Introduction to Probability and Linear Algebra

(CS-723)

Instructor: Saketh

Contents

Lecture 1

Abstract

This lecture defines the goals, scope, syllabus and evaluation scheme for the “Introduction to Probability and Linear Algebra” (IPL-09) course. The classical definition of probability is briefly reviewed and the need for an axiomatic approach is motivated.

1.1 Goals, Scope and Syllabus

This course introduces the student to various fundamental concepts in probability theory and linear algebra. The knowledge of such mathematical tools is essential in various fields of computer science like Machine Learning, Communication Networks, Computer Graphics and Vision etc. Though the treatment of the subject is mathematical, focus is more on the problem solving techniques rather than on the formalism. The syllabus is also tuned based on the needs of computer science:

1. Introduction to probability

   - Classical and axiomatic probability, probability spaces, conditional probability and independence

2. Random variables

   - definition, common examples, multivariate random variables, moments and moment generating functions, functions of random variables, conditional expectation

3. Sequences of random variables

   - Convergence and central limit theorem

4. Introduction to random processes
   - Markov chains and characterization
5. Topics in statistics
   - Hypothesis testing, concentration inequalities
6. Introduction to linear algebra
   - Vectors, vector spaces, bases, dimensionality and orthogonality, matrices, fundamental subspaces of matrix, rank-nullity theorem
7. Spectral decompositions
   - Eigen value and singular value decompositions, applications
8. Properties of matrices
   - Special matrices, norms and determinants

Reference text books for this course are: [3, 4, 5, 7, 8]. Also, the following video lecture series (available online) provide good insights into the subject [6, 1].

1.2 Evaluation Scheme

The grades (relative grading) will be decided based on the overall marks obtained in:

S.No.	Exam	Weightage	Date
1.	End-Semester	50%	16^{th}-29^{th} Nov'09
2.	Mid-Semester	20%	6^{th}-13^{th} Sep'09
3.	Two Quizes	10+10%	$\sim 22^{nd}$ Aug'09,$\sim 15^{th}$ Oct'09
4.	Assignments	10%	Weekly

1.3 Contact

The course page is at `http://www.cse.iitb.ac.in/saketh/teaching/cs723.html`. Office hours for the course are Wed and Fri, 3:30-5:00pm. During these hours the instructor will be available in his office (No. 306, Kanwal Rekhi Building) for clarifying specific queries that the students may have. The instructor can be contacted via phone: x7903 or email: saketh at cse also.

1.4 Introduction to Probability Theory

Probability theory is the branch of mathematics which aids in analyzing random phenomenon. Many real-world phenomena are too complicated to be studied systematically in their entirety. Examples vary from seemingly simple phenomenon like queues at ATMs/banks, vehicular traffic to more complicated ones like states of sub-atomic particles etc. A little thought must convince the reader that it is *impossible* to correctly predict the exact behaviour of such systems at the relevant time instances of interest. More importantly, in many cases this is not only a statement about the limitations of one's ability to measure particular quantities of a system but rather it is a statement about the nature of the system itself (please recollect Heisenberg's uncertainty principle from school days). In this course Probability is studied as a deductive theory which enables us to describe such phenomenon in terms of probabilities of events.

In the following the classical definition for probability is briefly reviewed (this is secondary school stuff). Later, the motivation for an axiomatic approach to probability is presented.

A *random experiment* is a *repeatable* experiment in which the *outcome* of the experiment is not known (can't be determined); however the set of all possible outcomes, called the *sample space* is known. Let the sample space be denoted by Ω. An *event* is a subset of the sample space. The set of all events is nothing but the power set of sample space i.e., 2^{Ω}. The classical definition of probability, which was used for several centuries, is (here E is an event):

Definition 1.4.1.

$$P(E) \equiv \frac{|E|}{|\Omega|} \tag{1.1}$$

where $|E|$ is the size (cardinality) of the set E i.e. number of outcomes in *favour* of event E and $|\Omega|$ is the total number of outcomes possible. Of course, the inherent assumption is that all the outcomes are *equally likely*[1]. This classical definition has multiple flaws: a) the definition itself uses the notion of probability (via the equally likely assumption) b) application is limited to cases where the "equal likely" assumption holds e.g. not suitable for a biased coin or loaded die etc.

