since q is a simple function.
Hence, $\int g \, d\mu = \sup_{q \in \mathcal{S}(g)} \int q d\, \mu \geq 0$, since supremum of a set of non-negative numbers is non-negative.

■

[PAI 3] If $g = 0$, μ.a.e., then $\int g \, d\mu = 0$. If $X = 0$ a.s., then $\mathbb{E}(X) = 0$.

Proof: Let $g = 0$ μ.a.e. be a simple function. Then, g has a canonical representation of the form $g = \sum_{i=1}^{k} a_i \mathbb{I}_{A_i}$, where $\mu(A_i) = 0$, for each i. Hence, $\int g \, d\mu = 0$.

Let $g = 0$ μ.a.e. be a non-negative function. Let $q \in \mathcal{S}(g) \Rightarrow q(\omega) \leq g(\omega), q(\omega) \geq 0$, $\forall q \in \mathcal{S}(g)$ and $\forall \omega$.
Since, $g = 0$ we have $q(\omega) = 0$, $\forall\, \omega$ i.e., $q(\omega) = 0$, μ.a.e..
Hence, $\int q \, d\mu = 0$, $\forall q \in \mathcal{S}(g) \Rightarrow \int g \, d\mu = 0$.

■

[PAI 4] **[Linearity]** $\int (g + h) \, d\mu = \int g \, d\mu + \int h \, d\mu$. And, $\mathbb{E}(X + Y) = E(X) + E(Y)$.

Proof: Let g and h be simple functions. We can write g and h in canonical representation form as:

$$g = \sum_{i=1}^{k} a_i \mathbb{I}_{A_i}, \qquad h = \sum_{j=1}^{m} b_j \mathbb{I}_{B_j},$$

where the sets A_i are disjoint, and the sets B_j are also disjoint. So, the sets $A_i \cap B_j$ are disjoint. Then, $g + h$ can be written as:

$$g + h = \sum_{i=1}^{k} \sum_{j=1}^{m} (a_i + b_j)\, \mathbb{I}_{A_i \cap B_j}.$$

So,

$$\begin{aligned}
\int (g+h)\, d\mu &\overset{(a)}{=} \sum_{i=1}^{k} \sum_{j=1}^{m} (a_i + b_j)\, \mu\left(A_i \cap B_j\right), \\
&= \sum_{i=1}^{k} a_i \sum_{j=1}^{m} \mu\left(A_i \cap B_j\right) + \sum_{j=1}^{m} b_j \sum_{i=1}^{k} \mu\left(A_i \cap B_j\right), \\
&\overset{(b)}{=} \sum_{i=1}^{k} a_i \mu(A_i) + \sum_{j=1}^{m} b_j \mu(B_j), \\
&\overset{(c)}{=} \int g\, d\mu + \int h\, d\mu.
\end{aligned}$$

Where (a) and (c) are due to definition of integral of simple functions, (b) is due to finite additivity of μ.

Proving linearity for general non-negative functions is not easy at this point. We will return to finish this proof after equipping ourselves with the Monotone Convergence Theorem.

■

[PAI 5] If $0 \le g \le h$ μ.a.e., then $\int g\, d\mu \le \int h\, d\mu$. In particular, if $0 \le X \le Y$ a.s., then $\mathbb{E}(X) \le \mathbb{E}(Y)$.

Proof: Let g and h be simple functions and $0 \le g \le h$ μ.a.e.. Then, we have $h = g + q$, for some simple function $q \ge 0$ μ.a.e.. But, we can write $q = q_+ - q_-$, where $q_+ \ge 0$ and $q_- \ge 0$, and $q_- = 0$ μ.a.e.. Here, q, q_+, q_- are all simple functions. Using the linearity property [PAI 4] and then properties [PAI 3], [PAI 2], we write

$$\int h\, d\mu = \int g\, d\mu + \int q_+\, d\mu - \int q_-\, d\mu = \int g\, d\mu + \int q_+\, d\mu \ge \int g\, d\mu.$$

Let g and h be non-negative functions and $0 \le g \le h$ μ.a.e.. Let $q \in \mathcal{S}(g) \Rightarrow q(\omega) \le g(w) \le h(\omega)\, \forall\, \omega \in \Omega \Rightarrow q \in \mathcal{S}(h)$. So, $\mathcal{S}(g) \subseteq \mathcal{S}(h)$. Hence, $\sup_{q \in \mathcal{S}(g)} \int q\, d\mu \le \sup_{q \in \mathcal{S}(h)} \int q d\mu \Rightarrow \int g\, d\mu \le \int h\, d\mu$.

■

[PAI 6] If $g = h$ μ.a.e., then $\int g\, d\mu = \int h\, d\mu$. If $X = Y$ a.s., then $\mathbb{E}(X) = \mathbb{E}(Y)$.

Proof: The proof follows from the above property since $g = h$ μ.a.e. $\Leftrightarrow g \le h$ μ.a.e. and $h \le g$ μ.a.e.

■

Example: The Dirichlet function and the zero function are equal μ.a.e.. So, they have the same integral equal to zero under Lebesgue measure.
Note that the measure with respect to which the integration is performed on both sides must be the same for the equality to hold.

[PAI 7] If $g \geq 0$ μ.a.e., and $\int g\, d\mu = 0$, then $g = 0$ μ.a.e.. If $X \geq 0$ a.s., and $\mathbb{E}(X) = 0$, then $X = 0$ a.s..

Proof: Let g be a simple function. Let $g \geq 0$ μ.a.e., and $\int g\, d\mu = 0$. We can write $g = g_+ - g_-$, where $g_+ \geq 0$ and $g_- \geq 0$. Then, $g_- = 0$ μ.a.e.. Using [PAI 3], we get $\int g_-\, d\mu = 0$.
Due to Linearity property[PAI 4], we can write $\int g_+\, d\mu = \int g\, d\mu + \int g_-\, d\mu = 0$.
Observe that g_+ is a simple function. So, g_+ has a canonical representation of the form: $g_+ = \sum_{i=1}^{k} a_i \mathbb{I}_{A_i}$, with $a_i > 0$ for each i. It follows that $\mu(A_i) = 0 \,\forall\, i$, since $\sum_{i=1}^{k} a_i \mu(A_i) = 0$. Due to finite additivity, we conclude that $\mu\left(\bigcup_{i=1}^{k} A_i\right) = 0$. Therefore, $g_+ = 0$ μ.a.e., and $g = 0$ μ.a.e..

Let g be a non-negative function. We use proof by contradiction method to prove this property. Suppose the contrary, i.e., $B = \{\omega | g(\omega) > 0\}$, where $\mu(B) > 0$.
Let $B_n = \{\omega | g(\omega) > \frac{1}{n}\}$. Clearly $B_n \subseteq B_{n+1} \,\forall n \in \mathbb{N}$ and $\bigcup_{i=1}^{\infty} B_n = B$. So, $\mu(B) = \mu\left(\bigcup_{i=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mu(B_n) > 0$ which implies that $\exists\, k \in \mathbb{N}$ such that $\mu(B_k) > 0$ (from properties of limits of sequences).
$\int g\, d\mu \geq \int_B g\, d\mu = \int g \mathbb{I}_B\, d\mu \geq \int g \mathbb{I}_{B_k}\, d\mu > \int \frac{1}{k} \mathbb{I}_{B_k}\, d\mu = \frac{1}{k}\mu(B_k) > 0$, which is a contradiction!

■

[PAI 8] **[Scaling]** Let $a \geq 0$. Then $\int (af)\, d\mu = a \int f d\mu$. If $a \geq 0$, then $\mathbb{E}(aX) = a\mathbb{E}(X)$.

Proof: Let f be a simple function. It is trivially true in this case (Why?). We should be careful for the case where $\int f\, d\mu = \infty$ and $a = 0$. We see that $af = 0 \Rightarrow \int (af)\, d\mu = 0 = 0 \times \infty$ (By convention for extended Reals!) $= a \int f\, d\mu$, so the property holds.

Let f be a non-negative function. If $a = 0$, then the result is obvious. So, consider the case $a > 0$. It can be easily seen that $q \in \mathcal{S}(f) \Leftrightarrow aq \in \mathcal{S}(af)$.
$\int (af)\, d\mu = \sup_{q' \in \mathcal{S}(af)} \int q'\, d\mu = \sup_{aq \in \mathcal{S}(af)} \int (aq)\, d\mu = \sup_{q \in \mathcal{S}(f)} \int (aq)\, d\mu = \sup_{q \in \mathcal{S}(f)} a \int q\, d\mu = a \sup_{q \in \mathcal{S}(f)} \int q\, d\mu = a \int f\, d\mu$.

■

With the help of point (3) in section 17.2.2 from Lecture #17, proving the above properties for general measurable functions is not difficult, and is left as an exercise to the reader.

As mentioned in a previous lecture, we now prove the Inclusion-Exclusion property of probability measure using indicator random variables and their expectation.

Inclusion-Exclusion property of probability measures:

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A_1, A_2, \cdots, A_n$ be elements of $\mathcal{F}$. Then,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n-1} \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right).$$

Proof:

$$\begin{aligned}
\mathbb{I}_{\bigcup_{i=1}^{n} A_i} &= 1 - \mathbb{I}_{\bigcap_{i=1}^{n} A_i^c}, \\
&= 1 - \prod_{i=1}^{n} \mathbb{I}_{A_i^c}, \\
&= 1 - \prod_{i=1}^{n} \left(1 - \mathbb{I}_{A_i}\right). \\
&= \sum_{i=1}^{n} \mathbb{I}_{A_i} - \sum_{i<j} \mathbb{I}_{A_i}\mathbb{I}_{A_j} + \cdots + (-1)^{n-1}\Pi_{i=1}^{n}\mathbb{I}_{A_i}.
\end{aligned}$$

Taking expectation on both sides of the above equation yields the desired result, since $\mathbb{I}_{A_i}\mathbb{I}_{A_j} = \mathbb{I}_{A_i \cap A_j}$. ■

Now, we summarize all the properties here:

[PAI 1]	$\int \mathbb{I}_A \, d\mu = \mu(A)$	$\mathbb{E}[\mathbb{I}_A] = \mathbb{P}(A)$
[PAI 2]	$g \geq 0, \Rightarrow \int g \, d\mu \geq 0$	$X \geq 0, \Rightarrow \mathbb{E}(X) \geq 0$
[PAI 3]	$g = 0 \, \mu$.a.e., $\Rightarrow \int g \, d\mu = 0$	$X = 0$ a.s., $\Rightarrow \mathbb{E}(X) = 0$
[PAI 4]	$\int (g+h) \, d\mu = \int g \, d\mu + \int h \, d\mu$	$\mathbb{E}(X+Y) = E(X) + E(Y)$
[PAI 5]	$0 \leq g \leq h \, \mu$.a.e., $\Rightarrow \int g \, d\mu \leq \int h \, d\mu$	$0 \leq X \leq Y$ a.s., $\Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$
[PAI 6]	$g = h \, \mu$.a.e., $\Rightarrow \int g \, d\mu = \int h \, d\mu$	$X = Y$ a.s., $\Rightarrow \mathbb{E}(X) = \mathbb{E}(Y)$
[PAI 7]	$g \geq 0 \, \mu$.a.e., and $\int g \, d\mu = 0, \Rightarrow g = 0$ a.e.	$X \geq 0$ a.s., and $\mathbb{E}(X) = 0, \Rightarrow X = 0$ a.s..
[PAI 8]	$a \geq 0$, $\int (af) \, d\mu = a \int f d\mu$	$a \geq 0$, $\mathbb{E}(aX) = a\mathbb{E}(X)$

18.2 Exercise:

1. Show that if $g : \Omega \to [0, \infty]$ satisfies $\int g \, d\mu < \infty$, then $g < \infty$, μ.a.e..
2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $g : \Omega \to \mathbb{R}$ be a non-negative measurable function. Let λ be a lebesgue measure. Let f be a non-negative measurable function on the real line such that $\int f \, d\lambda = 1$. For any Borel set A, if $\mathbb{P}_1(A) = \int_A f \, d\lambda$, then prove that $\mathbb{P}_1$ is a probability measure.
3. Let X_1, X_2, ..., X_n be i.i.d. random variables for which $\mathbb{E}[X_1^{-1}]$ exists. Show that if $m \leq n$, then $\mathbb{E}[\frac{S_m}{S_n}] = \frac{m}{n}$, where $S_m = X_1 + X_2 + ... + X_m$.
4. Consider the Real line endowed with the Borel σ-algebra, and let $c \in \mathbb{R}$ be fixed. Then the *Dirac measure* at c, denoted as δ_c, is defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as follows. For any Borel set A, $\delta_c(A) = 1$ if $c \in A$, and $\delta_c(A) = 0$ if $c \notin A$. It is quite easy to see that $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \delta_c)$ is a measure space (indeed, it is a probability space). The Dirac measure is referred to as *unit impulse* in the engineering literature, and sometimes (incorrectly) called a Dirac delta "function".

(a) Let g be a non-negative, measurable function. Show that $\int g \, \mathrm{d}\delta_c = g(c)$.

Now, let us define a *counting measure* on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as $\mu(A) = \sum_{n=1}^{\infty} \delta_n(A)$. In words, $\mu(A)$simply counts the number of natural numbers contained in the Borel set A. In engineering parlance, the counting measure is called an impulse train.

(b) Let g be a non-negative, measurable function. Show that $\int g \, \mathrm{d}\delta_c = \sum_{n=1}^{\infty} g(n)$. Thus, summation is just a special case of integration. In particular, summation is nothing but integral with respect to the counting measure!

EE5110 : Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 19: Monotone Convergence Theorem

Lecturer: Dr. Krishna Jagannathan *Scribes: Vishakh Hegde*

In this lecture, we present the Monotone Convergence Theorem (henceforth called MCT), which is considered one of the cornerstones of integration theory. The MCT gives us a sufficient condition for interchanging limit and integral. We also prove the linearity property of integrals using the MCT. Recall the $g_n \longrightarrow g$ μ.a.e. if $g_n(\omega) \longrightarrow g(\omega)$ $\forall \omega \in \Omega$ except possibly on a set of $\mu-$measure zero.

19.1 Monotone Convergence Theorem

Theorem 19.1 *Let $g_n \geq 0$ be a sequence of measurable functions such that $g_n \uparrow g$ μ.a.e. (That is, except perhaps on a set of μ-measure zero, we have $g_n(\omega) \to g(\omega)$, and $g_n(\omega) \leq g_{n+1}(\omega)$, $n \geq 1$). We then have $\int g_n d\mu \uparrow \int g d\mu$. In other words,*

$$\lim_{n\to\infty} \int g_n \, d\mu = \int g \, d\mu.$$

See Section 5.2 in Lecture 11 of [1] for the proof.

Example 19.2 *Consider $([0,1], \mathcal{B}, \lambda)$ and consider the sequence of functions given by,*

$$f_n(\omega) = \begin{cases} n, & \text{if } 0 < \omega \leq 1/n, \\ 0, & \text{otherwise.} \end{cases}$$

$$\int f_n d\lambda = 1, \forall n \Rightarrow \lim_{n\to\infty} \int f_n d\lambda = 1.$$

For $\omega > 0$, we have,

$$\lim_{n\to\infty} f_n(\omega) = 0.$$

For $\omega = 0$, we have,

$$\lim_{n\to\infty} f_n(\omega) = \infty.$$

Therefore we have,

$$\int f d\lambda = 0.$$

Hence we see that,

$$\int f d\lambda \neq \lim_{n\to\infty} \int f_n d\lambda.$$

Note that monotonicity does not hold in this example.

19.2 Linearity of Integrals

In this section, we will prove the linearity property of integrals, using the MCT. Recall that we stated the linearity property in the previous lecture as **PAI 4** but proved it only for simple functions. Here we prove it in full generality.

Let f and g be simple functions. Therefore we can express them as,

$$f = \sum_{i=1}^{n} a_i \mathbb{I}_{A_i},$$
$$g = \sum_{j=1}^{m} b_j \mathbb{I}_{B_j}.$$

Here A_i and B_i are $\mathcal{F}$ measurable sets and I_{A_i} and I_{B_j} are indicator variables. Summing f and g, we obtain,

$$f + g = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \mathbb{I}_{A_i \cap B_j}. \tag{19.1}$$

Note that f and g are canonical representations. This implies that A_i's are disjoint sets, and so are B_j's. Therefore $A_i \cap B_j$ are disjoint sets. Hence we have,

$$\begin{aligned}\int f + g \; d\mu &= \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \mu(A_i \cap B_j), \\ &= \sum_{i=1}^{n} a_i \sum_{j=1}^{m} \mu(A_i \cap B_j) + \sum_{j=1}^{m} b_j \sum_{i=1}^{n} \mu(A_i \cap B_j).\end{aligned}$$

By finite additivity property, we have,

$$\begin{aligned}\int f + g \; d\mu &= \sum_{i=1}^{n} a_i \mu(A_i) + \sum_{j=1}^{m} b_j \mu(B_j), \\ &= \int f \; d\mu + \int g \; d\mu.\end{aligned}$$

Next, we need to prove linearity for non-negative measurable functions. Let f_n and g_n (with $n \geq 1$) be sequences of simple functions where, $f_n \uparrow f$ and $g_n \uparrow g$. Such a simple sequence always exist for every non-negative measurable function, as we will show in the next section. Now, since f_n and g_n are monotonic, $f_n + g_n$ is monotonic. Then we can show that $(f_n + g_n) \uparrow (f + g)$. Using MCT, we have,

$$\int (f + g) d\mu = \lim_{n \to \infty} \int (f_n + g_n) d\mu. \tag{19.2}$$

But f_n and g_n are simple functions. We know that, for simple functions,

$$\int (f_n + g_n) d\mu = \int f_n d\mu + \int g_n d\mu.$$

Thus,

$$\begin{aligned}\lim_{n \to \infty} \int (f_n + g_n) d\mu &= \lim_{n \to \infty} \int f_n d\mu + \lim_{n \to \infty} \int g_n d\mu, \\ &\overset{MCT}{=} \int f d\mu + \int g d\mu.\end{aligned}$$

This implies that,

$$\int (f+g)d\mu = \int f d\mu + \int g d\mu. \tag{19.3}$$

This proves linearity for non-negative functions.

For arbitrary measurable functions f and g, we can write them as $f = f_+ - f_-$ and $g = g_+ - g_-$ where f_+, f_-, g_+ and g_- are non-negative measurable functions. A similar proof can then be worked out which completes the proof of linearity.

19.3 Integration using simple functions

Our earlier definition $\int g d\mu = \sup_{q \in S(g)} \int q d\mu$ helped us to prove some properties of abstract integrals quite easily. However, it does not give us a practical way of performing the integration. In this section, we present a method to explicitly compute the integral, using the MCT. First, we approximate the function to be integrated using simple functions from below. Specifically, define

$$g_n(\omega) = \begin{cases} n, & \text{if } g(\omega) \geq n, \\ \frac{i}{2^n}, & \text{if } \frac{i}{2^n} \leq g(\omega) < \frac{i+1}{2^n}; i \in \{0, 1, \ldots, n2^n - 1\}. \end{cases} \tag{19.4}$$

Thus, the function to be integrated in quantized to $n2^n$ levels. Next, we note here that $g_n(\omega)$ is a simple function since it can be written as

$$g_n(\omega) = \sum_{i=0}^{n2^n-1} \frac{i}{2^n} \mathbb{I}_{\{\omega : \frac{i}{2^n} \leq g(\omega) < \frac{i+1}{2^n}\}} + n\mathbb{I}_{\{g_n(\omega) \geq n\}}. \tag{19.5}$$

Claim 1: We can easily show that:

- $g_n(\omega) \to g(\omega)\ \forall \omega \in \Omega$.
- $g_n(\omega) \leq g_{n+1}(\omega)\ \forall \omega \in \Omega$ and $\forall n \in \mathbb{N}$.