Also, it can be shown that probability actually depends on the choice of the set of possible outcomes and events: consider the problem described in the *Bertrand's paradox*: Given a circle of radius R, determine the probability that

[1] *Principle of insufficient reason* or *Principle of maximum entropy*

the length of a randomly selected chord AB is greater than $\sqrt{3}R$. There are multiple ways of solving this problem; a couple of them are discussed below:

1. Assume the random chord AB is perpendicular to a specific diametrical chord of the circle, say FG. Note that though this reduces the number of possibilities, the number favourable cases also are correspondingly reduced. Hence this restriction has no effect on the probability which is to be calculated. It is easy to see that if the mid-point (M) of the chord AB is at a distance less than $\frac{R}{2}$ from the center of the circle, then the length of the chord is indeed greater than $\sqrt{3}R$. Now, consider the set of outcomes as the various positions of M on the line FG i.e., $\Omega^{(1)} = [-R, R]$. Above observation shows that the event of interest is $E^{(1)} = [-\frac{R}{2}, \frac{R}{2}]$. Hence the probability of the event $E^{(1)}$ is $\frac{|E^{(1)}|}{|\Omega^{(1)}|} = \frac{1}{2}$. Here length is taken as the "measure" for size of sets, which are intervals in this case.

2. Again consider the random chord AB. It is easy to see that if the mid-point M of the chord AB lies inside a concentric circle of radius $\frac{R}{2}$, then its length is indeed greater than $\sqrt{3}R$. Now, consider the set of outcomes as the various positions of M i.e., points in the original circle. With this, $\Omega^{(2)} = \{(r, \theta) | r \in [0, R],\ \theta \in [0, 2\pi]\}$ and $E^{(2)} = \{(r, \theta) | r \in [0, \frac{R}{2}],\ \theta \in [0, 2\pi]\}$. Considering area as the "measure" for the size of the sets (which are 2-d intervals in this case), the required probability is $\frac{|E^{(2)}|}{|\Omega^{(2)}|} = \frac{1}{4}$.

The above discussion shows that probability actually depends on the choice of the outcomes. In other words, the set of possible outcomes Ω needs to be defined before venturing into the calculation of probabilities. Also, note that (unknowingly) while calculating probabilities using area and length as the "measures" for the sizes of the events, we have assumed that events always take the form of (n-dimensional) intervals. This is because the concept of length and area are usually associated with intervals rather than for arbitrary sets. So do we need to assume that the set of possible events must only be intervals (and not the power set of Ω) ? or do we need to generalize the notions of length and area ??. (Interested students please look into the wikipedia article on this problem at `http://en.wikipedia.org/wiki/Bertrand's_paradox_(probability)`).

An axiomatic definition of probability was proposed by A. N. Kolmogorov in the early 1930s which addresses the short comings of the classical definition. The definition assumes that the sample space (Ω) and set of events $(\mathcal{F})$ are given and then defines probability as a set function satisfying few axioms. This will be the topic of discussion for the next lecture.

Lecture 2

Abstract

The objective of this lecture is to introduce the axiomatic definition of probability (Kolmogorov's approach). Given the set of outcomes and events, probability is defined as a real-valued set function satisfying three fundamental axioms. Examples of probability functions which satisfy the three axioms for the cases with countable sample spaces are then discussed. The case of uncountable sample spaces motivates the need for "shrinking the size" of the set of events using the notion of a σ-algebra. The lecture ends discussing examples of σ-algebras and the refined axiomatic definition of probability using the notion of σ-algebra.

The discussion in the previous lecture, and in particular the problem described in *Bertrand's paradox*, clearly motivates the need for an improved definition of probability. As noted in the last class, it was A. N Kolmogorov in his pioneering work in the 1930s [2] gave such a definition of probability which is in wide-acceptance today. In this lecture, we study the Kolmogorov's axiomatic definition of probability.

An important learning from the problem in *Bertrand's paradox* was that the choice of the set of outcomes Ω is crucial in determining probabilities of events. Hence we define probability assuming the set of outcomes is fixed a-priori. Also, let us assume that the set of events, $\mathcal{F}$, is the power set 2^{Ω} itself. Now, we are in search of a rigorous definition of probability which generalizes the classical notion of probability (1.1). Kolmogorov identified three important properties that the classical probability satisfied:

Non-negativity If $E \in \mathcal{F}$, then $P(E) \geq 0$.

Unit measure $P(\Omega) = 1$.

σ-additivity If $E_i \in \mathcal{F}, i = 1, 2, \ldots$ and $E_i \cap E_j = \phi, i \neq j,\ i, j = 1, 2, \ldots$, then $P(\cup_{i=1}^{\infty} E_i) = \sum_i P(E_i)$. In other words, probability of countable union of disjoint (*mutually exclusive*) events is the sum of the probabilities of the individual events.

The properties non-negativity and unit-measure are merely concerned with the boundary (extremal) values of probability whereas σ-additivity is the vital property which actually captures the classical (intuitive) notion of probability that: "bigger" events have higher probabilities and the "smaller" ones have lower probabilities.

Now let us define probability as the real-valued set function $P : \mathcal{F} \mapsto \mathbb{R}$ which satisfies the above mentioned (three) principles of non-negativity, unit-measure and σ-additivity. Note that since $\mathcal{F} = 2^{\Omega}$, we are assured that $\Omega \in \mathcal{F}$ and $\cup_{i=1}^{\infty} E_i \in \mathcal{F}$ whenever $E_i \in \mathcal{F}, i = 1, 2, \ldots$. Hence the definition is indeed a valid one. Now let us verify if there is atleast one such function which satisfies these three principles (axioms).

Let us start with the case of finite sample space e.g., $\Omega = \{1, 2, \ldots, n\}, n \in \mathbb{N}$ (no. sixers in a match by Yuvi :). Suppose we denote $p_i \equiv P(i), i = 1, \ldots, n$ and define probability of any event as the sum of the probabilities of outcomes favourable to that event, i.e, $P(E) = \sum_{i: i \in E, i \in \Omega} p_i$. It is easy to see that if p_i are chosen such that $p_i \geq 0, \sum_{i=1}^{n} p_i = 1$, then all the three axioms are satisfied. This shows that there are infinitely many choices for the probability function (the classical probability just picks one such choice with $p_i = \frac{1}{n}, i = 1, \ldots, n$). The situation does not change much if the sample space is countably infinite e.g. $\Omega = \{1, 2, \ldots, n, n+1, \ldots\}$: now we just need to pick p_i such that $p_i \geq 0, \sum_{i=1}^{\infty} p_i = 1$. For e.g. choose $p_i = (1-q)q^{i-1}, q \in [0, 1]$ (geometric series family). Ofcourse there are other families of functions too. This discussion clearly shows that the new definition of probability is well-defined and indeed generalizes the classical notion of probability.

Before extending this strategy of assigning probabilities to each element of the sample space to the case of uncountable sample spaces, let us ask the following question: what is the maximum number of mutually exclusive events (m.e.e) that $\mathcal{F}$ can have each with probability atleast $\frac{1}{k}$ (here, $0 < k \leq 1$)? The answer to this question is: there can be atmost k m.e.e in $\mathcal{F}$ (why?). In other words, there can be atmost countable number of m.e.e in $\mathcal{F}$. Hence assigning probabilities to each element of an uncountable sample space is not feasible. In fact, Guisseppe Vitali (in around 1900s) showed[1] examples of "large" events to which probability assignment cannot be done i.e., the set of events $\mathcal{F}$ cannot be taken to be the

[1] Students looking for insights into a formal proof can look at `www.math.unl.edu/~gmeisters1/papers/Measure/measure.pdf`

whole of 2^Ω! (Again, recall the question posed towards the end of Lecture 1: while applying notion of length/area probability are we inherently assuming $\mathcal{F} < 2^\Omega$?)

The above discussion clearly shows that one must consider "smaller" sets of events than 2^Ω and which are compatible with the axiomatic definition of probability. A σ-algebra is a useful algebraic structure on the set of subsets of a set which facilitates this "shrinkage":

Definition 2.0.2. *σ-algebra over a set Ω is a non-empty collection $\mathcal{F}$ of subsets of Ω such that it is closed under:*

- *complementation i.e., $E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}$*
- *countable union i.e., $E_i \in \mathcal{F}, i = 1, 2, \ldots \Rightarrow \cup_{i=1}^{\infty} E_i \in \mathcal{F}$*