Therefore, using MCT, we have,

$$\begin{aligned} \int g d\mu &= \lim_{n\to\infty} \int g_n d\mu, \\ &= \lim_{n\to\infty} \sum_{i=0}^{n2^n-1} \frac{i}{2^n} \mu\left(\omega : \frac{i}{2^n} \leq g(\omega) \leq \frac{i+1}{2^n}\right) + n\mu\left(g_n(\omega) \geq n\right). \end{aligned}$$

Now, if g is bounded the second term $\mu(g_n(\omega) \geq n)$ will be 0 and if g is unbounded, it may or may not be finite.

This gives us an explicit way to compute the abstract integral.

19.4 Exercise:

1. Prove Claim 1.

2. Let X be a *non-negative* random variable (not necessarily discrete or continuous) with $\mathbb{E}[X] < \infty$.

   (a) Prove that $\lim_{n\to\infty} n\mathbb{P}(X > n) = 0$. [Hint: Write $\mathbb{E}[X] = \mathbb{E}[X\mathbb{I}_{\{X\leq n\}}] + \mathbb{E}[X\mathbb{I}_{\{X>n\}}]$.]

   (b) Prove that $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)\ \mathrm{d}x$. Yes, the integral on the right *is* just a plain old Riemann integral! [Hint: Write out $\mathbb{E}[X] = \int x\ \mathrm{d}\mathbb{P}_X$ as the limit of a sum, and use part (a) for the last term.]

   We say a random variable X is stochastically larger than a random variable Y, and denote by $X \geq_{st} Y$, if $\mathbb{P}(X > a) \geq \mathbb{P}(Y > a)\ \forall a \in \mathbb{R}$.

   (c) For non-negative random variables X and Y, show that if $X \geq_{st} Y$, then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

3. Show that $f(x) = x^{-\alpha}$ is integrable on $[0, \infty)$ for $\alpha > 1$.

19.5 References:

[1] David Gamarnick and John Tsitsiklis, "Introduction to Probability", *MIT OCW*, 2008.

EE 5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 20: Expectation of Discrete RVs, Expectation over Different Spaces

Lecturer: Dr. Krishna Jagannathan *Scribe: Arjun Bhagoji*

20.1 Expectations of Discrete RVs

A discrete random variable $X(\omega)$, (which only takes a countable set of values) can be represented as follows:

Definition 20.1 $X(\omega) = \sum_{i=1}^{\infty} a_i \mathbb{I}_{A_i}(\omega)$ *where* $X \geq 0$.

In the canonical representation, the a_i's are non-negative and distinct, and the A_i's are disjoint. It is easy to see that the A_i's partition the sample space. Let us now define a sequence of simple random variables, which approximate X from below.

Definition 20.2 *Define* $X_n(\omega) = \sum_{i=1}^{n} a_i \mathbb{I}_{A_i}(\omega)$.

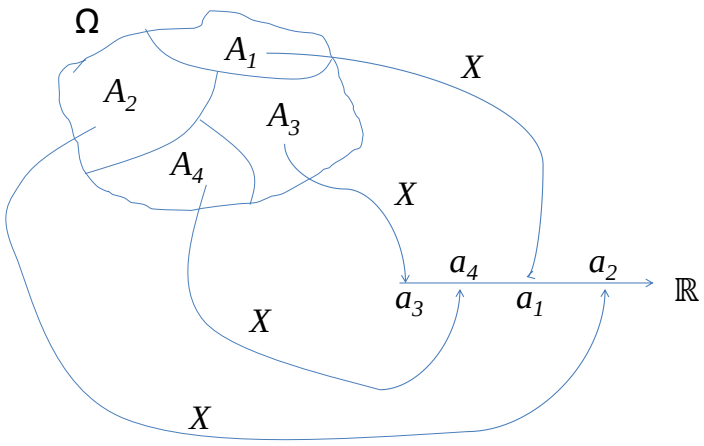

Figure 20.1: Simple random variable

Note that $\forall \omega$, $X_n(\omega) \leq X_{n+1}(\omega)$, where $n \geq 1$. Next, let us fix $\omega \in \Omega$. Since A_i's partition Ω, there exists $k \geq 1$ such that $\omega \in A_k$. Thus, $\forall n \geq k$, $X_n(\omega) = a_k$ and $\forall n < k$, $X_n(\omega) = 0$. Therefore,

$$\lim_{n \to \infty} X_n(\omega) = X(\omega) \qquad \forall \omega \in \Omega. \tag{20.1}$$

In other words, $X_n(\omega)$ is a sequence of simple functions converging monotonically to $X(\omega)$. Now applying

the Monotone Convergence Theorem (MCT) to the sequence of random variables X_n,

$$\begin{aligned}\mathbb{E}[X] &= \lim_{n\to\infty} \mathbb{E}[X_n], \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} a_i \mathbb{P}(A_i), \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} a_i \mathbb{P}(X = a_i), \\ \Rightarrow \mathbb{E}[X] &= \sum_{i=1}^{\infty} a_i \mathbb{P}(X = a_i). \end{aligned} \tag{20.2}$$

The limit of the sum is well-defined as X is a non-negative random variable and it either converges to some positive real number or goes to $+\infty$. If X is discrete but takes on both positive and negative values, we write $X = X_+ - X_-$, where $X_+ = \max(X, 0)$ and $X_- = -\min(X, 0)$. Then, we compute

$$\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-]. \tag{20.3}$$

The above is meaningful when *at least one* of the expectation on the right hand side is finite. We now give some examples.

1. $X \sim$ Geometric(p) - $\mathbb{E}[X] = \sum_{i=1}^{\infty} i(1-p)^{i-1}p = \frac{1}{p}$.
   This tells us that, for a geometric random variable, the expected number of trials for the first success to occur scales as $\frac{1}{p}$.

2. $\mathbb{P}(X = k) = \frac{6}{\pi^2}\frac{1}{k^2}$ for $k \geq 1$ - For this probability distribution,the expectation is calculated as

   $$\mathbb{E}[X] = \frac{6}{\pi^2} \sum_{i=1}^{\infty} i \left(\frac{1}{i^2}\right) = +\infty. \tag{20.4}$$

   In this example, we see that a random variable can have infinite expectation.

3. $\mathbb{P}(X = k) = \frac{3}{\pi^2}\frac{1}{k^2}$ for $k \in \mathbb{Z}/\{0\}$ - For this probability distribution, the expectation is calculated as $\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-]$. However, both the expectations $\mathbb{E}[X_+]$ and $\mathbb{E}[X_-]$ are infinite! Therefore, $\mathbb{E}[X]$ is not defined! This is an example of a discrete random variable with undefined expectation.

20.2 Connection between Riemann and Lebesgue integrals

The connection is given by the following theorem which we state without proof.

Theorem 20.3 *Let f be measurable and Riemann integrable over an interval $[a, b]$. Then,*

$$\int_{[a,b]} f \, d\lambda \text{ exists, and } \int_{[a,b]} f \, d\lambda = \int_a^b f(x) \, dx. \tag{20.5}$$

Here, λ is the Lebesgue measure on $\mathbb{R}$. The integral on the left is a Lebesgue integral while the one on the right is the standard Riemann integral.

20.3 Expectations on different spaces

We often want to compute the expectation of a function of a random variable, say $Y = f(X)$, where both X and Y are random variables and $f(\cdot)$ is a measurable function on $\mathbb{R}$. The following theorem asserts that the expectation can be computed over different spaces, to obtain the 'same answer.' For example, we can compute the expectation of Y by either working in the X-space or the Y-space to write (for discrete random variables)

$$\sum_i y_i \mathbb{P}(Y = y_i) = \sum_i f(a_i)\mathbb{P}(X = a_i), \tag{20.6}$$

where $y_i = f(a_i)$. This is just a special case of the following theorem

Theorem 20.4 *Denote the probability measure on the sample space by $\mathbb{P}$, on the range space of X as $\mathbb{P}_X$ and on range space of Y as $\mathbb{P}_Y$. Then, $\int Y\, d\mathbb{P} = \int f\, d\mathbb{P}_X = \int y\, d\mathbb{P}_Y$ where $Y = f(X)$ and the integrals are over the respective spaces.*

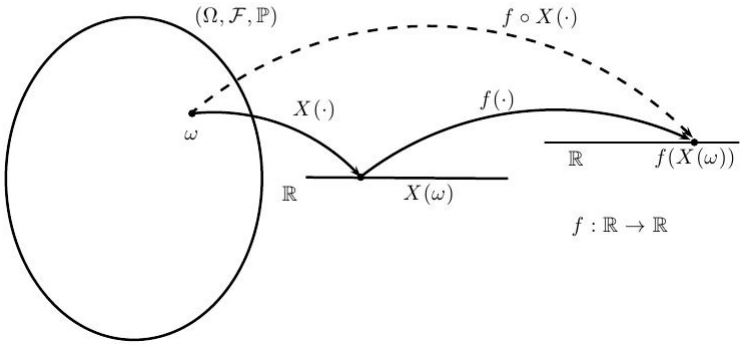

Figure 20.2: Different spaces considered

Proof: Let f be a simple function which takes values $y_1, y_2 \cdots y_n$. Then,

$$\begin{aligned}\int Y\, d\mathbb{P} &= \sum_{i=1}^{n} y_i \mathbb{P}(\omega | Y(\omega) = y_i), \\ &= \sum_{i=1}^{n} y_i \mathbb{P}(\omega | f(X(\omega)) = y_i).\end{aligned}$$

Now, looking at the second integral, we have

$$\begin{aligned}\int f\, d\mathbb{P}_X &= \sum_{i=1}^{n} y_i \mathbb{P}_X(x \in \mathbb{R} | f(x) = y_i),\\ &= \sum_{i=1}^{n} y_i \mathbb{P}_X(f^{-1}(y_i)),\\ &= \sum_{i=1}^{n} y_i \mathbb{P}(\omega | \omega : X(\omega) \in f^{-1}(y_i)),\\ &= \sum_{i=1}^{n} y_i \mathbb{P}(\omega | f(X(\omega)) = y_i).\end{aligned}$$

Now, we extend the above to the case when f is a non-negative measurable function. Let $\{f_n\}$ be a sequence of simple functions such that $f_n \uparrow f$ according to the construction given in the previous lecture. Thus, $f_n(X) \uparrow f(X)$ and,

$$\begin{aligned}\int Y\, d\mathbb{P} &= \int (f \cdot X)\, d\mathbb{P},\\ &= \int f(X)\, d\mathbb{P},\\ &= \lim_{n\to\infty} \int f_n(X)\, d\mathbb{P} \qquad \text{(by MCT)},\\ &= \lim_{n\to\infty} \int f_n\, d\mathbb{P}_X \qquad (\because \text{simple function}),\\ &= \int f\, d\mathbb{P}_X \qquad \text{(by MCT)}.\end{aligned}$$

This can now be simply extended to the case where g takes both negative and positive values. ■

A simple corollary of this theorem is that $\int X\, d\mathbb{P} = \int x\, d\mathbb{P}_X$.

20.4 Exercise

1. *[Dimitri P.Bertsekas]* Let X be a random variable with PMF $p_X(x) = \frac{x^2}{a}$, if $x = -3, -2, -1, 0, 1, 2, 3$ and zero otherwise. Compute a and $\mathbb{E}[X]$.
2. *[Dimitri P.Bertsekas]* As an advertising campaign, a chocolate factory places golden tickets in some of its candy bars, with the promise that a golden ticket is worth a trip through the chocolate factory, and all the chocolate you can eat for life. If the probability of finding a golden ticket is p, find the expected number of bars you need to eat to find a ticket.
3. *[Dimitri P.Bertsekas]* On a given day, your golf score takes values from the range 101 to 110, with probability 0.1, independent of other days. Determined to improve your score, you decide to play on three different days and declare as your score the minimum X of the scores X_1, X_2 and X_3 on the different days. By how much has your expected score improved as a result of playing on three days?
4. *[Papoulis]* A biased coin is tossed and the first outcome is noted. Let the probability of head occuring be p and that of a tail be $q = 1 - p$. The tossing is continued until the outcome is the complement of the first outcome, thus completing the first run. Let X denote the length of the first run. Find the PMF of X and show that $\mathbb{E}[X] = \frac{p}{q} + \frac{q}{p}$.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 21: Expectation of CRVs, Fatou's Lemma and DCT

Lecturer: Krishna Jagannathan *Scribe: Jainam Doshi*

In the present lecture, we will cover the following three topics:

- Integration of Continous Random Variables
- Fatou's Lemma
- Dominated Convergence Theorem (DCT)

21.1 Integration of Continuous Random Variables

Theorem 21.1 *Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $X : \Omega \to \mathbb{R}$ be a continuous random variable. Let g be a measurable function which is either non-negative or satisfies $\int |g| d\mathbb{P}_X < \infty$. Then,*

$$\mathbb{E}[g(X)] = \int g f_X d\lambda.$$

In particular, if $g(x) = x$, i.e. the identity map, we have

$$\mathbb{E}[X] = \int x f_X d\lambda.$$

Proof: Let us first consider the case of g being a simple function i.e. $g = \sum_{i=1}^{K} a_i I_{A_i}$ for some measurable disjoint subsets A_i over the real line. We then have

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int g \; d\mathbb{P}_X \\
&= \sum_{i=1}^{K} a_i \mathbb{P}_X(A_i) && [g \text{ is a simple function}] \\
&= \sum_{i=1}^{K} a_i \int_{A_i} f_X d\lambda && [\text{From Radon-Nikodym Theorem}] \\
&= \sum_{i=1}^{K} \int_{A_i} a_i \; f_X d\lambda && [a_i \text{ is a constant}] \\
&= \sum_{i=1}^{K} \int_{\Omega} (a_i I_{A_i} f_X) d\lambda && [I_{A_i} \text{ is the indicator random variable of event } A_i]
\end{aligned}$$

$$= \int_\Omega \sum_{i=1}^{K} (a_i I_{A_i} f_X) d\lambda \quad \text{[Interchanging finite summation and integral]}$$
$$= \int_\Omega \left(\sum_{i=1}^{K} (a_i I_{A_i}) \right) f_X d\lambda$$
$$= \int_\Omega (g f_X) d\lambda.$$

Thus we have proved the above theorem for simple functions. We now assume g to be a non-negative measurable function which may not necessarily be simple.

Let g_n be an increasing sequence of non-negative simple functions that converge to g point wise. One way of coming up with such a sequence was discussed in the previous lecture. We then have,

$$\mathbb{E}[g(X)] = \lim_{n\to\infty} \int g_n d\mathbb{P}_X \quad \text{[From MCT]}$$
$$= \lim_{n\to\infty} \int g_n f_X d\lambda \quad \text{[From result for simple functions]}$$
$$= \int g f_X d\lambda. \quad \text{[From MCT, since } g_n f_X \uparrow g f_X \text{]}$$

For arbitrary g which are absolutely integrable, a similar proof can be worked out by writing $g = g_+ - g_-$ and proceeding. ■

Example 1: Let X be an exponential random variable with parameter μ. Find $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.
Solution: Recall that for an exponential random variable with parameter μ, $f_X(x) = \mu e^{-\mu x}$. Thus, we have

$$\mathbb{E}[X] = \int x f_X d\lambda = \int_0^\infty x \mu e^{-\mu x} dx = \frac{1}{\mu}.$$

$$\mathbb{E}[X^2] = \int x^2 f_X d\lambda = \int_0^\infty x^2 \mu e^{-\mu x} dx = \frac{2}{\mu^2}.$$

Example 2: Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Find $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.
Solution: Recall that the density of X is given by $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Thus, we have

$$\mathbb{E}[X] = \int x f_X d\lambda = \int_{-\infty}^\infty x \ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu.$$

$$\mathbb{E}[X^2] = \int x^2 f_X d\lambda = \int_{-\infty}^\infty x^2 \ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \mu^2 + \sigma^2.$$

Example 3: Let X be a one-sided Cauchy random variable i.e. $f_X(x) = \frac{2}{\pi}\frac{1}{1+x^2}$ for $x \geq 0$. Find $\mathbb{E}[X]$.
Solution: We have

$$\mathbb{E}[X] = \int x f_X d\lambda = \int_0^\infty x \ \frac{2}{\pi}\frac{1}{1+x^2} dx = \infty.$$

Example 4: Let X be a two-sided Cauchy random variable i.e., $f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ for $\forall x \in \mathbb{R}$. Find $\mathbb{E}[X]$.
Solution: In this case the random variable X takes both positive and negative values. Hence, we need to find $\mathbb{E}[X_+]$ and $\mathbb{E}[X_-]$ seperately and then evaluate $\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-]$. Recall that $X_+ = \max(X, 0)$ and $X_- = -\min(X, 0)$. Thus,

$$X_+(\omega) = 0 \text{ for } \omega \in A = \{\omega \in \Omega | X(\omega) < 0\},$$
$$X_+(\omega) = X(\omega) \text{ for } \omega \in A^c.$$

Similarly,

$$X_-(\omega) = 0 \text{ for } \omega \in B = \{\omega \in \Omega | X(\omega) > 0\},$$
$$X_-(\omega) = -X(\omega) \text{ for } \omega \in B^c.$$

It is easy to see that $\mathbb{P}(A) = \mathbb{P}(B) = 0.5$. Next, we have

$$\mathbb{E}[X_+] = \int x d\mathbb{P}_{X_+}$$
$$\mathbb{E}[X_+] = 0 \times \mathbb{P}(A) + \int_0^\infty x \ \frac{1}{\pi}\frac{1}{1+x^2} dx = \infty.$$

Similarly,

$$\mathbb{E}[X_-] = \int x d\mathbb{P}_{X_-}$$
$$\mathbb{E}[X_-] = 0 \times \mathbb{P}(B) + \int_{-\infty}^0 -x \ \frac{1}{\pi}\frac{1}{1+x^2} dx = \infty.$$

Thus, we have a case of $\infty - \infty$ and $\mathbb{E}[X]$ is undefined.

Note that in Example 2 also, X takes both positive and negative values and we should find $\mathbb{E}[X_+]$ and $\mathbb{E}[X_-]$ seperately and evaluate $\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-]$. But in that case both $\mathbb{E}[X_+]$ and $\mathbb{E}[X_-]$ are finite, allowing us to integrate with respect to the pdf $f_X(x)$ from $-\infty$ to ∞ directly.

Note: For the two sided Cauchy,

$$\int_{-\infty}^{\infty} \frac{1}{\pi}\frac{1}{1+x^2} dx \ \triangleq \ \lim_{\substack{M_1 \to -\infty \\ M_2 \to \infty}} \int_{M_1}^{M_2} \frac{1}{\pi}\frac{1}{1+x^2} dx.$$

The above limit does not exist and hence the integral is not defined.