From this definition, it is easy to show that any σ-algebra atleast contains two events: $\{\phi\}$ (impossible event) and $\{\Omega\}$ (certain event). Also, one can show that a σ-algebra is closed under countable intersection (use De Morgan's Laws). E.g. of σ-algebras are $\{\{\phi\}, \{\Omega\}\}$, $\{\{\phi\}, E, E^c, \{\Omega\}\}$, 2^Ω and so on. As we discussed earlier, if the σ-algebra is taken to be 2^Ω, then probability assignment cannot be done. On the other extreme, the case of $\mathcal{F} = \{\{\phi\}, \{\Omega\}\}$ is trivial. What one needs in practice is a σ-algebra which includes "most" of the events of interest. For e.g. if $\Omega = \mathbb{R}$ we may want the σ-algebra to atleast include all kinds of intervals (so that length can then be employed as the measure for size of events). Usually, the smallest σ-algebra containing the "interesting" events (say intervals) is taken as $\mathcal{F}$. Note that it is easy to construct such a σ-algebra: starting with the given events, we just need to supply all the events which make it closed under complementation and countable unions (we will see an example of such a σ-algebra over $\Omega = \mathbb{R}$ in a later lecture). We call such a "smallest" σ-algebra as the one *generated* by the events under consideration. However the question of which is the "largest" σ-algebra for which probability assignment can be done is an important one (and beyond the scope of this course). Now we are in a position to formally state the axiomatic definition of probability:

Definition 2.0.3. *Given a sample space Ω and a σ-algebra $\mathcal{F}$ over Ω, probability is a real-valued set function $P : \mathcal{F} \mapsto \mathbb{R}$ satisfying the following three axioms:*

Non-negativity $E \in \mathcal{F} \Rightarrow P(E) \geq 0$.

Unit measure $P(\Omega) = 1$.

σ-additivity $E_i \in \mathcal{F}, i = 1, 2, \ldots, \; E_i \cap E_j = \phi, i \neq j, \; i, j = 1, 2, \ldots \Rightarrow P(\cup_{i=1}^{\infty} E_i) = \sum_i P(E_i)$

The triplet $(\Omega, \mathcal{F}, P)$ *is known as the probability space.*

In the subsequent lecture we will explore some properties of the defined probability function.

Lecture 3

Abstract

We begin by proving some interesting properties of the probability function defined in last lecture. The issue of continuity of the probability function is dealt with to some extent. The lecture ends with some discussion on the notion of conditional probability and independence of events.

3.1 Consequences of the Axioms

Let us look at some of the consequences of the three axioms of probability:

1. $E \in \mathcal{F} \Rightarrow P(E^c) = 1 - P(E)$

   $$\because \underbrace{E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}}_{\mathcal{F} \text{ is } \sigma-\text{algebra}}, \underbrace{1 = P(\Omega)}_{\text{Unit Measure}} = \underbrace{P(E \cup E^c) = P(E) + P(E^c)}_{\text{Additivity}}$$

   In particular, $P(\{\phi\}) = 0$ (substitute Ω for E).

2. $E_1 \subseteq E_2 \in \mathcal{F} \Rightarrow P(E_1) \leq P(E_2)$

   $\because E_1, E_2 \in \mathcal{F} \Rightarrow E_2 - E_1 \equiv \underbrace{E_2 \cap E_1^c \in \mathcal{F}}_{\mathcal{F} \text{ is } \sigma-\text{algebra}}, \underbrace{P(E_2) = P(E_1) + P(E_2 - E_1)}_{\textit{Additivity}}, \underbrace{P(E_2 - E_1) \geq 0}_{\text{Non-negativity}}$

   In particular, $E \in \mathcal{F} \Rightarrow P(E) \leq 1$.

3. $E_1, E_2 \in \mathcal{F} \Rightarrow P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$. This is because the following three identities are true: $P(E_1 \cup E_2) = P(E_1 - E_2) + P(E_2 - E_1) +$

$P(E_1 \cap E_2), P(E_1) = P(E_1 - E_2) + P(E_1 \cap E_2), P(E_2) = P(E_2 - E_1) + P(E_1 \cap E_2)$. Note that each of these in turn follow from the additivity property. Also, refer the assignment for more generalizations and bounds derivable from this.

3.1.1 Property of Sequential Continuity

We wish to know whether the probability function defined in defn. 2.0.3 is "continuous". To do this, let us first discuss the notion of convergence of sequence of sets. In this lecture we will be concerned only with special sequences known as the *monotonic sequences* and discuss general convergence issues at a later stage.

A sequence of events $E_i \in \mathcal{F}, i = 1, 2, \ldots$, is said to be *monotonically non-decreasing* iff $E_1 \subseteq E_2 \subseteq \ldots E_n \subseteq E_{n+1} \subseteq \ldots$. Similarly, we define a *monotonically non-increasing* sequence as the one satisfying: $E_1 \supseteq E_2 \supseteq \ldots E_n \supseteq E_{n+1} \supseteq \ldots$. A sequence is called *monotonic* is it is either monotonically non-decreasing or non-increasing. Now, we define the limits of monotonic sequences as follows:

Definition 3.1.1. *If $\{E_n\}$ is a non-decreasing sequence, then $\lim_{n\to\infty} E_n \equiv \cup_{i=1}^{\infty} E_i$ and if $\{E_n\}$ is non-increasing, then $\lim_{n\to\infty} E_n \equiv \cap_{i=1}^{\infty} E_i$.*

Note that for any monotonic sequence in $\mathcal{F}$, the limiting event, $E \equiv \lim_{n\to\infty} E_n$, indeed belongs to $\mathcal{F}$ because $\mathcal{F}$ is a σ-algebra (and hence closed under countable unions and intersections). Infact, one could alternatively define σ-algebra as that collection of subsets of Ω, in which all monotonic sequences converge. Now, we can ask the following questionn for monotonic sequences: is $P(\lim_{n\to\infty} E_n) = \lim_{n\to\infty} P(E_n)$? The answer turns out to be yes and this property is known as the property of *sequential continuity* and is proved below considering the case of monotonically non-decreasing sequences (proof in the other case is similar):

Proof. Let us define $E \equiv \lim_{n\to\infty} E_n \equiv \cup_{i=1}^{\infty} E_i$. Then, for any $n \in \mathbb{N}$, we have $E = E_n \cup_{i=n}^{\infty} (E_{i+1} - E_i)$. By the σ-additivity axiom, $1 \geq P(E) = P(E_n) + \sum_{i=n}^{\infty} P(E_{i+1} - E_i)$. In particular, if $n = 1$, then we obtain: $\sum_{i=1}^{\infty} P(E_{i+1} - E_i) \leq 1$. This says that the (infinite) series sum $\sum_{i=1}^{\infty} P(E_{i+1} - E_i)$ is bounded above and thus the "tail sum" must go to zero. In other words, $\lim_{n\to\infty} \sum_{i=n}^{\infty} P(E_{i+1} - E_i) = 0$ and hence $P(E) = \lim_{n\to\infty} P(E_n)$. □

In the lecture, we discussed a simple application of the property of sequential continuity: "the probability of never seeing a head in a series of coin tosses is zero".

3.2 Conditional Probability

Many times, probabilities of events change when it is known that a certain other event has happened. A trivial example is: probability of seeing a head in coin toss experiment is 1 if somebody already told you that the event H occured. The reason why the probability value changed is because the set of possible outcomes effectively got changed. Such situations are modeled using conditional probability, which is defined as follows:

Definition 3.2.1. *Given a probability space* $(\Omega, \mathcal{F}, P)$*, and an event* B *with non-zero probability of occurance i.e.,* $P(B) > 0$*, we define a new probability function known as the conditional probability given* B *as follows:*

$$P_B(A) \equiv P(A/B) \equiv \frac{P(A \cap B)}{P(B)} \ \forall \ A \in \mathcal{F}$$

It is a trivial excercise to verify that the conditional probability function defined above, satisfies non-negativity, unit-measure and σ-additivity: let $E_i, i = 1, 2, \ldots \in \mathcal{F}$ be m.e.e. (mutually exclusive events), then $P_B(\cup_{i=1}^{\infty} E_i) = \frac{P\left(\{\cup_{i=1}^{\infty} E_i\} \cap B\right)}{P(B)} = \frac{P\left(\cup_{i=1}^{\infty} (E_i \cap B)\right)}{P(B)}$ (why?). Now each of $E_i \cup B$ is indeed m.e.e. because E_i themselves are. Hence, $P_B(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P_B(E_i)$. Note that, $P_B(B) = 1$ and for any $C \cap B = \phi$, $P_B(C) = 0$. In other words, it is equivalent to shrinking the set of possible outcomes from Ω to B. Also, it is easy to verify that $P(A \cap B) = P(A/B)P(B)$ and $P(\cap_{i=1}^{n} E_i) = P(E_1)P(E_2/E_1)P(E_3/E_1 \cap E_2) \ldots P(E_n/E_1 \cap \ldots E_{n-1})$. These formulae are useful whenever conditional probabilities are easier to calculate than probabilities of intersection of events.