21.2 Fatou's Lemma

Before we state Fatou's lemma, let us motivate it with an elementary result.

Lemma 21.2 *Let X and Y be random variables. Then,*

$$\mathbb{E}\left[\min(X,Y)\right] \leq \min\left(\mathbb{E}[X], \mathbb{E}[Y]\right).$$

$$\mathbb{E}\left[\max(X,Y)\right] \geq \max\left(\mathbb{E}[X], \mathbb{E}[Y]\right).$$

Proof: By definition, we have

$$\min(X,Y) \leq X.$$
$$\min(X,Y) \leq Y.$$

Taking expectations on both the sides,

$$\mathbb{E}\left[\min(X,Y)\right] \leq \mathbb{E}[X].$$
$$\mathbb{E}\left[\min(X,Y)\right] \leq \mathbb{E}[Y].$$

Combining the above two equations, we get

$$\mathbb{E}\left[\min(X,Y)\right] \leq \min\left(\mathbb{E}[X], \mathbb{E}[Y]\right).$$

The other statment of the lemma involving maximum of X and Y can be proved in a similar way and is left to the reader as an exercise.

■

The above lemma can be generalized to any finite collection of random variables and a similar proof can be worked out. Fatou's Lemma generalizes this idea for a sequence of random variables.

Lemma 21.3 ***Fatou's Lemma:*** *Let Y be a random variable that satisfies $\mathbb{E}[|Y|] < \infty$. Then the following holds,*

- *If $Y \leq X_n$, for all n, then $\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}[X_n]$.*
- *If $Y \geq X_n$, for all n, then $\mathbb{E}\left[\limsup_{n\to\infty} X_n\right] \geq \limsup_{n\to\infty} \mathbb{E}[X_n]$.*

Proof: Let us start by proving the first statement. For some n we have

$$\inf_{k\geq n} X_k - Y \leq X_m - Y, \qquad \forall\ m \geq n.$$

Taking expectations,

$$\mathbb{E}\left[\inf_{k\geq n} X_k - Y\right] \leq \mathbb{E}[X_m - Y], \qquad \forall\ m \geq n.$$

Taking infimum with respect to m on R.H.S, we obtain

$$\mathbb{E}\left[\inf_{k\geq n} X_k - Y\right] \leq \inf_{m\geq n} \mathbb{E}[X_m - Y], \qquad \forall\ m \geq n.$$

Let $Z_n = \inf_{k\geq n} X_k - Y$. Note that $Z_n \geq 0$ since $X_m \geq Y\ \forall\ m$ and Z_n is a non-decreasing sequence of random variables.

Also, $Z = \lim_{n\to\infty} Z_n = \liminf_{n\to\infty} X_n - Y$. By MCT, we have

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n - Y\right] \leq \liminf_{n\to\infty} \mathbb{E}[X_n - Y].$$

As $\mathbb{E}[|Y|] < \infty$, we can invoke linearity of expectation to get the first result of Fatou's lemma.

The second statement can be proved similarly and is left to the reader as an exercise.

■

21.3 Dominated Convergence Theorem

The DCT is an important result which asserts a sufficient condition under which we can interchange a limit and integral.

Theorem 21.4 *Consider a sequence of random variables X_n that converges almost surely to X. Suppose there exists a random variable Y such that $|X_n| \leq Y$ almost surely for all n and $\mathbb{E}[Y] < \infty$. Then, we have*

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

Proof: We have $|X_n| \leq Y$ which implies $-Y \leq X_n \leq Y$. We can now apply Fatou's lemma to obtain

$$\mathbb{E}[X] = \mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}[X_n] \leq \limsup_{n\to\infty} \mathbb{E}[X_n] \leq \mathbb{E}\left[\limsup_{n\to\infty} X_n\right] = \mathbb{E}[X].$$

Thus, all the inequalities in the above equation must be met with equalities and we have

$$\mathbb{E}[X] = \liminf_{n\to\infty} \mathbb{E}[X_n] = \limsup_{n\to\infty} \mathbb{E}[X_n],$$

which proves that the limit, $\lim_{n\to\infty} \mathbb{E}[X_n]$ exists and is given by

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

Thus we see that Dominated Convergence theorem (DCT) is a direct consequence of Fatou's Lemma. The name "dominated" is intuitive because we need $|X_n|$ to be bounded by some random variable Y almost surely for every n. However, we do not require X_n's to be monotonically increasing as in the case of MCT.

■

Corollary 21.5 *A special case of DCT is known as Bounded Convergence theorem (BCT). Here, the random variable Y is taken to be a constant random variable. BCT states that if there exists a constant $c \in \mathbb{R}$ such that $|X_n| \leq c$ almost surely for all n, then $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$.*

21.4 Exercise

1. *[MIT OCW problem set]* A workstation consists of three machines, M_1, M_2 and M_3 , each of which will fail after an amount of time T_i which is an independent exponentially distributed random variable, with parameter 1. Assume that the times to failure of the different machines are independent. The workstation fails as soon as both of the following have happened:

   (a) Machine M_1 has failed.

   (b) Atleast one of the machines M_2 or M_3 has failed.

   Find the expected value of the time to failure of the workstation.

2. *[Assignment problem, University of Cambridge]* Let Z be an exponential random variable with parameter $\lambda = 1$ and $Z_{int} = \lfloor Z \rfloor$. Compute $\mathbb{E}[Z_{int}]$.

3. *[Prof. Pollak, Purdue University]* Suppose S_k and S_n are the prices of a financial instrument on days k and n, respectively. For $k < n$, the gross return $G_{k,n}$ between days k and n is defined as $G_{k,n} = \frac{S_n}{S_k}$ and is equal to the amount of money you would have on day n if you invested \$1 on day k. Let $G_{k,k+1}$ be lognormal random variable with parameters μ and σ^2, $\forall k \geq 1$, and the random variables $G_{j,j+1}$ and $G_{k,k+1}$ are independent and identically distributed $\forall k \neq j$. Find the expected total gross return from day 1 to day n.

EE5110 : Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 22: Variance and Covariance

Lecturer: Dr. Krishna Jagannathan *Scribes: R.Ravi Kiran*

In this lecture, we will introduce the notions of variance and covariance of random variables. Earlier, we learnt about the expected value of a random variable which gives an idea of the average value. The idea of variance is useful in describing the extent to which the random variable deviates about its mean on either side. The covariance is a property that characterizes the extent of dependence between two random variables.

22.1 Variance

As stated earlier, the variance quantifies the extent to which the random variable deviates about the mean. Mathematically, the variance is defined as follows :

Definition 22.1 *Let X be a random variable with $\mathbb{E}[X] < \infty$. The **variance** of X is defined as*

$$Var(X) = \sigma_X^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

*σ_X is referred to as the **standard deviation** of the random variable X.*

22.1.1 Properties of Variance

We will now study a few properties of the variance of a random variable.

First and foremost, we can clearly see that for any real valued random variable X, $g(X) = (X - \mathbb{E}[X])^2 \geq 0$. Thus it is easy to see that $\sigma_X^2 \geq 0$ from property **PAI 2** from Lecture #18. In fact, we can make the following stronger statement regarding the variance of a random variable.

Lemma 22.2 *Let X be a real valued random variable. Then, $Var(X) = 0$ if and only if X is a constant almost surely.*

Proof: We will first prove the sufficiency criterion in the above statement. That is, assume that X is a constant valued random variable almost surely. Thus, it is evident that $X = \mathbb{E}[X]$ almost surely, consequently implying that $\sigma_X^2 = 0$.

To prove the necessity condition in the statement, assume that X is a random variable with zero variance. Thus, we have the following :

$$\begin{aligned} \sigma_X^2 = \quad & \mathbb{E}[(X - \mathbb{E}[X])^2] \quad = 0. \\ \implies \quad & \int (X - \mathbb{E}[X])^2 d\mathbb{P}_X \quad = 0. \end{aligned} \tag{22.1}$$

Applying **PAI 7** from Lecture #18 to (22.1), we can conclude that $(X - \mathbb{E}[X])^2 = 0$ almost surely. Thus, we have $X = \mathbb{E}[X]$ almost surely.

■

Now, using some simple algebra, we make a few useful observations.

$$\begin{aligned}\sigma_X^2 &= \mathbb{E}\left[(X-\mathbb{E}[X])^2\right], \\ &= \mathbb{E}\left[(X^2)+(\mathbb{E}[X])^2-2X\mathbb{E}[X]\right], \\ &\overset{(a)}{=} \mathbb{E}\left[X^2\right]-2\mathbb{E}[X].\mathbb{E}[X]+(\mathbb{E}[X])^2, \\ &= \mathbb{E}\left[X^2\right]-(\mathbb{E}[X])^2, \end{aligned} \tag{22.2}$$

where (a) follows from the linearity of expectation (**PAI 4** from Lecture #18). Now using the fact that $\sigma_X^2 \geq 0$ and (22.2), we can see that $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$. The term $\mathbb{E}\left[X^2\right]$ is referred to as the **second moment** of the random variable X.

An interesting digression:

Theorem 22.3 (Jensen's Inequality) *Let X be any real valued random variable and let $h(\cdot)$ be a function of the random variable. Then,*

1. *If $h(\cdot)$ is convex, then $\mathbb{E}[h(X)] \geq h(\mathbb{E}[X])$.*
2. *If $h(\cdot)$ is concave, then $\mathbb{E}[h(X)] \leq h(\mathbb{E}[X])$.*
3. *If $h(\cdot)$ is linear, then $\mathbb{E}[h(X)] = h(\mathbb{E}[X])$.*

A guided proof of Jensen's inequality will be encountered in your homework.

Since $f(x) = x^2$ is a convex function, we can invoke Theorem 22.3 and observe that $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$.

Let us look at a few examples.

Example 1: Let X be a Bernoulli random variable with parameter p i.e.,

$$X = \begin{cases} 1 & \text{w.p. } p, \\ 0 & \text{w.p. } 1-p. \end{cases}$$

Find the variance of X.
Solution: We have

$$\begin{aligned}\mathbb{E}[X] &= p \times 1 + (1-p) \times 0, \\ &= p.\end{aligned}$$

Next,

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

Finally,

$$\begin{aligned}\sigma_X^2 &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \\ &= p - p^2, \\ &= p(1-p).\end{aligned}$$

Example 2: Let X be a discrete valued random variable with Poisson distribution of parameter λ. That is, $\mathbb{P}(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}, \forall k \in \mathbb{Z}^+ \cup \{0\}$. Find the variance of X.
Solution: We have

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

Next,

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=0}^{\infty} k^2\frac{e^{-\lambda}\lambda^k}{k!},\\ &= \sum_{k=1}^{\infty}\frac{\lambda(k-1+1)e^{-\lambda}\lambda^{(k-1)}}{(k-1)!},\\ &= \sum_{k=2}^{\infty}\lambda^2\frac{e^{-\lambda}\lambda^{(k-2)}}{(k-2)!}+\sum_{k=1}^{\infty}\lambda\frac{e^{-\lambda}\lambda^{(k-1)}}{(k-1)!},\\ &= \lambda^2+\lambda.\end{aligned}$$

Finally,

$$\begin{aligned}\sigma_X^2 &= \mathbb{E}[X^2]-\left(\mathbb{E}[X]\right)^2,\\ &= \lambda^2+\lambda-\left(\lambda\right)^2,\\ &= \lambda.\end{aligned}$$

Example 3: Let X be a discrete random variable with $\mathbb{P}(X=k)=\frac{1}{\zeta(3)}\frac{1}{k^3}$ for $k \in \mathbb{N}$, where $\zeta(\cdot)$ is the Riemann zeta function. Find σ_X^2.
Solution: We have,

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^{\infty} k\mathbb{P}(X=k),\\ &= \sum_{k=1}^{\infty} k\frac{1}{\zeta(3)}\frac{1}{k^3},\\ &= \frac{1}{\zeta(3)}\sum_{k=1}^{\infty}\frac{1}{k^2},\\ &= \frac{1}{\zeta(3)}\frac{\pi^2}{6}.\end{aligned}$$

Next, we have

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=1}^{\infty} k^2 \mathbb{P}(X=k),\\ &= \sum_{k=1}^{\infty} k^2 \frac{1}{\zeta(3)} \frac{1}{k^3},\\ &= \frac{1}{\zeta(3)} \sum_{k=1}^{\infty} \frac{1}{k},\\ &= \infty.\end{aligned}$$

Finally,

$$\begin{aligned}\sigma_X^2 &= \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2,\\ &= \infty.\end{aligned}$$

The above example is a case of a random variable with finite expected value but infinite variance!

Example 4: Let X be a uniform random variable in the interval $[a, b]$. Find the variance of X.
Solution: Recall that the density of X is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b,\\ 0 & \text{otherwise .}\end{cases}$$

Now, we have

$$\begin{aligned}\mathbb{E}[X] &= \int x f_X(x) dx,\\ &= \int_a^b x \ \frac{1}{b-a} dx,\\ &= \frac{a+b}{2}.\end{aligned}$$

Next, we have

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \int x^2 f_X(x) dx,\\ &= \int_a^b x^2 \ \frac{1}{b-a} dx,\\ &= \frac{(b^3-a^3)}{3(b-a)},\\ &= \frac{a^2+ab+b^2}{3}.\end{aligned}$$

Finally,

$$\begin{aligned}\sigma_X^2 &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \\ &= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4}, \\ &= \frac{b^2-2ab+a^2}{12}, \\ &= \frac{(b-a)^2}{12}.\end{aligned}$$

Example 5: Let X be an exponentially distributed random variable with parameter μ. Find σ_X^2.
Solution: Recall that for an exponential random parameter $f_X(x) = \mu e^{-\mu x}$ for $x \geq 0$.

$$\begin{aligned}\mathbb{E}[X] &= \int x f_X(x) dx, \\ &= \int_0^\infty x \ \mu e^{-\mu x} dx, \\ &= \frac{1}{\mu}.\end{aligned}$$

Next, we have

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \int x^2 f_X(x) dx, \\ &= \int_0^\infty x^2 \ \mu e^{-\mu x} dx, \\ &= \frac{2}{\mu^2}.\end{aligned}$$

Finally,

$$\begin{aligned}\sigma_X^2 &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \\ &= \frac{2}{\mu^2} - \left(\frac{1}{\mu}\right)^2, \\ &= \frac{1}{\mu^2}.\end{aligned}$$

Example 6: Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Find σ_X^2.
Solution: From Example 2 in Lecture #21, we know that $\mathbb{E}[X] = \mu$ and $\mathbb{E}\left[X^2\right] = \mu^2 + \sigma^2$. Thus, we have

$$\begin{aligned}\sigma_X^2 &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2, \\ &= \mu^2 + \sigma^2 - (\mu)^2, \\ &= \sigma^2.\end{aligned}$$

Note that the normal distribution is parametrized by the expected value μ and the variance σ^2.

22.2 Covariance

Having looked at variance, a term that characterizes the extent of deviation of a single random variable around its expected value, we now define and study the covariance of two random variables X and Y, a term that quantifies the extent of dependence between the two random variables.

Definition 22.4 *Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Further, let $\mathbb{E}[X] < \infty$ and $\mathbb{E}[Y] < \infty$. The **covariance** of X and Y is given by*

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

Definition 22.5 *Let X and Y be random variables. X and Y are said to be **uncorrelated** if $cov(X,Y) = 0$, i.e, if $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

Thus, two random variables are uncorrelated if the expectation of their product is the product of their expectations. The following theorem asserts that independent random variables are uncorrelated.

Theorem 22.6 *If X and Y are independent random variables with $\mathbb{E}[|X|] < \infty$, $\mathbb{E}[|Y|] < \infty$. Then $\mathbb{E}[XY]$ exists, and $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ i.e., $cov(X,Y) = 0$.*

Proof: We will prove this theorem in three steps. We will first assume that the random variables X and Y are simple and thus can be represented as follows :

$$X = \sum_{i=1}^{n} x_i \mathbb{I}_{A_i} \quad \text{and} \quad Y = \sum_{i=1}^{m} y_i \mathbb{I}_{B_i}.$$

Assuming canonical representation of the random variables X and Y, we have

$$XY = \sum_{i=1}^{n}\sum_{j=1}^{m} (x_i y_j) \mathbb{I}_{(A_i \cap B_j)}.$$

Thus, we have,

$$\begin{aligned} \mathbb{E}[XY] &= \int XY \, d\mathbb{P}, \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} (x_i y_j) \mathbb{P}(A_i \cap B_j). \end{aligned} \tag{22.3}$$

Next, as X and Y are independent random variables, $\sigma(X)$ and $\sigma(Y)$ are independent $\sigma-$algebras. Also, $A_i = \{\omega \in \Omega | X(\omega) = a_i\} \in \sigma(X)$ and $B_j = \{\omega \in \Omega | X(\omega) = b_j\} \in \sigma(Y)$. By definition of independent $\sigma-$algebras,

$$\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)\mathbb{P}(B_j), \ \forall i, j. \tag{22.4}$$

Using (22.4) in (22.3), we get

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j \mathbb{P}(A_i)\mathbb{P}(B_j), \\ &= \left(\sum_{i=1}^{n} x_i \mathbb{P}(A_i)\right)\left(\sum_{j=1}^{m} y_j \mathbb{P}(B_j)\right), \\ &= \mathbb{E}[X]\mathbb{E}[Y]. \end{aligned}$$

We will now extend the proof to non-negative random variables. Let X and Y be non-negative random variables. Let the sequences of simple random variables, X_n and Y_n, be such that $X_n \uparrow X$ and $Y_n \uparrow Y$. We know that such a sequence exists from section 3 in Lecture #19. Also, by construction, it is easy to see that X_n and Y_n are independent. Consequently, we have $X_nY_n \uparrow XY$. Thus,

$$\mathbb{E}[XY] \overset{MCT}{=} \lim_{n\to\infty} \mathbb{E}[X_nY_n] \overset{(a)}{=} \left(\lim_{n\to\infty} \mathbb{E}[X_n]\right)\left(\lim_{n\to\infty} \mathbb{E}[Y_n]\right) \overset{MCT}{=} \mathbb{E}[X]\mathbb{E}[Y], \tag{22.5}$$

where (a) follows from the independence of X_n and Y_n and since both the limits exist.

Finally, for the case of X and Y possibly being negative, let $X = X_+ - X_-$, and let $Y = Y_+ - Y_-$ where X_+, X_-, Y_+ and Y_- are as defined in Lecture #17. Then

$$\begin{aligned}
\mathbb{E}[XY] &= \mathbb{E}[X_+Y_+] + \mathbb{E}[X_-Y_-] - \mathbb{E}[X_+Y_-] - \mathbb{E}[X_-Y_+], &(22.6)\\
&= \mathbb{E}[X_+]\mathbb{E}[Y_+] + \mathbb{E}[X_-]\mathbb{E}[Y_-] - \mathbb{E}[X_+]\mathbb{E}[Y_-] - \mathbb{E}[X_-]\mathbb{E}[Y_+], &(22.7)\\
&= \mathbb{E}[X]\mathbb{E}[Y]. &(22.8)
\end{aligned}$$

where (22.6) and (22.8) follow from the linearity of expectations (**PAI 4** from Lecture #18) and (22.7) follows from (22.5). Note that X_+ and X_- are functions of X, and Y_+ and Y_- are functions of Y. Since X and Y are independent, all the pairs of random variables inside expectation in RHS of (22.6) are independent.[1] Thus, we have proved that independent random variables are uncorrelated.