Now, let $E_i, i = 1, \ldots, n \in \mathcal{F}$ such that they are m.e.e. and $\cup_{i=1}^{n} E_i = \Omega$. Such a collection of events is known as a *partition* of the sample space Ω. Let A be another event in $\mathcal{F}$ and hence $A = \cup_{i=1}^{n} (A \cap E_i)$. Since each of $A \cap E_i$ are again m.e.e., we have that $P(A) = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i}^{n} P(A/E_i)P(E_i)$. This is known as the total probability rule. Now, further, $P(E_i/A) = \frac{P(E_i \cap A)}{P(A)} = \frac{P(A/E_i)P(E_i)}{P(A)} = \frac{P(A/E_i)P(E_i)}{\sum_{j}^{n} P(A/E_j)P(E_j)}$. This is known as the Baye's rule. The probabilities $P(E_j)$ are known as the *prior probabilities*, $P(A/E_j)$ are known as class-conditional probabilities or aposterior probabilities and $P(E_j/A)$ are known as posterior probabilities. Usually, the posterior probabilities are of interest and are difficult to estimate whereas the prior and class-conditional probabilities are easy to estimate (In later lectures concrete examples will be given).

Example: In a certain population, the probability of a person having disease is p. A new diagnostic test was deviced which has a rate of success q i.e., with

probability q it identifies a correctly the state of a person (disease or normal). In order to deploy the diagnostic, it is required that the ratio of probability of person having disease and probability of being normal given that the diagnostics report presence of disease, is high. Compute this ratio.
Solution: Let D be the event a person has desease and T be the event that the diagnostic test reports presence of disease. $P(D/T) = \frac{P(T/D)P(D)}{P(T)}$ and $P(D^c/T) = \frac{P(T/D^c)P(D^c)}{P(T)}$. Hence the required fraction is $\frac{P(D/T)}{P(D^c/T)} = \frac{P(T/D)P(D)}{P(T/D^c)P(D^c)} = \frac{qp}{(1-q)(1-p)}$. Note that the decision making criteria does not involve $P(T)$.

Two events are said to be independent if $P(E_1 \cap E_2) = P(E_1)P(E_2)$. In case $P(E_1), P(E_2) \neq 0$, this implies that $P(E_1/E_2) = P(E_1), P(E_2/E_1) = P(E_2)$. In other words occurance of one event does not change anything to affect the probability of the other. This notion can be extended to a set of n events $E_i, i = 1, \ldots, n$. However one needs to ensure possible pairs, triples, etc. of these events are independent. This amounts to $2^n - n - 1$ conditions! Pair-wise independence may not imply higher order independences and vice-versa.

Lecture 4

Abstract

In this lecture, we construct a σ-algebra over $\mathbb{R}$ known as the Borel σ-algebra, which we always take as the set of the events while working with $\Omega = \mathbb{R}$. We introduce the concept of a random variable and then formally define it. Distribution function of a random variable is defined and its properties are studied.

Many times we may want to quantify the outcomes of random expts. in terms of (real) numbers. The reason may be that such a quantification is very natural to the problem at hand or it simply may be to aid study of probability functions defined over standardized (numeric) sample spaces. For e.g. consider the expt. of two coin tosses where the sample space is $\Omega = \{HH, HT, TH, TT\}$. One way to quantify this sample space is to consider the function X defined as "the number of heads". It is easy to see that $X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0$. Such a function which quantifies the sample space is known as a *random variable*. Now given the original probability space $(\Omega, \mathcal{F}, P)$, one can calculate the probability P_X of an event B in the new sample space (i.e. set of real numbers $\mathbb{R}$) using this simple notion: $P_X(B) = P(X^{-1}(B)) = P(\{\omega \in \Omega : X(\omega) \in B\})$. The idea is that probability of events in the new sample space are computed as the probability of their pre-images in the original sample space. For e.g. in the above example, $P_X([-3, 1]) = P(X^{-1}([-3, 1])) = P(\{\omega \in \Omega : X(\omega) \in [-3, 1]\}) = P(\{TT, HT, TH\})$. Note that this probability is known only if $\{TT, HT, TH\} \in \mathcal{F}$. Pathological examples of $\mathcal{F}$, which are σ-algebras, can be constructed where $\{TT, HT, TH\} \notin \mathcal{F}$ e.g., $\mathcal{F} = \{\{\phi\}, \Omega\}$ or $\mathcal{F} = \{\{HH, TT\}, \{HT, TH\}, \Omega, \{\phi\}\}$ etc. Hence while defining random variable X we must ensure that such pathological $\mathcal{F}$ would not effect the probability calculations. Also, we need to fix an appropriate σ-algebra for $\mathbb{R}$ (we know that taking the set of events as $2^{\mathbb{R}}$ wont work!). The σ-algebra we choose is called the Borel σ-algebra and is described in the following section.

4.1 Borel σ-algebra

Suppose we consider the following (basic/elementary) events in $\mathbb{R}$: $\mathcal{A} = \{(-\infty, x] : x \in \mathbb{R}\}$. We call the σ-algebra generated by $\mathcal{A}$ as the Borel σ-algebra and always denote it by $\mathcal{B}$. In other words, $\mathcal{B}$ is the "smallest" σ-algebra which contains $\mathcal{A}$. It is easy to see that $\mathcal{B}$ contains the following kinds of events (this is no where near the exhaustive list):

1. $(-\infty, x]$ ($\because$ they are in $\mathcal{A}$ itself)

2. (x, ∞) ($\because$ complements of events in $\mathcal{A}$)

3. $\mathbb{R}$ and $\{\phi\}$ ($\because$ union and intersection of complementary intervals of kind 1,2)

4. $(x_1, x_2]$ ($\because$ intersection of intervals of kind 1,2)

5. (x_1, x_2) ($\because (x_1, x_2) = \cup_{n=1}^{\infty}(x_1, x_2 - \frac{1}{n}]$)

6. $[x_1, x_2)$ ($\because [x_1, x_2) = \cap_{n=1}^{\infty}(x_1 - \frac{1}{n}, x_2)$)

7. $[x_1, x_2]$ ($\because$ intersection of intervals of kind 4,6)

8. $\{x\}$ (solve the assignment problem)

9. All countable sets of reals ($\because$ countable union of singletons)

It can be shown that $\mathcal{B} \neq 2^{\mathbb{R}}$ and infact, there exist many probability functions that can be defined on $(\mathbb{R}, \mathcal{B})$ (proof is beyond our scope). Hence $\mathcal{B}$ is a σ-algebra which is rich enough to model various events on $\mathbb{R}$ and is also of manageable size in the sense that probability functions can be defined. Therefore we always take it as the default set of events for $\Omega = \mathbb{R}$.

4.2 Random Variables

Now we are in position to define a random variable:

Definition 4.2.1. *Given a probability space $(\Omega, \mathcal{F}, P)$, we define random variable as a function $X : \Omega \mapsto \mathbb{R}$, where $X^{-1}(B) \in \mathcal{F} \ \forall \ B \in \mathcal{B}$. The induced probability space of the random variable is $(\mathbb{R}, \mathcal{B}, P_X)$, where $P_X(B) \equiv P(X^{-1}(B)) \ \forall \ B \in \mathcal{B}$.*

One can now easily verify that the probability function P_X indeed satisfies the non-negativity as well as the unit-measure axioms of probability. The σ-additivity of P_X follows from that of P and the fact that the pre-image, X^{-1}, preserves the set operations:

Proof. We want to show that $P_X(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} P_X(B_i) \ \forall \ B_i \in \mathcal{B} \ \ni \ B_i \cap B_j = \phi \ (i \neq j)$. We have,

$$\begin{aligned} P_X(\cup_{i=1}^{\infty} B_i) &= P(X^{-1}(\cup_{i=1}^{\infty} B_i)) && (\because \text{ defn. of } P_X) \\ &= P(\cup_{i=1}^{\infty} X^{-1}(B_i)) && (\because X^{-1} \text{ preserves set operations}) \\ &= \sum_{i=1}^{\infty} P(X^{-1}(B_i)) && (\because B_i \cap B_j = \phi \Rightarrow X^{-1}(B_i) \cap X^{-1}(B_j), \ \forall \ i \neq j) \\ &= \sum_{i=1}^{\infty} P_X(B_i) && (\because \text{ defn. of } P_X) \end{aligned}$$

□

This discussion shows that the induced probability function is indeed a valid probability function according to the axiomatic definition. Now, the condition $X^{-1}(B) \in \mathcal{F} \ \forall \ B \in \mathcal{B}$ is a very mild one (i.e., it is an issue in very rare/pathalogical cases). Infact, one need not verify the condition for forall $B \in \mathcal{B}$; it is enough to verify $X^{-1}(A) \in \mathcal{F} \ \forall \ A \in \mathcal{A}$. In other words, it is enough to check the condition for the basic events which generated the Borel σ-algebra. This is because for any function, the pre-image preserves set operations and both $\mathcal{F}, \mathcal{B}$ are indeed σ-algebras. Also (again without formal proof) one can show that probabilities of any event in $\mathcal{B}$ can be computed from the probabilities of the basic events: $P_X((-\infty, x])$. Here's an example: $P((x_1, x_2]) = P((-\infty, x_2]) - P((-\infty, x_1])$. Since these probabilities are of such an importance, we give a name to these as a function of x: the *Distribution Function*.