■

Caution: While independence guarantees that two random variables are uncorrelated, the converse is not necessarily true i.e., two uncorrelated random variables may or may not be independent. We show this by a counter example.

Let $X \sim unif[-1, 1]$ and $Y = X^2$ be two random variables. It can be shown that X and Y are not independent. However,

$$\begin{aligned}
cov(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],\\
&= \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\mathbb{E}\left[X^2\right],\\
&\overset{(a)}{=} 0 - 0,\\
&= 0,
\end{aligned}$$

where (a) follows since X is symmetric around 0.
Thus, we have an example where two random variables X and Y are uncorrelated but not independent.

Proposition 22.7 *Consider two random variables X and Y. Then, we have*

$$Var(X + Y) = Var(X) + Var(Y) + 2cov(X, Y).$$

Proof:

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

[1] Let X and Y be independent random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Also, let $f(\cdot)$ and $g(\cdot)$ be measurable functions from $\mathbb{R}$ to $\mathbb{R}$. Then, $f(X)$ and $g(Y)$ are independent random variables.

■

It is easy to see that if X and Y are uncorrelated, then $Var(X+Y) = Var(X) + Var(Y)$. This can of course be extended to the sum of any finite number of random variables.

Definition 22.8 *Let X and Y be random variables. Then, the **correlation coefficient** for the two random variables is defined as :*

$$\rho_{X,Y} \triangleq \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Theorem 22.9 ***Cauchy-Schwartz Inequality*** *For any two random variables X and Y, $-1 \leq \rho_{X,Y} \leq 1$. Further, if $\rho_{X,Y} = 1$, then there exists $a > 0$ such that $Y - \mathbb{E}[Y] = a\,(X - \mathbb{E}[X])$ and if $\rho_{X,Y} = -1$, then there exists $a < 0$ such that $Y - \mathbb{E}[Y] = a\,(X - \mathbb{E}[X])$.*

Proof: Let $\widetilde{X} = X - \mathbb{E}[X]$ and $\widetilde{Y} = Y - \mathbb{E}[Y]$. Now we know that,

$$\mathbb{E}\left[\left(\widetilde{X} - \frac{\mathbb{E}[\widetilde{X}\widetilde{Y}]}{\mathbb{E}[\widetilde{Y}^2]}\widetilde{Y}\right)^2\right] \overset{(a)}{\geq} 0, \tag{22.9}$$

$$\mathbb{E}\left[\widetilde{X}^2 - 2\widetilde{X}\frac{\mathbb{E}[\widetilde{X}\widetilde{Y}]}{\mathbb{E}[\widetilde{Y}^2]}\widetilde{Y} + \frac{\left(\mathbb{E}[\widetilde{X}\widetilde{Y}]\right)^2}{\left(\mathbb{E}[\widetilde{Y}^2]\right)^2}\widetilde{Y}^2\right] \geq 0,$$

$$\mathbb{E}\left[\widetilde{X}^2\right] - \frac{(\mathbb{E}[\widetilde{X}\widetilde{Y}])^2}{\mathbb{E}[\widetilde{Y}^2]} \overset{(b)}{\geq} 0,$$

$$\mathbb{E}\left[\widetilde{X}^2\right] \geq \frac{(\mathbb{E}[\widetilde{X}\widetilde{Y}])^2}{\mathbb{E}[\widetilde{Y}^2]},$$

$$-1 \leq \frac{\mathbb{E}\left[\widetilde{X}\widetilde{Y}\right]}{\sqrt{\mathbb{E}\left[\widetilde{X}^2\right]}\sqrt{\mathbb{E}\left[\widetilde{Y}^2\right]}} \leq 1, \tag{22.10}$$

where (a) follows from **PAI 2** of Lecture #18 and (b) follows from linearity and scaling property of expectation (**PAI 4** and **PAI 8** of Lecture #18). From definition, $\mathbb{E}[\widetilde{X}^2] = Var(X)$ and $\mathbb{E}[\widetilde{Y}^2] = Var(Y)$. Further, we can observe that $\mathbb{E}[\widetilde{X}\widetilde{Y}] = cov(X,Y)$.Thus, it is easy to see that

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\mathbb{E}\left[\widetilde{X}\widetilde{Y}\right]}{\sqrt{\mathbb{E}\left[\widetilde{X}^2\right]}\sqrt{\mathbb{E}\left[\widetilde{Y}^2\right]}}\,. \tag{22.11}$$

Combining (22.10) and (22.11), we get

$$-1 \leq \rho_{X,Y} \leq 1.$$

Note that $\rho_{X,Y} = 1$ or $\rho_{X,Y} = -1$ when the (22.9) is met with equality. This happens when $\widetilde{X} = \frac{\mathbb{E}[\widetilde{X}\widetilde{Y}]}{\mathbb{E}[\widetilde{Y}^2]}\widetilde{Y}$ almost surely which proves the second part of the theorem.

■

The discussion regarding Cauchy-Schwartz inequality above has a close connection with Hilbert Spaces. As one may recall from a course in Linear Algebra, a Hilbert Space is a complete vector space endowed with an inner product.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{L}_2$ be a collection of all zero-mean, real-valued random variables defined over this space with finite second moment. It can be shown that $\mathcal{L}_2$ with addition of functions and scalar multiplication (obeyed except perhaps on a set of measure zero) is a Hilbert Space. The associated inner product is the covariance function. We say that two random variables from $\mathcal{L}_2$ are *equivalent* if they agree, except perhaps on a set of measure zero. That is, $X \sim Y$ (read as X is equivalent to Y) if $\mathbb{P}(X = Y) = 1$, for any $X, Y \in \mathcal{L}_2$. Thus, $\mathcal{L}_2$ is partitioned into several such equivalence classes by the aforementioned equivalence relation.

In light of this discussion, the covariance function can be interpreted as the dot product of the Hilbert space, and the correlation coefficient is interpreted as the cosine of the angle between two random variables in this Hilbert space. In particular uncorrelated random variables are orthogonal! The interested reader is referred to sections 7 through 11 of chapter 6 in [1] for a more detailed treatment of this topic; this viewpoint is especially useful in estimation theory.

22.3 Exercise

1. *[Papoulis]* Let a and b be positive integers with $a \leq b$, and let X be a random variable that takes as values, with equal probability, the powers of 2 in the interval $[2^a, 2^b]$. Find the expected value and variance of X.

2. *[Papoulis]* Suppose that X and Y are random variables with the same variance. Show that $X - Y$ and $X + Y$ are uncorrelated.

3. *[Papoulis]* Suppose that a random variable X satisfies $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] = 1$, $\mathbb{E}[X^3] = 0$ and $\mathbb{E}[X^4] = 3$ and let $Y = a + bX + cX^2$. Find the correlation co-efficient $\rho_{X,Y}$.

4. *[Assignment problem, University of Cambridge]* Take $0 \leq r \leq 1$. Let X and Y be independent random variables taking values ± 1 with probabilities $\frac{1}{2}$. Set $Z = X$, with probability r and $Z = Y$, with probability $1 - r$. Find $\rho_{X,Z}$.

5. *[Papoulis]* Let $X_1, X_2, ..., X_n$ be independent random variables with non-zero finite expectations. Show that
$$\frac{\text{var}(\prod_{i=1}^{n} X_i)}{\prod_{i=1}^{n} \mathbb{E}[X_i]^2} = \prod_{i=1}^{n} \Big(\frac{\text{var}(X_i)}{\mathbb{E}[X_i]^2} + 1 \Big) - 1$$
.

References

[1] David Williams, "Probability with Martingales", *Cambridge University Press*, Fourteenth Printing, 2011.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 23: Conditional Expectation

Lecturer: Dr. Krishna Jagannathan *Scribe: Sudharsan Parthasarathy*

Let X and Y be discrete random variables with joint probability mass function $p_{X,Y}(x,y)$, then the conditional probability mass function was defined in previous lectures as

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

assuming $p_Y(y) > 0$. Let us define

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

$\psi(y) = \mathbb{E}[X|Y=y]$ changes with y. The random variable $\psi(Y)$ is the conditional expectation of X given Y and denoted as $\mathbb{E}[X|Y]$.

Let X and Y be continuous random variables with joint probability density function $f_{X,Y}(x,y)$. Recall the conditional probability density function

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

when $f_Y(y) > 0$. Define

$$\mathbb{E}[X|Y=y] = \int_x x f_{X|Y}(x|y)\mathrm{d}x.$$

The random variable $\psi(Y)$ is the conditional expectation of X given Y and denoted as $\mathbb{E}[X|Y]$.

Example 1: Find $\mathbb{E}[Y|X]$ if the joint probability density function is $f_{X,Y}(x,y)=\frac{1}{x}$; $0 < y \leq x \leq 1$.

Solution: $f_X(x)=\int_0^x \frac{1}{x}dy = 1$, $0 \leq x \leq 1$

$f_{Y|X}(y|x)=\frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{x}$, $0 < y \leq x$

$\mathbb{E}[Y|X=x] = \int_0^x y f_{Y|X}(y|x)\mathrm{d}y = \int_0^x \frac{y}{x}\mathrm{d}y = \frac{x}{2}$

The conditional expectation $\mathbb{E}[Y|X] = \frac{X}{2}$.

Theorem 23.1 *Law of Iterated Expectation:*

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

Proof: We prove the result for discrete random variables. We have

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

■

Similarly law of iterated expectation for jointly continuous random variables can also be proved.

Application of the law of iterated expectation:

$S_N = \sum_{i=1}^{N} X_i$, where $\{X_1, ...X_N\}$ are independent and identically distributed random variables. N is a non-negative random variable independent of $X_i \ \forall \ i \in \{1, ..N\}$. From the law of iterative expectation, $\mathbb{E}[S_N] = \mathbb{E}_N[\mathbb{E}[S_N|N]]$. Consider

$$\mathbb{E}[S_N|N=n] = \mathbb{E}\left[\sum_{i=1}^{N} X_i | N=n\right] \tag{23.1}$$

$$= \mathbb{E}\left[\sum_{i=1}^{n} X_i | N=n\right]. \tag{23.2}$$

As N is independent of X_i, $\mathbb{E}\left[\sum_{i=1}^{n} X_i|N=n\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\mathbb{E}[X]$.

Thus $\mathbb{E}[S_N|N] = N\mathbb{E}[X]$, $\mathbb{E}[S_N] = \mathbb{E}[N]\mathbb{E}[X]$.

Theorem 23.2 *Generalized form of Law of Iterated Expectation:*

For any measurable function g with $\mathbb{E}[|g(X)|] < \infty$,

$$\mathbb{E}[Yg(X)] = \mathbb{E}[\mathbb{E}[Y|X]g(X)].$$

Proof: We prove the result for discrete random variables. We have

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

■

Exercise: Prove $\mathbb{E}[Yg(X)] = \mathbb{E}[\mathbb{E}[Y|X]g(X)]$ if X and Y are jointly continuous random variables.

This theorem implies that

$$\mathbb{E}[(Y - \mathbb{E}[Y|X])g(X)] = 0. \tag{23.3}$$

The conditional expectation $\mathbb{E}[Y|X]$ can be viewed as an estimator of Y given X. $Y - \mathbb{E}(Y|X)$ is then the *estimation error* for this estimator. The above theorem implies that the estimation error is uncorrelated with every function of X.

Observe that in this lecture, we have not dealt with conditional expectations in a general framework. Instead, we have separately defined it for discrete and jointly continuous random variables. In a more general development of the topic, (23.3) is in fact taken as the defining property of the conditional expectation. Specifically, for any $g(X)$, one can prove the existence and uniqueness (up to measure zero) of a $\sigma(X)$-measurable random variable $\psi(X)$, that satisfies $\mathbb{E}[(\psi(X) - Y)g(X)] = 0$. Such a $\psi(X)$ is then defined as the conditional expectation $\mathbb{E}[Y|X]$. For a more detailed discussion, refer Chapter 9 in [1].

Minimum Mean Square Error Estimator:

We have seen that $\mathbb{E}[Y|X]$ is an estimator of Y given X. In the next theorem we will prove that this is indeed an optimal estimate of Y given X, in the sense that the conditional expectation minimizes the mean-squared error.

Theorem 23.3 *If $\mathbb{E}(Y^2) < \infty$, then for any measurable function g,*

$$\mathbb{E}[(Y - \mathbb{E}[Y|X])^2] \leq \mathbb{E}[(Y - g(X))^2].$$

Proof:

$$\begin{aligned}\mathbb{E}[(Y - g(X))^2] &= \mathbb{E}[(Y - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - g(X))^2] + 2\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - g(X))] \\ &\geq \mathbb{E}[(Y - \mathbb{E}[Y|X])^2].\end{aligned}$$

This is because $\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - g(X))] = 0$ (by (23.3)), and $\mathbb{E}[(\mathbb{E}[Y|X] - g(X))^2] \geq 0$.

$\mathbb{E}[(Y - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - g(X))] = 0$ as from (23.3) we know that $\mathbb{E}[(\mathbb{E}[Y|X] - Y)\psi(X)] = 0$. Here $\psi(X) = (\mathbb{E}[Y|X] - g(X))$. ■

From (23.3) we observe that the estimation error $Y - (\mathbb{E}[Y|X])$ is orthogonal to any measurable function of X. In the Hilbert Space of square integrable random variables, $\mathbb{E}[Y|X]$ can be viewed as the projection of Y onto the subspace $\mathcal{L}_2(\sigma(X))$ of $\sigma(X)$ measurable random variables. As depicted in Figure 23.1, it is quite intuitive that the conditional expectation (which is the projection of Y onto the subspace) minimizes the mean-squared error among all random variables from the subspace $\mathcal{L}_2(\sigma(X))$.

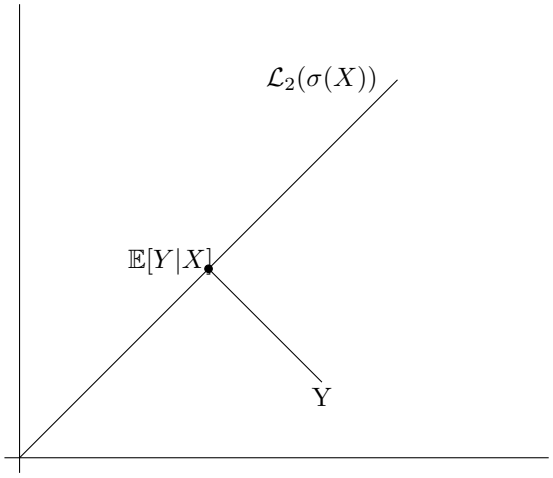

Figure 23.1: Geometric interpretation of MMSE

23.1 Exercises

1. Prove the law of iterated expectation for jointly continuous random variables.
2. (i) Given is the table for Joint PMF of random variables X and Y.

	X=0	X =1
Y=0	$\frac{1}{5}$	$\frac{2}{5}$
Y=1	$\frac{2}{5}$	0

   Let $Z = \mathbb{E}[X|Y]$ and $V = Var(X|Y)$. Find the PMF of Z and V, and compute $\mathbb{E}[Z]$ and $\mathbb{E}[V]$.

   (ii) Consider a sequence of i.i.d. random variables $\{Z_i\}$ where $\mathbb{P}(Z_i = 0) = \mathbb{P}(Z_i = 1) = \frac{1}{2}$. Using this sequence, define a new sequence of random variables $\{X_n\}$ as follows:
   $X_0 = 0$,
   $X_1 = 2Z_1 - 1$, and
   $X_n = X_{n-1} + (1 + Z_1 + ... + Z_{n-1})(2Z_n - 1)$ for $n \geq 2$.
   Show that $\mathbb{E}[X_{n+1}|X_0, X_1, ..., X_n] = X_n$ a.s. for all n.
3. (a) *[MIT OCW problem set]* The number of people that enter a pizzeria in a period of 15 minutes is a (nonnegative integer) random variable K with known moment generating function $M_K(s)$. Each person who comes in buys a pizza. There are n types of pizzas, and each person is equally likely to choose any type of pizza, independently of what anyone else chooses. Give a formula, in terms of $M_K(.)$, for the expected number of different types of pizzas ordered.

   (b) John takes a taxi to home everyday after work. Every evening, he waits by the road to get a taxi but every taxi that comes by is occupied with a probability 0.8 independent of each other . He counts the number of taxis he missed till he gets an unoccupied taxi. Once he gets inside the taxi, he throws a fair six faced die for a number of times equal to the number of taxis he missed. He counts the output of the die throws and gives a tip to the driver equal to that. Find the expected amount of tip that John gives everyday.

References

[1] D. Williams, "Probability with Martingales",*Cambridge University Press*, Fourteenth Printing, 2011.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 24: Probability Generating Functions

Lecturer: Dr. Krishna Jagannathan *Scribe: Debayani Ghosh*

24.1 Probability Generating Functions (PGF)

Definition 24.1 *Let X be an integer valued random variable. The probability generating function (PGF) of X is defined as :*

$$G_X(z) \triangleq \mathbb{E}[z^X] = \sum_i z^i \mathbb{P}(X = i).$$

24.1.1 Convergence

For a non-negative valued random variable, there exists R, possibly $+\infty$, such that the PGF converges for $|z| < R$ and diverges for $|z| > R$ where $z \in \mathbb{C}$. $G_X(z)$ certainly converges for $|z| < 1$ and possibly in a larger region as well. Note that,

$$|G_X(z)| = \left| \sum_i z^i \mathbb{P}(X = i) \right| \leq \sum_i |z|^i.$$

This implies that $G_X(z)$ converges absolutely in the region $|z| < 1$. Generating functions can be defined for random variables taking negative as well as positive integer values. Such generating functions generally converge for values of z satisfying $\alpha < |z| < \beta$ for some α, β such that $\alpha \leq 1 \leq \beta$.

Example 1 : Consider the Poisson random variable X with probability mass function

$$\mathbb{P}(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}, \quad i \geq 0.$$

Find the PGF of X.

Solution : The PGF of X is

$$G_X(z) = \sum_{i=1}^{\infty} \frac{z^i \lambda^i e^{-\lambda}}{i!} = e^{\lambda(z-1)}, \quad \forall z \in \mathbb{C}.$$

Example 2 : Consider the geometric random variable X with probability mass function

$$\mathbb{P}(X = i) = (1-p)^{i-1} p, \quad i \geq 1.$$

Find the PGF of X.

Solution : The PGF of X is

$$\begin{aligned} G_X(z) &= \sum_{i=1}^{\infty} (1-p)^{i-1} p z^i, \\ &= \frac{pz}{1 - z(1-p)}, \quad \text{if} \quad |z| < \frac{1}{1-p}. \end{aligned}$$

24.1.2 Properties

1. $G_X(1) = 1$.
2. $\left.\frac{\mathrm{d}G_X(z)}{\mathrm{d}z}\right|_{z=1} = \mathbb{E}[X]$.

   Proof : From definition

   $$G_X(z) = \mathbb{E}\left[z^X\right] = \sum_i z^i \mathbb{P}(X = i).$$

   Now,

   $$\begin{aligned}\frac{\mathrm{d}G_X(z)}{\mathrm{d}z} &= \frac{\mathrm{d}}{\mathrm{d}z}\sum_i z^i \mathbb{P}(X = i),\\ &\overset{(a)}{=} \sum_i \frac{\mathrm{d}}{\mathrm{d}z} z^i \mathbb{P}(X = i),\\ &= \sum_i i z^{i-1} \mathbb{P}(X = i),\end{aligned}$$

   where the interchange of differentiation and summation in (a) is a consequence of absolute convergence of the series $\sum_i z^i \mathbb{P}(X = i)$. Thus,

   $$\left.\frac{\mathrm{d}G_X(z)}{\mathrm{d}z}\right|_{z=1} = \mathbb{E}[X].$$

3. $\left.\frac{\mathrm{d}^k G_X(z)}{\mathrm{d}z^k}\right|_{z=1} = \mathbb{E}\left[X\left(X-1\right)\left(X-2\right)\cdots\left(X-k+1\right)\right]$.
4. If X and Y are independent and $Z = X + Y$, then $G_Z(z) = G_X(z)G_Y(z)$. The ROC for the PGF of z is the intersection of the ROCs of the PGFs of X and Y.
   Proof :

   $$G_Z(z) = \mathbb{E}[z^Z] = \mathbb{E}[z^{X+Y}] = \mathbb{E}[z^X.z^Y].$$

   Since X and Y are independent, they are uncorrelated. This implies that

   $$\mathbb{E}[z^X.z^Y] = \mathbb{E}[z^X]\mathbb{E}[z^Y] = G_X(z)G_Y(z).$$

   Hence proved.

5. **Random sum of discrete RVs** : Let $Y = \sum_{i=1}^{N} X_i$, where X_i's are i.i.d discrete positive integer valued random variables and N is independent of X_i's. The PGF of Y is $G_Y(z) = G_N(G_X(z))$.
   Proof :

   $$G_Y(z) = \mathbb{E}[z^Y] = \mathbb{E}\left[\mathbb{E}\left[z^Y|N\right]\right] \quad \text{(By law of iterated expectation).}$$

   Now,

   $$\mathbb{E}\left[z^Y|N = n\right] = \mathbb{E}\left[z^{\sum_i x_i}|N = n\right] = \mathbb{E}\left[G_X(z)^N\right].$$

   This implies that

   $$G_Y(z) = G_N\left(G_X(z)\right).$$

24.2 Exercise

1. Find the PMF of a random variable X whose probability generating function is given by

$$G_X(z) = \frac{(\frac{1}{3}z+\frac{2}{3})^4}{z}$$

2. Suppose there are X_0 individuals in initial generation of a population. In the n^{th} generation, the X_n individuals independently give rise to numbers of offspring $Y_1^{(n)}, Y_2^{(n)}, ..., Y_{X_n}^{(n)}$, where $Y_1^{(n)}, Y_2^{(n)}, ..., Y_{X_n}^{(n)}$ are i.i.d. random variables. The total number of individuals produced at the $(n+1)^{st}$ generation will then be $X_{n+1} = Y_1^{(n)} + Y_2^{(n)} + ... + Y_{X_n}^{(n)}$. Then, $\{X_n\}$ is called a branching process. Let X_n be the size of the n^{th} generation of a branching process with family-size probability generating function $G(z)$, and let $X_0 = 1$. Show that the probability generating function $G_n(z)$ of X_n satisfies $G_{n+1}(z) = G(G_n(z))$ for $n \geq 0$. Also, prove that $\mathbb{E}[X_n] = \mathbb{E}[X_{n-1}]G'(1)$.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 25: Moment Generating Function

Lecturer: Dr. Krishna Jagannathan *Scribe: Subrahmanya Swamy P*

In this lecture, we will introduce Moment Generating Function and discuss its properties.

Definition 25.1 *The moment generating function (MGF) associated with a random variable X, is a function, $M_X : \mathbb{R} \to [0, \infty]$ defined by $M_X(s) = \mathbb{E}\left[e^{sX}\right]$.*

The domain or region of convergence (ROC) of M_X is the set $D_X = \{s | M_X(s) < \infty\}$. In general, s can be complex, but since we did not define expectation of complex valued random variables, we will restrict ourselves to real valued s. Note that $s = 0$ is always a point in the ROC for any random variable, since $M_X(0) = 1$.

Cases:

- If X is discrete with pmf $p_X(x)$, then $M_X(s) = \sum_x e^{sx} p_X(x)$.
- If X is continuous with density $f_X(\cdot)$, then $M_X(s) = \int e^{sx} f_X(x)\, dx$.

Example 25.2 *Exponential random variable*

$$f_X(x) = \mu e^{-\mu x}, \qquad x \geq 0,$$

$$M_X(s) = \int_0^\infty e^{sx} \mu e^{-\mu x} dx = \begin{cases} \frac{\mu}{\mu - s}, & \text{if } s < \mu, \\ +\infty, & \text{otherwise.} \end{cases}$$

The Region of Convergence for this example is, $\{s | M_X(s) < \infty\}$, i.e., $s < \mu$.

Example 25.3 *Std. Normal random variable*

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \quad x \in \mathbb{R}, \\ M_X(s) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{\frac{-x^2}{2}} dx, \\ &= e^{\frac{s^2}{2}}, \quad s \in \mathbb{R}. \end{aligned}$$

The Region of Convergence for this example is the entire real line.

Example 25.4 *Cauchy random variable*

$$f_X(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R},$$

$$M_X(s) = \frac{1}{\pi} \int_{-\infty}^{\infty} e^{sx} \frac{1}{1 + x^2} dx = \begin{cases} 1, & \text{if } s = 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

The Region of Convergence for this example is just the point $s = 0$.

Remark 2: The above examples can be interpreted as follows.

- In Example 25.2, we have the product of two exponentials. Thus, the MGF converges when the product is decreasing.
- In Example 25.3, there is a 'competition' between $e^{-\frac{x^2}{2}}$ and e^{sx}. Since the first term from the Gaussian decreases faster than e^{sx} increases (for any s), the integral always converges.
- In Example 25.4, for $s \neq 0$, an exponential competes with a decreasing polynomial, as a result of which the integral diverges.

It is an interesting question whether or not we can uniquely find the CDF of a random variable, given the moment generating function and its ROC. A quick look at Example 25.4 reveals that if the MGF is finite only at $s = 0$ and infinite elsewhere, it is not possible to recover the CDF uniquely. To see this, one just needs to produce another random variable whose MGF is finite only at $s = 0$. (Do this!) On the other hand, if we can specify the value of the moment generating function even in a tiny interval, we can uniquely determine the density function. This result follows essentially because the MGF, when it exists in an interval, is *analytic*, and hence possesses some nice properties. The proof of the following theorem is rather involved, and uses the properties of an analytic function.

Theorem 25.5 *(Without Proof)*

i) Suppose $M_X(s)$ is finite in the interval $[-\epsilon, \epsilon]$ for some $\epsilon > 0$, then M_X uniquely determines the CDF of X.

ii) If X and Y are two random variables such that, $M_X(s) = M_Y(s) \ \ \forall s \in [-\epsilon, \epsilon], \epsilon > 0$ then X and Y have the same CDF.

25.1 Properties

1. $M_X(0) = 1$.
2. *Moment Generating Property:* We shall state this property in the form of a theorem.

   Theorem 25.6 *Supposing $M_X(s) < \infty$ for $s \in [-\epsilon, \epsilon]$, $\epsilon > 0$ then,*

   $$\frac{d}{ds} M_X(s)\Big|_{s=0} = \mathbb{E}[X]. \tag{25.1}$$

   More generally,

   $$\frac{d^m}{ds^m} M_X(s)\Big|_{s=0} = \mathbb{E}[X^m]\ ; \ \ m \geq 1.$$

   Proof: (25.1) can be proved in the following steps.

   $$\frac{d}{ds} M_X(s) = \frac{d}{ds}\mathbb{E}[e^{sX}] \overset{(a)}{=} \mathbb{E}[\frac{d}{ds} e^{sX}] = \mathbb{E}[Xe^{sX}],$$

   where, (a) is obtained by the interchange of the derivative and the expectation. This follows from the use of basic definition of the derivative, and then invoking the DCT; see Lemma 25.7 (d). ■

Lemma 25.7 *Suppose that X is a non-negative random variable and $M_X(s) < \infty$, $\forall s \in (-\infty, a]$, where a is a positive number, then*

(a) $\mathbb{E}[X^k] < \infty$, *for every* k.

(b) $\mathbb{E}[X^k e^{sX}] < \infty$, *for every* $s < a$.

(c) $\frac{e^{hX}-1}{h} \leq Xe^{hX}$.

(d) $\mathbb{E}[X] = \mathbb{E}[\lim_{h\downarrow 0} \frac{e^{hX}-1}{h}] = \lim_{h\downarrow 0} \frac{\mathbb{E}[e^{hX}]-1}{h}$.

Proof: Given that X is a non-negative random variable with a Moment Generating Function such that $M_X(s) < \infty$, $\forall s \in (-\infty, a]$, for some positive a.

(a) For a positive number a, $x^k \leq e^{ax}$, $\forall k \in \mathbb{Z}^+ \cup \{0\}$. Therefore, $\mathbb{E}[X^k] = \int x^k \mathrm{d}\mathbb{P}_X \leq \int e^{ax} \mathrm{d}\mathbb{P}_X$. However, $\int e^{ax} \mathrm{d}\mathbb{P}_X = M_X(a) < \infty$. Therefore, $\mathbb{E}[X^k] < \infty$.

(b) For s $< a$, $\exists \epsilon > 0$ such that $M_X(s+\epsilon) < \infty \Rightarrow \int e^{sx} e^{\epsilon x} \mathrm{d}\mathbb{P}_X < \infty$. But since $\epsilon > 0$, as $x \to \infty$, $x^k \leq e^{\epsilon x}$. Therefore, $\mathbb{E}[X^k e^{sX}] = \int x^k e^{sx} \mathrm{d}\mathbb{P}_X \leq \int e^{sx} e^{\epsilon x} \mathrm{d}\mathbb{P}_X < \infty \Rightarrow \mathbb{E}[X^k e^{sX}] < \infty$.

(c) To prove that $\frac{e^{hX}-1}{h} \leq Xe^{hX}$.
Let $hX = Y$. Therefore, re-arranging the terms, we need to prove that $e^Y - Ye^Y \leq 1$. Or equivalently, it is enough to prove that, $g(Y) = e^Y(Y-1) \geq -1$.
$g(Y)$ has a minima at $Y = 0$, and the minimum value, *i.e.,* $g(0)$ = -1.
$\Rightarrow g(Y) \geq -1$,
$\Rightarrow e^Y(Y-1) \geq -1$.
Hence proved.

(d) Define $X_h = \frac{e^{hX}-1}{h}$.
$\lim_{h\downarrow 0} X_h = X$ i.e. $X_h \to X$ point-wise. Since $\mathbb{E}[X^k e^{sX}] < \infty$ is true, when $s = h$ and $k = 1$, we get $\mathbb{E}[Xe^{hX}] < \infty$. Since X_h is dominated by Xe^{hX}, $\mathbb{E}[Xe^{hX}] < \infty$ and $\lim_{h\downarrow 0} X_h = X$, applying DCT we get $\mathbb{E}[X] = \mathbb{E}[\lim_{h\downarrow 0} X_h] = \mathbb{E}[\lim_{h\downarrow 0} \frac{e^{hX}-1}{h}] = \lim_{h\downarrow 0} \mathbb{E}\left[\frac{e^{hX}-1}{h}\right] = \lim_{h\downarrow 0} \frac{\mathbb{E}[e^{hX}]-1}{h}$. Therefore, $\mathbb{E}[X] = \mathbb{E}[\lim_{h\downarrow 0} \frac{e^{hX}-1}{h}] = \lim_{h\downarrow 0} \frac{\mathbb{E}[e^{hX}]-1}{h}$.
Hence proved.

■

3. If $Y = aX + b$, $a, b \in \mathbb{R}$, then $M_Y(s) = e^{sb} M_X(as)$. For example, $X \sim \mathcal{N}(0,1)$, $Y = \sigma X + \mu$ $\Rightarrow Y \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow M_Y(s) = e^{\mu s} e^{\sigma^2 \frac{s^2}{2}}$, $s \in \mathbb{R}$.

4. If X and Y are independent and $Z = X + Y$, then $M_Z(s) = M_X(s) M_Y(s)$.
Proof: $\mathbb{E}[e^{sZ}] = \mathbb{E}[e^{sX+sY}] = \mathbb{E}[e^{sX} e^{sY}] = \mathbb{E}[e^{sX}]\mathbb{E}[e^{sY}]$. ■

Consider the following examples:

(a) $X_1 \sim N(\mu_1, \sigma_1^2)$; $X_2 \sim N(\mu_2, \sigma_2^2)$; and X_1, X_2 are independent. $Z = X_1 + X_2$;

$$\begin{aligned} M_{X_1}(s) &= e^{\left(\mu_1 s + \frac{\sigma_1^2 s^2}{2}\right)}, \\ M_{X_2}(s) &= e^{\left(\mu_2 s + \frac{\sigma_2^2 s^2}{2}\right)}, \\ M_Z(s) &= M_{X_1}(s) M_{X_2}(s), \\ &= e^{\left((\mu_1+\mu_2)s + \frac{(\sigma_1^2+\sigma_2^2)s^2}{2}\right)}. \end{aligned}$$

$\Rightarrow Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

(b) $X_1 \sim \exp(\mu); X_2 \sim \exp(\lambda)$, $\lambda \neq \mu$ and X_1, X_2 are independent. $Z = X1 + X2$;

$$\begin{aligned}
M_{X_1}(s) &= \frac{\mu}{\mu - s}, \\
M_{X_2}(s) &= \frac{\lambda}{\lambda - s}, \\
M_Z(s) &= M_{X_1}(s) M_{X_2}(s), \\
&= \frac{\mu\lambda}{(\mu - s)(\lambda - s)}, \quad \text{ROC is } \ s < \min(\lambda, \mu) \\
\Rightarrow f_Z(x) &= \frac{\mu}{\mu - \lambda} \lambda e^{-\lambda x} - \frac{\lambda}{\mu - \lambda} \mu e^{-\mu x}, \\
&= \left(\frac{\mu\lambda}{\mu - \lambda} \right) \left(e^{-\lambda x} - e^{-\mu x} \right), \quad x \geq 0.
\end{aligned}$$

5. $Z = \sum_{i=1}^{N} X_i$, X_i are i.i.d and N is independent of X_i.

$$\begin{aligned}
M_Z(s) = \mathbb{E}[e^{sZ}] &= \mathbb{E}\left[\mathbb{E}\left[e^{sZ} | N \right] \right], \\
&= \mathbb{E}\left[\left(M_X(s) \right)^N \right],
\end{aligned}$$

If we write in terms of the PGF and MGF of N, then,

$$\begin{aligned}
M_Z(s) &= G_N(M_X(s)), \\
&= M_N(\log M_X(s)).
\end{aligned}$$

For example, $X_i \sim \exp(\mu); N \sim \text{Geom}(p)$ and $Z = \sum_{i=1}^{N} X_i$. Then the distribution of Z is computed as follows:

$$\begin{aligned}
M_X(s) &= \frac{\mu}{\mu - s}, \quad s < \mu, \\
G_N(\xi) &= \frac{p\xi}{1 - (1-p)\xi}, \quad |\xi| < \frac{1}{1-p}, \\
M_Z(s) &= G_N(M_X(s)), \\
&= \frac{p\left(\frac{\mu}{\mu - s} \right)}{1 - (1-p)\left(\frac{\mu}{\mu - s} \right)}, \\
&= \frac{\mu p}{\mu p - s}, \ \ s < \mu p, \\
\Rightarrow Z &\sim \exp(\mu p).
\end{aligned}$$

25.2 Exercise

1. (a) *[Dimitri P.Bertsekas]* Find the MGF associated with an integer-valued random variable X that is uniformly distributed in the range $\{a, a+1, ..., b\}$.

(b) *[Dimitri P.Bertsekas]* Find the MGF associated with a continuous random variable X that is uniformly distributed in the range $[a, b]$.

2. *[Dimitri P.Bertsekas]* A non-negative interger-valued random variable X has one of the following MGF:

   (a) $M(s) = e^{2(e^{e^s-1}-1)}$.

   (b) $M(s) = e^{2(e^{e^s}-1)}$.

   (a) Explain why one of the 2 cannot possibly be a MGF.

   (b) Use the true MGF to find $\mathbb{P}(X = 0)$.

3. Find the variance of a random variable X whose moment generating function is given by

$$M_X(s) = e^{3e^s-3}$$

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 26: Characteristic Functions

Lecturer: Dr. Krishna Jagannathan *Scribe: Aseem Sharma and Ajay M.*

The characteristic function of a random variable X is defined as

$$\begin{aligned} C_X(t) =& \mathbb{E}[e^{itX}] \\ =& \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)], \end{aligned}$$

which can also be written as

$$C_X(t) = \int e^{itx} d\mathbb{P}_X.$$

If X is a continuous random variable with density function $f_X(x)$, then

$$C_X(t) = \int e^{itx} f_X(x)\, dx.$$

The advantage with the characteristic function is that it always exists, unlike the moment generating function, which can be infinite everywhere except $s = 0$.

Example 1: Let X be an exponential random variable with parameter μ. Find its characteristic function.

Solution: Recall that for an exponential random variable with parameter μ, $f_X(x) = \mu e^{-\mu x}$. Thus, we have

$$\begin{aligned} C_X(t) =& \int_{x=0}^{\infty} \mu e^{-\mu x} e^{itx} dx \\ =& \frac{\mu}{\mu - it}. \end{aligned}$$

We have evaluated the above integral essentially by pretending that $\mu - it$ is a real number. Although this happens to produce the correct answer in this case, the correct method of evaluating a characteristic function is by performing contour integration. Indeed, in the next example, it is not possible to obtain the correct answer by pretending that it is a real number (which is not).

Example 2: Let X be a Cauchy random variable. Find its characteristic function.

Solution: The density function for a Cauchy random variable is

$$f_X(x) = \frac{1}{\pi(1 + x^2)}.$$

Therefore,

$$\begin{aligned} C_X(t) =& \int_{x=-\infty}^{\infty} \frac{e^{itx}}{\pi(1 + x^2)} dx \\ =& e^{-|t|}. \end{aligned}$$

The above expression is not entirely trivial to obtain. Indeed, it requires considering two separate contour integrals for $t > 0$ and $t < 0$, and invoking Cauchy's residue theorem to evaluate the contour integrals. (For details, see *http://www.wpressutexas.net/forum/attachment.php?attac hmentid=408&d=1296667390.*)

However, it is also possible to obtain the characteristic function of the Cauchy random variable by invoking a Fourier transform duality trick from your undergraduate signals and systems course. (Do it!)

Recall also that the moment generating function of a Cauchy random variable does not converge anywhere except at $s = 0$. On the other hand, we find here that the characteristic function for the Cauchy random variable exists everywhere. This is essentially because the integral defining the chracteristic function converges absolutely, and hence uniformly, for all $t \in R$. Characteristic functions are thus particularly useful in handling heavy-tailed random variables, for which the corresponding moment generating functions do not exist.