4.3 Distribution Function of Random Variable

The distribution function $F_X(x)$ of a random variable X is a real valued function on reals defined as $F_X(x) = P_X((-\infty, x])$. Recall that $P_X((-\infty, x]) = P(\{\omega \in$

$\Omega : X(\omega) \leq x\}$). The short hand notation for the last probability term is usually: $P[X \leq x]$ (abuse of notation!). In other words, $F_X(x) \equiv P[X \leq x]$. It is easy to see that the distribution function for the "number of heads" random variable is:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{4} & 0 \leq x < 1 \\ \frac{3}{4} & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

Also, the distribution function satisfies these properties:

1. $0 \leq F_X(x) \leq 1 \; \forall x$ ($\because$ distribution function at each x is after all a probability).

2. $F_X(-\infty) = P(\phi) = 0$ and $F_X(\infty) = P(\Omega) = 1$.

3. $x_1 \leq x_2 \Rightarrow (-\infty, x_1) \subseteq (\infty, x_2) \Rightarrow F_X(x_1) = P((-\infty, x_1)) \leq P((-\infty, x_2)) = F_X(x_2)$. In other words, F_X is monotonically non-decreasing function.

4. $F_X(x)$ is right continuous and has left limit.

The property that F_X is right continuous etc. can be proved using the results on continuity of P_X. However since we have studied continuity issues of probability functions (like P_X) only for the case of monotonic sequences, we will not be able to provide a formal proof of this at this stage. What we provide below is a justification considering monotonic sequences alone:

Consider a sequence $\{x_n = x + a_n\}$ where $\{a_n\}$ is any non-negative sequence that monotonically decreases to zero (for e.g., $\{a_n\} = \frac{1}{n}$). So, $x_n \rightarrow x$ from the right monotonically i.e., $x_n \downarrow x$. Now consider the sequence of intervals $\{I_n =$ (-$\infty, x_n)\}$. Since this is a monotonically non-increasing sequence of events we have $\lim_{n\rightarrow\infty} I_n = \cap_{n=1}^{\infty} I_n = (-\infty, x]$ and hence by sequential continuity property we have $F_X(x) = P((-\infty, x]) = P(\lim_{n\rightarrow\infty} I_n) = \lim_{n\rightarrow\infty} P(I_n) = \lim_{x_n \downarrow x} F_X(x_n)$. Similarly, one can prove that $\lim_{x_n \uparrow x} F_X(x) = P[X < x]$. However since $P[X \leq x] = P[X < x] + P[X = x]$, unless $P[X = x] = 0$ it will not happen that $P[X < x] = P[X \leq x]$. In case $P[X = x] = 0$, then F_X is continuous at x (both from right and left).

Infact, any function which satisfies these four properties is called a distribution function. One can also show (again not in this course) that for every distribution function there exists a random variable. However there might be multiple random variables with the same distribution function (an example is in the assignments). In the next lecture we will study distribution functions which are not left continuous i.e. $P[X = x] \neq 0$ and arise in the case of special kind of random variables known as *Discrete random variables.*

Bibliography

[1] Mrityunjoy Chakraborty. Probability and Random Processes Lecture Videos. Available at http://nptel.iitm.ac.in/video.php?courseId=1056&p=1.

[2] A. N. Kolmogorov.

[3] A. Papoulis and S. U. Pillai. *Probability, Random Variables, and Stochastic Processes*. Tata Mc-Graw Hill, 4 edition, 2002.

[4] S. M. Ross. *Introduction to Probability and Statistics for Engg. and Scientists*. Academic Press, 3 edition, 2004.

[5] S. M. Ross. *Introduction to Probability Models*. Academic Press, 9 edition, 2006.

[6] Gilbert Strang. Linear Algebra Lecture Videos. Available at http://web.mit.edu/18.06/www/Video/video-fall-99.html, 2000.

[7] Gilbert Strang. *Linear Algebra and its Applications*. Cengage Learning, 4 edition, 2006.

[8] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley Cambridge Press, 4 edition, 2009.