Let us next discuss some properties of characteristic functions.

26.1 Properties of characteristic functions

26.1.1 Elementary properties

1) If $Y = aX + b$, $C_Y(t) = e^{ibt}C_X(at)$.

2) If X and Y are independent random variables and $Z = X + Y$, then $C_Z(t) = C_X(t)C_Y(t)$.

3) If $M_X(s) < \infty$ for $s \in [-\epsilon, \epsilon]$, then $C_X(t) = M_X(it)$ for all $t \in \mathbb{R}$.

Example 3: Let $X \sim \mathcal{N}(0, 1)$. The moment generating function is

$$M_X(s) = e^{\frac{s^2}{2}}.$$

Then, the characteristic function is

$$C_X(t) = M_X(it) = e^{\frac{-t^2}{2}}.$$

For a non-standard Gaussian, $Y \sim \mathcal{N}(\mu, \sigma^2)$, we can now invoke property 1) and conclude that $C_Y(t) = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2}\right)$.

26.1.2 Defining properties

Theorem 26.1 *A characteristic function $C_X(t)$ satisfies the following properties:*

1) $C_X(0) = 1$ and $|C_X(t)| \leq 1$, $\forall t \in \mathbb{R}$.

2) $C_X(t)$ is uniformly continuous on $\mathbb{R}$, *i.e.*, $\forall t \in \mathbb{R}$, $\exists$ a $\psi(h) \downarrow 0$ as $h \to 0$ such that

$$|C_X(t + h) - C_X(t)| \leq \psi(h).$$

3) $C_X(t)$ is a non-negative definite kernel, *i.e.*, for any n, any real $t_1, t_2, \ldots, t_n$, and any complex $z_1, z_2, \ldots, z_n$, we have

$$\sum_{j,k} z_j C_X(t_j - t_k)\overline{z_k} \geq 0.$$

Proof:

1)

$$|C_X(t)| = |\int e^{itx} d\mathbb{P}_X| \leq \int |e^{itx}| d\mathbb{P}_X = 1.$$

2)

$$\begin{aligned} |\mathbb{E}[e^{i(t+h)X}] - \mathbb{E}[e^{itX}]| &= |\mathbb{E}[e^{itX}(e^{ihX} - 1)]| \\ &\leq \mathbb{E}[|e^{ihX} - 1|]. \end{aligned}$$

Let $|e^{ihX} - 1| = y(h)$ and $\mathbb{E}[y(h)] = \psi(h)$. We now need to show that $\psi(h) \downarrow 0$ as $h \downarrow 0$. Note that $y(h) \to 0$ as $h \to 0$. Further,

$$\begin{aligned} y(h) &= |e^{ihX} - 1| \\ &= \sqrt{(\cos(hX) - 1)^2 + (\sin(hX))^2} \\ &= \sqrt{2 - 2\cos(hX)} \\ &= 2\sin\left(\frac{hX}{2}\right) \\ &\leq 2. \end{aligned}$$

Since $y(h)$ is bounded above by 2, applying DCT, we thus have $\psi(h) \to 0$ as $h \to 0$.

3)

$$\begin{aligned} \sum_{j,k} z_j C_X(t_j - t_k)\overline{z_k} &= \sum_{j,k} \int z_j e^{i(t_j - t_k)X} \overline{z_k}\, d\mathbb{P}_X \\ &= \sum_{j,k} \int z_j e^{it_j X} (\overline{z_k e^{it_k X}})\, d\mathbb{P}_X \\ &= \mathbb{E}[\sum_{j,k} z_j e^{it_j X} (\overline{z_k e^{it_k X}})] \\ &\geq \mathbb{E}[\sum_{j} |z_j e^{it_j X}|^2] \\ &\geq 0. \end{aligned}$$

■

The significance of 3) may not be apparent at a first glance. However, these three properties are considered as the defining properties of a characteristic function, because these properties are also *sufficient* for an arbitrary function to be the characteristic function of some random variable. This important result is known as Bochner's theorem, which is beyond our scope.

Theorem 26.2 ***(Bochner's theorem)*** *A function $C(\cdot)$ is a characteristic function of a random variable if and only if it satisfies the properties of theorem 26.1.*

26.2 Inversion Theorems

The following inverse theorems are presented without proof, since the proofs require some sophisticated machinery from harmonic analysis and complex variables. Essentially, they state that the CDF of a random variable can be recovered from the characteristic function.

Theorem 26.3

(i) Let X be a continuous random variable, having a probability density function $f_X(x)$ and the corresponding characteristic function be

$$C_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x) dx. \tag{26.1}$$

The probability density function, $f_X(x)$ can be obtained from the characteristic function as

$$f_X(x) = \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} e^{-itx} C_X(t) dt, \tag{26.2}$$

at every point where $f_X(x)$ is differentiable.

(ii) The sufficient (but not necessary) condition for the existence of a probability density function is that the characteristic function should be absolutely integrable, i.e.,

$$\int_{-\infty}^{\infty} |C_X(t)| dt < \infty. \tag{26.3}$$

(iii) Let $C_X(t)$ be a valid characteristic function of a random variable X with a cumulative distribution function $F_X(x)$. We define,

$$\hat{F}_X(x) = \frac{1}{2}\left(F_X(x) + \lim_{y\uparrow x} F_X(y)\right) \text{ for some } y, \tag{26.4}$$

then

$$\hat{F}_X(b) - \hat{F}_X(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-iat} - e^{-ibt}}{T} C_X(t) dt \quad \forall \ a \text{ and } b. \tag{26.5}$$

In part (iii) above, the function $\hat{F}_X(x)$ coincides with the CDF $F_X(x)$ at all points where the CDF is continuous. At points of discontinuity, it is easy to see that $\hat{F}_X(x)$ takes the value at the mid-point of the right and left limits of the CDF. Equation (26.5) says that the function $\hat{F}_X(x)$ can be recovered from the characteristic function. Finally, since the CDF is right-continuous, we can recover $F_X(x)$ from $\hat{F}_X(x)$.

26.3 Moments from the Characteristic Function

Theorem 26.4

(i) Let X be a random variable having a characteristic function $C_X(t)$. If $\frac{\mathrm{d}^k C_X(t)}{\mathrm{d}t^k}$ exists at $t = 0$, then

(a) $\mathbb{E}[|X^k|] < \infty$ *when* k *is even.*

(b) $\mathbb{E}[|X^k - 1|] < \infty$ *when* k *is odd.*

(ii) If $\mathbb{E}[|X^k|] < \infty$*, then*

$$i^k \mathbb{E}[X^k] = \left. \frac{\mathrm{d}^k C_X(t)}{\mathrm{d}t^k} \right|_{t=0}. \tag{26.6}$$

Further,

$$C_X(t) = \sum_{j=0}^{k} \frac{\mathbb{E}\left[X^j\right]}{j!} (it)^j + \mathcal{O}\left(t^k\right), \tag{26.7}$$

where the error, $\mathcal{O}\left(t^k\right)$ *means that* $\mathcal{O}\left(t^k\right) / \left(t^k\right) \to 0$ *as* $t \to 0$*.*

Note: Since $C_X(t) = \int e^{itx} d\mathbb{P}_X$ converges uniformly, we are justified in 'taking the derivative inside the integral.'

26.4 Exercise:

1. *[Papoulis]* Use characteristic function definition to find the distribution of $Y = aX^2$, if X is Gaussian with zero mean and variance σ^2.
2. *[Papoulis]* Use characteristic function definition to find the distribution of $Y = \sin(X)$, if X is uniformly distributed in $(-\pi/2, \pi/2)$.

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 27: Concentration Inequalities

Lecturer: Dr. Krishna Jagannathan *Scribe: Arjun Nadh*

A concentration inequality is a result that gives us a probability bound on certain random variables taking atypically large or atypically small values. While concentration of probability measures is a vast topic, we will only discuss some foundational concentration inequalities in this lecture.

27.1 Markov's Inequality

If X is a non-negative random variable, with $\mathbb{E}[X] < \infty$, then for any $\alpha > 0$,

$$\mathbb{P}(X > \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}.$$

Clearly, this inequality is meaningful only when $\alpha > \mathbb{E}[X]$.

Proof:

$$\begin{aligned}\mathbb{E}[X] &\overset{(a)}{=} \mathbb{E}[X\mathbb{I}_{\{X\leq\alpha\}}] + \mathbb{E}[X\mathbb{I}_{\{X>\alpha\}}],\\ &\overset{(b)}{\geq} \mathbb{E}[X\mathbb{I}_{\{X>\alpha\}}],\\ &\geq \alpha\mathbb{P}(X > \alpha).\end{aligned}$$

where (a) follows from linearity of expectations. Since X is a non-negative random variable $\mathbb{E}[X\mathbb{I}_{\{X\leq\alpha\}}] \geq 0$ and thus (b) follows. ■

Markov Inequality is probably the most fundamental concentration inequality, although it is usually quite loose. After all, the bound decays rather slowly, as $1/\alpha$. Tighter bounds can be derived under stronger assumptions on the random variable. For example, when the variance is finite, we have Chebyshev's inequality.

27.2 Chebyshev Inequality

If X is a random variable with expectation μ and variance $\sigma^2 < \infty$, then

$$\mathbb{P}(\mid X - \mu \mid > k\sigma) \leq \frac{1}{k^2}, \quad k > 0.$$

This can also be written as

$$\mathbb{P}(\mid X - \mu \mid > c) \leq \frac{\sigma^2}{c^2}, \quad c > 0.$$

Proof: The proof follows by applying Markov's inequality to the non-negative random variable $\mid X - \mu \mid^2$.

$$\begin{aligned}
\mathbb{P}(\mid X - \mu \mid^2 > (k\sigma)^2) &\leq \frac{\mathbb{E}(\mid X - \mu \mid^2)}{(k\sigma)^2}, \\
&= \frac{\sigma^2}{(k\sigma)^2}, \\
&= \frac{1}{k^2}, \\
\Rightarrow \mathbb{P}(\mid X - \mu \mid > (k\sigma)) &\leq \frac{1}{k^2}.
\end{aligned}$$

■

Note that the Chebyshev's bound decays as $1/k^2$, an improvement over the basic Markov inequality. As one might imagine, exponentially decaying bounds can be derived by invoking the Markov inequality, as long as the moment generating function exists in a neighbourhood of the origin. This result is known as the Chernoff bound, which we present briefly.

27.3 Chernoff Bound

Let $M_X(s) = \mathbb{E}[e^{sX}]$ and assume that $M_X(s) < \infty$ for $s \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$. Then

$$\mathbb{P}(X > \alpha) \leq e^{-\Lambda^*(\alpha)},$$

where $\Lambda^*(\alpha) = \sup_{s>0}(s\alpha - \log M_X(s))$.

Proof: For any $s > 0$,

$$\mathbb{P}(X > \alpha) = \mathbb{P}(e^{sX} > e^{s\alpha}).$$

By Markov's Inequality,

$$\begin{aligned}
\mathbb{P}(X > \alpha) &\leq \frac{\mathbb{E}[e^{sX}]}{e^{s\alpha}}, \\
\mathbb{P}(X > \alpha) &\leq M_X(s)e^{-s\alpha}, \quad \forall s > 0 \text{ and } s \in D_X, \qquad (27.1)
\end{aligned}$$

where $D_X = \{s \mid M_X(s) < \infty\}$.

In (27.1), note that the bound decays exponentially in α for *every* $s > 0$ belonging to D_X. The tightest such exponential bound is obtained by infimising the right hand side:

$$\begin{aligned}
\mathbb{P}(X > \alpha) &\leq \inf_{s>0} M_X(s)e^{-s\alpha}, \\
&= e^{-\sup_{s>0}(s\alpha - \log M_X(s))}.
\end{aligned}$$

Thus

$$\mathbb{P}(X > \alpha) \leq e^{-\Lambda^*(\alpha)}.$$

■

This gives us an exponentially decaying bound for the 'positive tail' $\mathbb{P}(X > \alpha)$. Similarly we can prove a Chernoff bound for the negative tail $\mathbb{P}(X < \alpha)$ by taking $s < 0$.

27.4 Exercise

1. Let X_1, X_2, ..., X_n be i.i.d. random variables with PDF f_X. Then the set of random variables X_1, X_2, ..., X_n is called a *random sample* of size n of X. The sample mean is defined as

$$\overline{X_n} = \frac{1}{n}(X_1 + X_2 + ... + X_n).$$

   Let X_1, X_2,.. X_n be a random sample of X with mean μ and variance σ^2. How many samples of X are required for the probability that the sample mean will not deviate from the true mean μ by more than $\sigma/10$ to be at least .95?

2. A biased coin, which lands heads with probability $\frac{1}{10}$ each time it is flipped, is flipped 200 times consecutively. Give an upper bound on the probability that it lands heads at least 120 times.

3. A post-office handles 10,000 letters per day with a variance of 2,000 letters. What can be said about the probability that this post office handles between 8,000 and 12,000 letters tomorrow? What about the probability that more than 15,000 letters come in?

EE5110: Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 28: Convergence of Random Variables and Related Theorems

Lecturer: Dr. Krishna Jagannathan *Scribe: Gopal, Sudharsan, Ajay, Swamy, Kolla*

An important concept in Probability Theory is that of convergence of random variables. Since the important results in Probability Theory are the limit theorems that concern themselves with the asymptotic behaviour of random processes, studying the convergence of random variables becomes necessary. We begin by recalling some definitions pertaining to convergence of a sequence of real numbers.

Definition 28.1 *Let $\{x_n, n \geq 1\}$ be a real-valued sequence, i.e., a map from $\mathbb{N}$ to $\mathbb{R}$. We say that the sequence $\{x_n\}$ converges to some $x \in \mathbb{R}$ if there exists an $n_0 \in \mathbb{N}$ such that for all $\epsilon > 0$,*

$$|x_n - x| < \epsilon, \forall\ n \geq n_0.$$

We say that the sequence $\{x_n\}$ converges to $+\infty$ if for any $M > 0$, there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $x_n > M$.
We say that the sequence $\{x_n\}$ converges to $-\infty$ if for any $M > 0$, there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $x_n < -M$.

We now define various notions of convergence for a sequence of random variables. It would be helpful to recall that random variables are after all deterministic functions satisfying the measurability property. Hence, the simplest notion of convergence of a sequence of random variables is defined in a fashion similar to that for regular functions.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of real-valued random variables defined on this probability space.

Definition 28.2 [Definition 0 (Point-wise convergence or sure convergence)]
A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to converge point-wise or surely to X if

$$X_n(\omega) \to X(\omega), \quad \forall\ \omega \in \Omega.$$

Note that for a fixed ω, $\{X_n(\omega)\}_{n\in\mathbb{N}}$ is a sequence of real numbers. Hence, the convergence for this sequence is same as the one in definition (28.1). Also, since X is the point-wise limit of random variables, it can be proved that X is a random variable, *i.e.*, it is an $\mathcal{F}$-measurable function. This notion of convergence is exactly analogous to that defined for regular functions. Since this notion is too strict for most practical purposes, and neither does it consider the measurability of the random variables nor the probability measure $\mathbb{P}(\cdot)$, we define other notions incorporating the said characteristics.

Definition 28.3 [Definition 1 (Almost sure convergence or convergence with probability 1)]
A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to converge almost surely or with probability 1 (denoted by a.s. or w.p. 1) to X if

$$\mathbb{P}\left(\{\omega | X_n(\omega) \to X(\omega)\}\right) = 1.$$

Almost sure convergence demands that the set of ω's where the random variables converge have a probability one. In other words, this definition gives the random variables "freedom" not to converge on a set of zero measure! Hence, this is a weakened notion as compared to that of sure convergence, but a more useful one.

In several situations, the notion of almost sure convergence turns out to be rather strict as well. So several other notions of convergence are defined.

Definition 28.4 [Definition 2 (convergence in probability)]
A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to converge in probability (denoted by i.p.) to X if

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| > \epsilon\right) = 0, \quad \forall\ \epsilon > 0.$$

As seen from the above definition, this notion concerns itself with the convergence of a sequence of probabilities!

At the first glance, it may seem that the notions of almost sure convergence and convergence in probability are the same. But the two definitions actually tell very different stories! For almost sure convergence, we collect all the ω's wherein the convergence happens, and demand that the measure of this set of ω's be 1. But, in the case of convergence in probability, there is no direct notion of ω since we are looking at a sequence of probabilities converging. To clarify this, we do away with the short-hand for probabilities (for the moment) and obtain the following expression for the definition of convergence in probability:

$$\lim_{n\to\infty} \mathbb{P}\left(\{\omega||X_n(\omega) - X(\omega)| > \epsilon\}\right) = 0, \quad \forall\ \epsilon > 0.$$

Since the notion of convergence of random variables is a very intricate one, it is worth spending some time pondering the same.

Definition 28.5 [Definition 3 (convergence in r^{th} mean)]
A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to converge in r^{th} mean to X if

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^r\right] = 0.$$

In particular, when $r = 2$, the convergence is a widely used one. It goes by the special name of *convergence in the mean-squared sense.*

The last notion of convergence, known as convergence in distribution, is the weakest notion of convergence. In essence, we look at the distributions (of random variables in the sequence in consideration) converging to some distribution (when the limit exists). This notion is extremely important in order to understand the Central Limit Theorem (to be studied in a later lecture).

Definition 28.6 [Definition 4 (convergence in distribution or weak convergence)]
A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ is said to converge in distribution to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad \forall\ x \in \mathbb{R} \text{ where } F_X(\cdot) \text{ is continuous.}$$

That is, the sequence of distributions must converge at all points of continuity of $F_X(\cdot)$. Unlike the previous four notions discussed above, for the case of convergence in distribution, the random variables need not be defined on a single probability space!

Before we look at an example that serves to clarify the above definitions, we summarize the notations for the above notions.

(1) *Point-wise Convergence:* $X_n \xrightarrow{\text{p.w.}} X$.

(2) *Almost sure Convergence:* $X_n \xrightarrow{\text{a.s.}} X$ or $X_n \xrightarrow{\text{w.p. } 1} X$.

(3) *Convergence in probability:* $X_n \xrightarrow{\text{i.p.}} X$.

(4) *Convergence in r^{th} mean:* $X_n \xrightarrow{r} X$. When $r = 2$, $X_n \xrightarrow{\text{m.s.}} X$.

(5) *Convergence in Distribution:* $X_n \xrightarrow{\text{D}} X$.

Example: Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \lambda)$ and a sequence of random variables $\{X_n, n \geq 1\}$ defined by

$$X_n(\omega) = \begin{cases} n, & \text{if } \omega \in \left[0, \frac{1}{n}\right], \\ 0, & \text{otherwise.} \end{cases}$$

Since the probability measure specified is the Lebesgue measure, the random variable can be re-written as

$$X_n = \begin{cases} n, & \text{with probability } \frac{1}{n}, \\ 0, & \text{with probability } 1 - \frac{1}{n}. \end{cases}$$

Clearly, when $\omega \neq 0$, $\lim_{n\to\infty} X_n(\omega) = 0$ but it diverges for $\omega = 0$. This suggests that the limiting random variable must be the constant random variable 0. Hence, except at $\omega = 0$, the sequence of random variables converges to the constant random variable 0. Therefore, this sequence does not converge surely, but converges almost surely.
For some $\epsilon > 0$, consider

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}\left(|X_n| > \epsilon\right) &= \lim_{n\to\infty} \mathbb{P}\left(X_n = n\right), \\ &= \lim_{n\to\infty} \left(\frac{1}{n}\right), \\ &= 0. \end{aligned}$$

Hence, the sequence converges in probability.
Consider the following two expressions:

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}\left[|X_n|^2\right] &= \lim_{n\to\infty} \left(n^2 \times \frac{1}{n} + 0\right), \\ &= \infty. \end{aligned}$$

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\left[|X_n|\right] &= \lim_{n\to\infty}\left(n\times\frac{1}{n}+0\right),\\ &= 1.\end{aligned}$$

Since the above limits do not equal 0, the sequence converges neither in the mean-squared sense, nor in the sense of first mean.
Considering the distribution of X_n's, it is clear (through visualization) that they converge to the following distribution:

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

Also, this happens at each $x\neq 0$ i.e. at all points of continuity of $F_X(\cdot)$. Hence, the sequence of random variables converge in distribution.

So far, we have mentioned that certain notions are weaker than certain others. Let us now formalize the relations that exist among various notions of convergence.

It is immediately clear from the definitions that point-wise convergence implies almost sure convergence. Figure (28.1) is a summary of the implications that hold for any sequence for random variables. No other implications hold in general. We prove these, in a series of theorems, as below.

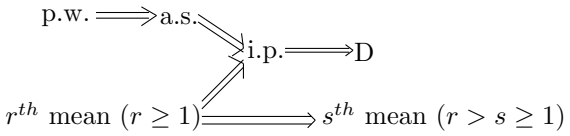

Figure 28.1: Implication Diagram

Theorem 28.7 $X_n \xrightarrow{r} X \implies X_n \xrightarrow{\text{i.p.}} X, \quad \forall\, r\geq 1.$

Proof: Consider the quantity $\lim_{n\to\infty}\mathbb{P}\left(|X_n-X|>\epsilon\right)$. Applying Markov's inequaliy, we get

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\left(|X_n-X|>\epsilon\right) &\leq \lim_{n\to\infty}\frac{\mathbb{E}\left[|X_n-X|^r\right]}{\epsilon^r},\ \forall\epsilon>0,\\ &\overset{(a)}{=} 0,\end{aligned}$$

where (a) follows since $X_n \xrightarrow{r} X$. Hence proved. ■

Theorem 28.8 $X_n \xrightarrow{\text{i.p.}} X \implies X_n \xrightarrow{\text{D}} X.$

Proof: Fix an $\epsilon>0$.

$$\begin{aligned}F_{X_n}(x) &= \mathbb{P}(X_n\leq x),\\ &= \mathbb{P}(X_n\leq x, X\leq x+\epsilon)+\mathbb{P}(X_n\leq x, X>x+\epsilon),\\ &\leq F_X(x+\epsilon)+\mathbb{P}(|X_n-X|>\epsilon).\end{aligned}$$

Similarly,

$$\begin{aligned} F_X(x-\epsilon) &= \mathbb{P}(X \le x-\epsilon), \\ &= \mathbb{P}(X \le x-\epsilon, X_n \le x) + \mathbb{P}(X \le x-\epsilon, X_n > x), \\ &\le F_{X_n}(x) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Thus,

$$F_X(x-\epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \le F_{X_n}(x) \le F_X(x+\epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

As $n \to \infty$, since $X_n \xrightarrow{\text{i.p.}} X$, $\mathbb{P}(|X_n - X| > \epsilon) \to 0$. Therefore,

$$F_X(x-\epsilon) \le \liminf_{n\to\infty} F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x) \le F_X(x+\epsilon), \ \forall \epsilon > 0.$$

If F is continuous at x, then $F_X(x-\epsilon) \uparrow F_X(x)$ and $F_X(x+\epsilon) \downarrow F_X(x)$ as $\epsilon \downarrow 0$. Hence proved. ■

Theorem 28.9 $X_n \xrightarrow{r} X \implies X_n \xrightarrow{s} X$, *if* $r > s \ge 1$.

Proof: From Lyapunov's Inequality [1, Chapter 4], we see that $(\mathbb{E}[|X_n - X|^s])^{1/s} \le (\mathbb{E}[|X_n - X|^r])^{1/r}$, $r > s \ge 1$. Hence, the result follows. ■

Theorem 28.10 $X_n \xrightarrow{\text{i.p.}} X \not\Longrightarrow X_n \xrightarrow{\text{r}} X$ *in general.*

Proof: Proof by counter-example:
Let X_n be an independent sequence of random variables defined as

$$X_n = \begin{cases} n^3, & \text{w.p. } \frac{1}{n^2}, \\ 0, & \text{w.p. } 1\text{-}\frac{1}{n^2}. \end{cases}$$

Then, $\mathbb{P}(|X_n| > \epsilon) = \frac{1}{n^2}$ for large enough n, and hence $X_n \xrightarrow{\text{i.p.}} 0$. On the other hand, $\mathbb{E}[|X_n|] = n$, which diverges to infinity as n grows unbounded. ■

Theorem 28.11 $X_n \xrightarrow{\text{D}} X \not\Longrightarrow X_n \xrightarrow{\text{i.p.}} X$ *in general.*

Proof: Proof by counter-example:
Let X be a Bernoulli random variable with parameter 0.5, and define a sequence such that X_i=X $\forall$ i. Let Y=$1 - X$. Clearly, $X_i \xrightarrow{\text{D}} Y$. But, $|X_i - Y|$=1, $\forall$ i. Hence, X_i does not converge to Y in probability. ■

Theorem 28.12 $X_n \xrightarrow{\text{i.p.}} X \not\Longrightarrow X_n \xrightarrow{\text{a.s.}} X$ *in general.*

Proof: Proof by counter-example:
Let $\{X_n\}$ be a sequence of independent random variables defined as

$$X_n = \begin{cases} 1, & \text{w.p. } \frac{1}{n}, \\ 0, & \text{w.p. } 1\text{-}\frac{1}{n}. \end{cases}$$

$\lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) = \lim_{n\to\infty} \mathbb{P}(X_n = 1) = \lim_{n\to\infty} \frac{1}{n} = 0$. So, $X_n \xrightarrow{\text{i.p.}} 0$.
Let A_n be the event that $\{X_n = 1\}$. Then, A_n's are independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. By Borel-Cantelli Lemma 2, w.p. 1 infinitely many A_n's will occur, *i.e.*, $\{X_n = 1\}$ i.o.. So, X_n does not converge to 0 almost surely. ■

Theorem 28.13 $X_n \xrightarrow{s} X \implies X_n \xrightarrow{r} X$ *if* $r > s \geq 1$ *in general.*

Proof: Proof by counter-example:
Let $\{X_n\}$ be a sequence of independent random variables defined as

$$X_n = \begin{cases} n, & \text{w.p. } \frac{1}{n^{\frac{r+s}{2}}}, \\ 0, & \text{w.p. } 1\text{-}\frac{1}{n^{\frac{r+s}{2}}}. \end{cases}$$

Hence, $\mathbb{E}[|X_n^s|] = n^{\frac{s-r}{2}} \to 0$. But, $\mathbb{E}[|X_n^r|] = n^{\frac{r-s}{2}} \to \infty$. ■

Theorem 28.14 $X_n \xrightarrow{\text{m.s.}} X \implies X_n \xrightarrow{\text{a.s.}} X$ *in general.*

Proof: Proof by counter-example:
Let $\{X_n\}$ be a sequence of independent random variables defined as

$$X_n = \begin{cases} 1, & \text{w.p. } \frac{1}{n}, \\ 0, & \text{w.p. } 1\text{-}\frac{1}{n}. \end{cases}$$

$\mathbb{E}[X_n^2]=\frac{1}{n}$. So, $X_n \xrightarrow{\text{m.s.}} 0$. As seen previously (during the proof of Theorem (28.12)), X_n does not converge to 0 almost surely. ■

Theorem 28.15 $X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\text{m.s.}} X$ *in general.*

Proof: Proof by counter-example:
Let $\{X_n\}$ be a sequence of independent of random variables defined as

$$X_n(\omega) = \begin{cases} n, & \omega \in (0, \frac{1}{n}), \\ 0, & \text{otherwise.} \end{cases}$$

We know that X_n converges to 0 almost surely. $\mathbb{E}[X_n^2]=n \longrightarrow \infty$. So, X_n does not converge to 0 in the mean-squared sense. ■

Before proving the implication $X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\text{i.p.}} X$, we derive a sufficient condition followed by a necessary and sufficient condition for almost sure convergence.

Theorem 28.16 *If* $\forall\ \epsilon > 0$, $\sum_n \mathbb{P}(A_n(\epsilon)) < \infty$, *then* $X_n \xrightarrow{\text{a.s.}} X$, *where* $A_n(\epsilon) = \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}$.

Proof: By Borel-Cantelli Lemma 1, $A_n(\epsilon)$ occurs finitely often, for any $\epsilon > 0$ w.p. 1. Let $B_m(\epsilon) = \bigcup_{n \geq m} A_n(\epsilon)$. Therefore,

$$\mathbb{P}(B_m(\epsilon)) \leq \sum_{n=m}^{\infty} \mathbb{P}(A_n(\epsilon)).$$

So, $\mathbb{P}(B_m(\epsilon)) \to 0$ as $m \to \infty$, whenever $\sum_n \mathbb{P}(A_n(\epsilon)) < \infty$. An equivalent way of proving almost sure convergence is to first consider $\lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n\geq m} \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}\right) = 0$, $\forall\ \epsilon > 0$. Hence, $\mathbb{P}(B_m(\epsilon)) \to 0$ as $m \to \infty$. This implies almost sure convergence.

■

Theorem 28.17 $X_n \xrightarrow{\text{a.s.}} X$ *iff* $\mathbb{P}(B_m(\epsilon)) \to 0$ *as* $m \to \infty$, $\forall\ \epsilon > 0$.

Proof:

Let $A(\epsilon) = \{\ \omega \in \Omega : \omega \in A_n(\epsilon)$ for infinitely many values of n $\} = \bigcap_m \bigcup_{n=m}^{\infty} A_n(\epsilon)$.
If $X_n \xrightarrow{\text{a.s.}} X$, then it is easy to see that $\mathbb{P}(A(\epsilon))=0$, $\forall\ \epsilon > 0$.
Then, $\mathbb{P}\left(\bigcap_{m=1}^{\infty} B_m(\epsilon)\right) = 0$.
Since $\{B_m(\epsilon)\}$ is a nested, decreasing sequence, it follows from the continuity of probability measures that $\lim_{m\to\infty} \mathbb{P}(B_m(\epsilon)) = 0$.

Conversely, let $C=\{\ \omega \in \Omega : X_n(\omega) \to X(\omega)$ as n $\to \infty\}$. Then,

$$\begin{aligned}
\mathbb{P}(C^c) &= \mathbb{P}\left(\bigcup_{\epsilon>0} A(\epsilon)\right) \\
&= \mathbb{P}\left(\bigcup_{m=1}^{\infty} A\left(\frac{1}{m}\right)\right) && \text{as } A(\epsilon) \subseteq A(\epsilon') \text{ if } \epsilon \geq \epsilon' \\
&\leq \sum_{m=1}^{\infty} \mathbb{P}\left(A\left(\frac{1}{m}\right)\right).
\end{aligned}$$

Also, $\mathbb{P}(A(\epsilon)) = \lim_{m\to\infty} \mathbb{P}(B_m(\epsilon))=0$. Consider

$$\begin{aligned}
\mathbb{P}\left(A\left(\frac{1}{k}\right)\right) &= \mathbb{P}\left(\bigcap_{m=1}^{\infty} B_m\left(\frac{1}{k}\right)\right) \\
&= \lim_{m\to\infty} \mathbb{P}\left(B_m\left(\frac{1}{k}\right)\right) \\
&= 0. && \forall\ k \geq 1, \text{ by assumption}
\end{aligned}$$

So, $\mathbb{P}(C^c) = 0$. Hence, $\mathbb{P}(C) = 1$. ■

Corollary 28.18 $X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\text{i.p.}} X$.

Proof: $X_n \xrightarrow{\text{a.s.}} X \implies \lim_{m\to\infty} \mathbb{P}(B_m(\epsilon))=0$.
As $A_m(\epsilon) \subseteq B_m(\epsilon)$, it is implied that $\lim_{m\to\infty} \mathbb{P}(A_m(\epsilon))=0$.
Hence, $X_n \xrightarrow{i.p.} X$. ■

Theorem 28.19 *If* $X_n \xrightarrow{i.p.} X$, *then there exists a deterministic, increasing subsequence* $n_1, n_2, n_3, \ldots$ *such that* $X_{n_i} \xrightarrow{a.s.} X$ *as* $i \longrightarrow \infty$.

Proof: The reader is referred to Theorem 13 in Chapter 7 of [1] for a proof. ■

Example: Let $\{X_n\}$ be a sequence of independent random variables defined as

$$X_n = \begin{cases} 1, & \text{w.p. } \frac{1}{n}, \\ 0, & \text{w.p. } 1\text{-}\frac{1}{n}. \end{cases}$$

It is easy to verify that, $X_n \xrightarrow{i.p.} 0$, but $X_n \overset{a.s.}{\not\to} X$. However, if we consider the subsequence $\{X_1, X_4, X_9, \dots\}$, this (sub)sequence of random variables converges almost surely to 0. This can be verified as follows.
Let $n_i = i^2$, $Y_i = X_{n_i} = X_{i^2}$.
Thus, $\mathbb{P}(Y_i = 1) = \mathbb{P}(X_{i^2} = 1) = \frac{1}{i^2}$.
$\Rightarrow \sum_{i\in\mathbb{N}} \mathbb{P}(Y_i) = \sum_{i\in\mathbb{N}} \frac{1}{i^2} < \infty$. Hence, by BCL-1, $X_i^2 \overset{a.s.}{\to} 0$.
Although this is not a *proof* for the above theorem, it serves to verify the statement via a concrete example.

Theorem 28.20 [Skorokhod's Representation Theorem]
Let $\{X_n, n \geq 1\}$ and X be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that X_n converges to X in distribution. Then, there exists a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$, and random variables $\{Y_n, n \geq 1\}$ and Y on $(\Omega', \mathcal{F}', \mathbb{P}')$ such that,

a) $\{Y_n, n \geq 1\}$ and Y have the same distributions as $\{X_n, n \geq 1\}$ and X respectively.

b) $Y_n \overset{a.s.}{\to} Y$ as $n \to \infty$.

Theorem 28.21 [Continuous Mapping Theorem]
If $X_n \overset{D}{\to} X$, and $g : \mathbb{R} \longrightarrow \mathbb{R}$ is continuous, then $g(X_n) \overset{D}{\to} g(X)$.

Proof: By Skorokhod's Representation Theorem, there exists a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$, and $\{Y_n, n \geq 1\}$, Y on $(\Omega', \mathcal{F}', \mathbb{P}')$ such that, $Y_n \overset{a.s.}{\to} Y$. Further, from continuity of g,
$\{\omega \in \Omega' \mid g(Y_n(\omega)) \to g(Y(\omega))\} \supseteq \{\omega \in \Omega' \mid Y_n(\omega) \to Y(\omega)\}$,
$\Rightarrow \mathbb{P}(\{\omega \in \Omega' \mid g(Y_n(\omega)) \to g(Y(\omega))\}) \geq \mathbb{P}(\{\omega \in \Omega' \mid Y_n(\omega) \to Y(\omega)\})$,
$\Rightarrow \mathbb{P}(\{\omega \in \Omega' \mid g(Y_n(\omega)) \to g(Y(\omega))\}) \geq 1$,
$\Rightarrow g(Y_n) \overset{a.s.}{\to} g(Y)$,
$\Rightarrow g(Y_n) \overset{D}{\to} g(Y)$.
This completes the proof since, $g(Y_n)$ has the same distribution as $g(X_n)$, and $g(Y)$ has the same distribution as $g(X)$. ■

Theorem 28.22 *$X_n \overset{D}{\to} X$ iff for every bounded continuous function $g : \mathbb{R} \longrightarrow \mathbb{R}$, we have $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$.*

Proof: Here, we present only a partial proof. For a full treatment, the reader is referred to Theorem 9 in chapter 7 of [1].
Assume $X_n \overset{D}{\to} X$. From Skorokhod's Representation Theorem, we know that there exist random variables $\{Y_n, n \geq 1\}$ and Y, such that $Y_n \overset{a.s.}{\to} Y$. From Continuous Mapping Theorem, it follows that $g(Y_n) \overset{a.s.}{\to} g(Y)$, since g is given to be continuous. Now, since g is bounded, by DCT, we have $\mathbb{E}[g(Y_n)] \to \mathbb{E}[g(Y)]$. Since, $g(Y_n)$ has the same distribution as $g(X_n)$, and $g(Y)$ has the same distribution as $g(X)$, we have $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$. ■

Theorem 28.23 *If $X_n \xrightarrow{D} X$, then $C_{X_n}(t) \longrightarrow C_X(t)$, $\forall t$.*

Proof: If $X_n \xrightarrow{D} X$, from Skorokhod's Representation Theorem, there exist random variables $\{Y_n\}$ and Y such that $Y_n \xrightarrow{a.s.} Y$.
So,

$$\cos(Y_n t) \longrightarrow \cos(Yt),\ \cos(X_n t) \longrightarrow \cos(Xt),\ \forall t.$$

As $\cos(\cdot)$ and $\sin(\cdot)$ are bounded functions,

$$\mathbb{E}[\cos(Y_n t)] + i\mathbb{E}[\sin(Y_n t)] \longrightarrow \mathbb{E}[\cos(Yt)] + i\mathbb{E}[\sin(Yt)], \ \ \forall t.$$

$$\Rightarrow C_{Y_n}(t) \longrightarrow C_Y(t), \ \ \forall t.$$

We get,

$$C_{X_n}(t) \longrightarrow C_X(t), \ \ \forall t,$$

since distributions of $\{X_n\}$ and X are same as those of $\{Y_n\}$ and Y respectively, from Skorokhod's Representation Theorem. ■

Theorem 28.24 *Let $\{X_n\}$ be a sequence of RVs with characteristic functions, $C_{X_n}(t)$ for each n, and let X be a RV with characteristic function $C_X(t)$. If $C_{X_n}(t) \longrightarrow C_X(t)$, then $X_n \xrightarrow{D} X$.*

Theorem 28.25 *Let $\{X_n\}$ be a sequence of RVs with characteristic functions $C_{X_n}(t)$ for each n, and suppose $\lim_{n\to\infty} C_{X_n}(t)$ exists $\forall t$, and is denoted by $\phi(t)$. Then, one of the following statements is true:*

(a) $\phi(\cdot)$ is discontinuous at $t = 0$, and in this case, X_n does not converge in distribution.

(b) $\phi(\cdot)$ is continuous at $t = 0$, and in this case, ϕ is a valid characteristic function of some RV X. Then $X_n \xrightarrow{D} X$.

Remark 28.26 *In order to prove that the $\phi(t)$ above is indeed a valid characteristic function, we need to verify the three defining properties of characteristic functions. However, in the light of Theorem* (28.25)*, it is sufficient to verify the continuity of $\phi(t)$ at $t = 0$. After all $\phi(t)$ is not an arbitrary function; it is the limit of the characteristic functions of X_ns, and therefore inherits some nice properties. Due to these inherited properties, it turns out it is enough to verify continuity at $t = 0$, instead of verifying all the conditions of Bochner's theorem!*

Note: Theorems (28.24) and (28.25) together are known as Continuity Theorem. For proof, refer to [1].

28.1 Exercises

1. (a) Prove that convergence in probability implies convergence in distribution, and give a counter-example to show that the converse need not hold.
   (b) Show that convergence in distribution to a constant random variable implies convergence in probability to that constant.

2. Consider the sequence of random variables with densities

$$f_{X_n}(x) = 1 \text{ - } \cos(2\pi n x), \ x \in (0, 1).$$

   Do X_n' s converge in distribution? Does the sequence of densities converge?

3. *[Grimmett]* A sequence $\{X_n, n \geq 1\}$ of random variables is said to be *completely convergent* to X if

$$\textstyle\sum_n \mathbb{P}(|X_n - X| > \epsilon) < \infty, \ \forall \epsilon > 0$$

Show that, for sequences of independent random variables, complete convergence is equivalent to almost sure convergence. Find a sequence of (dependent) random variables that converge almost surely but not completely.

4. Construct an example of a sequence of characteristic functions $\phi_n(t)$ such that the limit $\phi(t) = \lim_{n\to\infty} \phi_n(t)$ exists for all t, but $\phi(t)$ is not a valid characteristic function.

References

[1] G. G. D. Stirzaker and D. Grimmett. Probability and random processes. Oxford Science Publications, ISBN 0, 19(853665):8, 2001.

Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 29: The Laws of Large Numbers

Lecturer: Dr. Krishna Jagannathan *Scribe: Ravi Kolla, Vishakh Hegde and Arjun Bhagoji*

In this lecture, we study the laws of large numbers (LLNs), which are arguably the single most important class of theorems, which form the backbone of probability theory. In particular, the LLNs provide an intuitive interpretation for the expectation of a random variable as the 'average value' of the random variable. In the case of i.i.d. random variables that we consider in this lecture, the LLN roughly says that the sample average of a large number of i.i.d. random variables converges to the expected value. The sense of convergence in the weak law of large numbers is convergence in probability. The strong law of large numbers, as the name suggests, asserts the stronger notion of almost sure convergence.

29.1 Weak Law of Large Numbers

The earliest available proof of the weak law of large number dates to the year 1713, in the posthumously published work of Jacob Bernoulli. It asserts convergence in probability of the sample average to the expected value.

Theorem 29.1 (Weak Law of Large numbers) *Let $X_1, X_2, \dots$ be i.i.d random variables with finite mean, $\mathbb{E}[X]$. Let $S_n = \sum_{i=1}^{n} X_i$. Then,*

$$\frac{S_n}{n} \xrightarrow{i.p} \mathbb{E}[X].$$

Proof: First, we give a *partial* proof by assuming the variance of X to be finite *i.e.,* $\sigma_X^2 < \infty$. Since X_i's are i.i.d, $\mathbb{E}[S_n] = n\mathbb{E}[X]$, $Var\left(S_n\right) = nVar(X) \Rightarrow \mathbb{E}\left[\frac{S_n}{n}\right] = \mathbb{E}[X], Var\left(\frac{S_n}{n}\right) = \frac{\sigma_X^2}{n}$.

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}\left[X\right]\right| > \epsilon\right) &\leq \lim_{n\to\infty} \frac{Var\left(\frac{S_n}{n}\right)}{\epsilon^2} \quad \text{(By Chebyshev's Inequality)}, \\
&= \lim_{n\to\infty} \frac{\sigma_X^2}{n\epsilon^2}, \\
&= 0.
\end{aligned}$$

■

Next, we give a general proof using characteristic functions.

Proof: Assume that X_i (where $i = 1, 2, \dots, n, \dots$) are i.i.d random variables. The characteristic function of X_i be $C_{X_i}(t) \equiv C_X(t)$ for any $i \in \{1, 2, \dots, n\}$. Let $S_n = X_1 + X_2 + \dots X_n$ be the sum of these n i.i.d random variables. The following can be easily verified:

$$\begin{aligned}
C_{S_n} &= [C_X(t)]^n = \mathbb{E}[e^{itS_n}], \\
&= \mathbb{E}[e^{\frac{itnS_n}{n}}], \\
&= C_{\frac{S_n}{n}}(nt).
\end{aligned}$$

This implies that,

$$\begin{aligned} C_{\frac{S_n}{n}}(t) &= [C_X\left(\frac{t}{n}\right)]^n, \\ &= \left[1 + \frac{i\mathbb{E}[X]t}{n} + o\left(\frac{t}{n}\right)\right]^n. \end{aligned}$$

As $n \to \infty$, we have,

$$C_{\frac{S_n}{n}}(t) \to e^{i\mathbb{E}[X]t}, \ \forall t \in \mathbb{R}.$$

Note that, $e^{i\mathbb{E}[X]t}$ is a valid characteristic function. In fact, it is a characteristic function of a constant random variable which takes the value $\mathbb{E}[X]$. From the theorem on convergence of characteristic functions, we have

$$\frac{S_n}{n} \xrightarrow{D} \mathbb{E}[X].$$

Since $\mathbb{E}[X]$ is a constant[1], we have,

$$\frac{S_n}{n} \xrightarrow{i.p} \mathbb{E}[X].$$

■

29.2 Strong Law of Large Numbers

The Strong Law of Large Numbers (SLLN) gives us the condition when the sample average $\left(\frac{S_n}{n}\right)$ converges almost surely to the expected value.

Theorem 29.2 *If $\{X_i, i \geq 1\}$ is a sequence of i.i.d RVs with $\mathbb{E}[|X_i|] < \infty$, then $\frac{S_n}{n} \xrightarrow{a.s.} \mathbb{E}[X]$, i.e., $\mathbb{P}\left(\omega \middle| \frac{S_n(\omega)}{n} \to \mathbb{E}[X]\right) = 1$.*

Here, $S_n(\omega)$ is just $X_1(\omega) + X_2(\omega) \cdots + X_n(\omega)$. Thus, for a fixed $\omega \in \Omega$, $\left\{\frac{S_n(\omega)}{n}, n \geq 1\right\}$ is a sequence of real numbers. Then, there are the following three possibilities regarding the convergence of this sequence:

1. The sequence $\frac{S_n(\omega)}{n}$ does not converge as $n \to \infty$.
2. The sequence $\frac{S_n(\omega)}{n}$ converges to a value other than $\mathbb{E}[X]$, as $n \to \infty$.
3. The sequence $\frac{S_n(\omega)}{n}$ converges to $\mathbb{E}[X]$ as $n \to \infty$.

The SLLN asserts that the set of $\omega \in \Omega$ where the third possibility holds has a probability of 1. Also, the SLLN implies the WLLN because almost sure convergence implies convergence in probability. From Theorem 28.16, we obtain another way of stating the SLLN as given below

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m\geq n} \left\{\omega : \left|\frac{S_m(\omega)}{m} - \mathbb{E}[X]\right| > \epsilon\right\}\right) = 0, \quad \forall \epsilon > 0. \tag{29.1}$$

A general proof of the SLLN is rather long, so we will restrict ourselves to two partial proofs, each of which makes a stronger assumption than needed about the moments of the random variable X.

[1]Recall that convergence in probability is equivalent to convergence in distribution, when the limit is a constant.

29.3 Partial Proof 1 (assuming finite fourth moment)

Proof: Assume $\mathbb{E}[X_i^4] = \eta < \infty$ and without loss of generality, $\mathbb{E}[X] = 0$. The second assumption is not crucial. We want to show that $\frac{S_n}{n} \overset{a.s.}{\to} 0$.

Now,

$$\begin{aligned}
\mathbb{E}[S_n^4] &= \mathbb{E}[(X_1 + X_2 + \cdots X_n)^4], \\
&= n\eta + \binom{4}{2}\binom{n}{2}\mathbb{E}[X_1^2 X_2^2], \qquad (29.2) \\
&= n\eta + 6\binom{n}{2}\sigma^4, \\
&\leq n\eta + 3n^2\sigma^4. \qquad (29.3)
\end{aligned}$$

In (29.2), the coefficient of η is n because there are n terms of the form X_i^4. Terms of the form $X_i^3 X_j$ are not present as our assumption that $\mathbb{E}[X] = 0$ ensures that these terms go to zero. For the other surviving terms of the form $X_i^2 X_j^2$, the coefficient arises because there are $\binom{n}{2}$ ways to choose the distinct indices i and j, after which one can choose X_i from 2 out of the 4 terms being multiplied together, in which case X_j will come from the other two terms.

Now, we make use of the Markov inequality and substitute the inequality for $\mathbb{E}\left[S_n^4\right]$ from (29.3).

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{S_n}{n}\right|^4 > \epsilon\right) &\leq \frac{\mathbb{E}[S_n^4]}{n^4\epsilon}, \\
&\leq \frac{n\eta + 3n^2\sigma^4}{n^4\epsilon}, \\
&= \frac{\eta}{n^3\epsilon} + \frac{3\sigma^4}{n^2\epsilon}. \qquad (29.4)
\end{aligned}$$

Then, from (29.4),

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\left|\frac{S_n}{n}\right|^4 > \epsilon\right) \leq \sum_{n=1}^{\infty}\frac{\eta}{n^3\epsilon} + \frac{3\sigma^4}{n^2\epsilon} < \infty. \qquad (29.5)$$

Using the first Borel-Cantelli lemma, we can conclude

$$\begin{aligned}
&\left|\frac{S_n}{n}\right|^4 \xrightarrow{a.s.} 0, \\
\Rightarrow &\frac{S_n}{n} \xrightarrow{a.s.} 0.
\end{aligned}$$

29.4 Partial Proof 2 (assuming finite variance)

Assume $\sigma^2 < \infty$ and $\mathbb{E}[X] = \mu$. We begin by proving the SLLN for $X_i \geq 0$. From the partial proof of the Weak Law of Large Numbers, we have

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \leq \frac{\sigma_X^2}{n\epsilon^2}. \qquad (29.6)$$

To obtain a.s. convergence, consider a deterministic subsequence $n_i = i^2$, $i \geq 1$. Thus we get,

$$\mathbb{P}\left(\left|\frac{S_{i^2}}{i^2} - \mu\right| > \epsilon\right) \leq \frac{\sigma_X^2}{i^2\epsilon^2},$$

which implies that

$$\sum_{i=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{i^2}}{i^2} - \mu\right| > \epsilon\right) < \infty, \ \forall \epsilon > 0,$$

Using Borel-Cantelli lemma 1 we conclude that

$$\frac{S_{i^2}}{i^2} \xrightarrow{a.s.} \mu \quad \text{as } i \to \infty.$$

Let n be such that $i^2 \leq n \leq (i+1)^2$. Since $X_i \geq 0$,

$$\begin{aligned} S_{i^2} \leq S_n \leq S_{(i+1)^2}, \\ \Rightarrow \frac{S_{i^2}}{(i+1)^2} \leq \frac{S_n}{n} \leq \frac{S_{(i+1)^2}}{i^2}. \end{aligned}$$

Multiplying the expression on the left by i^2 in both the numerator and denominator, and similarly for the expression on the right, except by $(i+1)^2$, we get

$$\begin{aligned} \frac{S_{i^2}}{(i+1)^2}\frac{i^2}{i^2} \leq \frac{S_n}{n} \leq \frac{S_{(i+1)^2}}{i^2}\frac{(i+1)^2}{(i+1)^2}, \\ \frac{S_{i^2}}{i^2}\frac{i^2}{(i+1)^2} \leq \frac{S_n}{n} \leq \frac{S_{(i+1)^2}}{(i+1)^2}\frac{(i+1)^2}{i^2}, \end{aligned}$$

As $i \to \infty$, we have

$$\mu \leq \frac{S_n}{n} \leq \mu.$$

Thus, by the sandwich theorem, we get

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu.$$

To generalise to arbitrary RVs with a finite variance, we just write $X_n = X_n^+ - X_n^-$ and proceed as above since both X_n^+ and X_n^- have a finite variance and are non-negative.

■

29.5 Exercises

1. *[Gallager]* A town starts a mosquito control program and the random variable Z_n is the number of mosquitoes at the end of the n^{th} year (n = 0, 1, ...). Let X_n be the growth rate of mosquitoes in the year n i.e. $Z_n = X_n Z_{n-1}$, $n \geq 1$. Assume that $\{X_n, n \geq 1\}$ is a sequence of i.i.d. random variables with the PMF $\mathbb{P}(X = 2) = \frac{1}{2}$, $\mathbb{P}(X = \frac{1}{2}) = \frac{1}{4}$ and $\mathbb{P}(X = \frac{1}{4}) = \frac{1}{4}$. Suppose Z_0, the initial number of mosquitoes, is a known constant and assume, for simplicity and consistency, that Z_n can take non-integer values.

   (a) Find $\mathbb{E}[Z_n]$ and $\lim_{n\to\infty} \mathbb{E}[Z_n]$.

(b) Based on your answer to part (a), can you conclude whether or not the mosquito control program is successful? What would your conclusion be?

(c) Let $W_n = \log_2 X_n$. Find $\mathbb{E}[W_n]$ and $\mathbb{E}[\log_2 \frac{Z_n}{Z_0}]$.

(d) Show that there exists a constant α such that $\lim_{n\to\infty} \frac{1}{n}\log_2 \frac{Z_n}{Z_0} = \alpha$ almost surely.

(e) Show that there is a constant β such that $\lim_{n\to\infty} Z_n = \beta$ almost surely.

(f) Based on your answer to part (e), can you conclude whether or not the mosquito control program is successful? What would your conclusion be?

(g) How do you reconcile your answers to parts (b) and (f)?

2. Imagine a world in which the value of π is unknown. It is known that area of a circle is proportional to the square of the radius, but the constant of proportionality is unknown. Suppose you are given a uniform random variable generator, and you can generate as many i.i.d. samples as you need, devise a method to estimate the value of the proportionality constant without actually measuring the area/circumference of the circle.

EE5110 : Probability Foundations for Electrical Engineers **July-November 2015**

Lecture 30: The Central Limit Theorem

Lecturer: Dr. Krishna Jagannathan *Scribes: Vishakh Hegde*

30.1 Central Limit Theorem

In this section, we will state and prove the central limit theorem. Let $\{X_i\}$ be a sequence of i.i.d. random variables having a finite variance. From law of large numbers we know that for large n, the sum S_n is approximately as big as $n\mathbb{E}[X]$, i.e.,

$$\frac{S_n}{n} \xrightarrow{i.p.} \mathbb{E}[X],$$

$$\Rightarrow \quad \frac{S_n - n\mathbb{E}[X]}{n} \xrightarrow{i.p.} 0.$$

Thus whenever the variance of X_i is finite, the difference $S_n - n\mathbb{E}[X]$ grows slower as compared to n. The Central Limit Theorem (CLT) says that this difference scales as $\sqrt{n}$, and that the distribution of $\frac{S_n - n\mathbb{E}[X]}{\sqrt{n}}$ approaches a normal distribution as $n \to \infty$ *irrespective* of the distribution of X_i.

$$\frac{S_n - n\mathbb{E}[X]}{\sqrt{n}} \sim N(0, \sigma_X^2).$$

Theorem 30.1 (Central Limit Theorem) *Let $\{X_i\}$ be a sequence of i.i.d. random variables with mean $\mathbb{E}[X]$ and a non-zero variance $\sigma_X^2 < \infty$. Let $Z_n = \frac{S_n - n\mathbb{E}[X]}{\sigma_X \sqrt{n}}$. Then, we have $Z_n \xrightarrow{D} \mathcal{N}(0,1)$, i.e., $\lim_{n\to\infty} F_{Z_n}(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx$, $\forall z \in \mathbb{R}$.*

Proof: Let $Y_n = \frac{X_n - \mathbb{E}[X]}{\sigma_X}$. Let $Z_n = \frac{\sum_{i=1}^{n} Y_i}{\sqrt{n}}$. It is easy to see that Y_n has unit variance and zero mean, *i.e.,* $\mathbb{E}[Y_n] = 0$ and $\sigma_{Y_n}^2 = 1$.

$$\begin{aligned}
C_{Y_n}(t) &= 1 + it\mathbb{E}[Y_n] + \frac{i^2 t^2 \mathbb{E}[Y_n^2]}{2} + O(t^2), \\
C_{Y_n}(t) &= 1 + it(0) + \frac{i^2 t^2 (1)}{2} + o(t^2), \\
&= 1 - \frac{t^2}{2} + o(t^2), \\
C_{Z_n}(t) &= \left[C_{Y_n}\left(\frac{t}{\sqrt{n}}\right) \right]^n, \\
&= \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right) \right]^n \longrightarrow e^{\frac{-t^2}{2}} \ \forall t.
\end{aligned}$$

From the theorem on convergence of characteristic functions, Z_n converges to a standard Gaussian in distribution.

■

For example, if X_i's are discrete random variables, the CDFs will be step functions. As $n \to \infty$, these step functions will gradually converge to the error function (i.e. the steps will gradually decrease to form a continuous distribution as $n \to \infty$).

It is also important to understand what this theorem does *not* say. It is not saying that the probability density function converges to $\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$. Convergence in density function requires more stringent conditions which are stated in the Local Central Limit Theorem.

Theorem 30.2 (Local Central Limit Theorem) *Let $X_1, X_2, \ldots$ be i.i.d. random variables with zero mean and unit variance. Suppose further that their common characteristic function ϕ satisfies the following:*

$$\int_{-\infty}^{\infty} |\phi(t)|^r dt < \infty.$$

for some integer $r \geq 1$. The density function g_n of $U_n = \frac{(X_1+X_2+\ldots+X_n)}{\sqrt{n}}$ exists for $n \geq r$, and furthermore we have,

$$g_n(x) \to \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}},$$

as $n \longrightarrow \infty$, uniformly in $x \in \mathbb{R}$.

Proof: For a proof, refer to Section 5.10 in [1].

■

Let $X_1, X_2, \ldots$ be i.i.d. random variables with zero mean and unit variance. From CLT, we know that $U_n = \frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}$ is distributed as a standard Gaussian. We now look at yet another interesting result which deals with the largest value taken by U_m, $m \geq n$, for a large n.

Theorem 30.3 (The Law of the Iterated Logarithm) *Let $X_1, X_2, \ldots$ be i.i.d. random variables with zero mean and unit variance. Also, let $S_n = \sum_{i=1}^{n} X_i$ Then,*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\ log\ log\ n}} = 1\right) = 1.$$

Unlike the CLT which talks about distribution of U_n for a large, fixed n, law of iterated logarithm talks about the largest fluctuation in U_m, for $m \geq n$. In particular, it bounds the largest value taken by U_m beyond n. Formally, the subset of Ω for which this holds has a probability measure 1.

References

[1] G. G. D. Stirzaker and D. Grimmett. Probability and random processes. Oxford Science Publications, 2001.