Lecture 1: Introduction to Probability

Rajat Mittal

IIT Kanpur

Please look at the course policies mentioned in the course homepage. Most importantly, any immoral behavior like cheating and fraud will be punished with extreme measures and without any exception. This note is an introduction and written to give a feeling about what will be covered in the course (hence the terms are loosely defined). This introduction will make more and more sense as we progress through the course.

You have already taken CS201 and studied many parts of discrete mathematics. To remind you, the branch of mathematics which deals with "discrete" objects and structures is called *discrete mathematics*. Here, by discrete set, we mean that the elements are distinct and not connected. So we can say that the set has finite or countably infinite number of elements (the elements can be counted). To get the intuition, the set of natural numbers is a discrete set. On the other hand, the set of real numbers are continuous.

Discrete mathematics plays a fundamental role in Computer Science and is an essential background for almost all of the advanced courses like theory of computation, compilers, databases, operating systems, algorithms and data structures etc.. One of the main reason for its importance is that the information in a computer is stored and manipulated in a discrete fashion.

As part of CS201 you studied many different disciplines in discrete mathematics; specifically you focussed on combinatorics, graph theory, number theory and abstract algebra. In Combinatorics, we talked about the art of counting. In this course, we take that topic further and focus on *probability theory*.

1 An introduction to probability

There is no need to emphasize the importance of probability in science. Just to give a glimpse, probability is useful in statistics, physics, quantum mechanics, finance, artificial intelligence/machine learning, computer science and even gambling.

You have already been introduced to probability in high school, and must have calculated chance of events like 2 heads in succession and prime number on a throw of a dice. This course will take your knowledge a step further. How about looking at some of the questions we will be concerned with in this course?

Exercise 1. You role two dices, what is the probability that the two outcomes are co-prime.

This is not a very difficult one. There are total 36 possibilities. Given some time, you can figure out the number of favourable cases (find it). The ratio of these two numbers will give us the probability. Let us increase the stakes.

Exercise 2. Take a look at Fig. 1. What is the probability that a random chord is larger than the side of the equilateral triangle ABC?

One way is to pick the two endpoints of the chord. Without loss of generality, we can fix one end-point to be A. Then, the chord is longer than the side if and only if the other endpoint falls between B and C. So, the probability should be 1/3.

But wait, a chord can also be defined by its center. If the chord is bigger than the side, its center should fall in cocentric circle with half the radius (convince yourself). So, chord is larger with probability 1/4.

Also, a chord can be defined uniquely by its distance from the center. That will give us probability 1/2 (work it out). Which one is the correct answer? Let us look at another example from the continuous world.

Exercise 3. Suppose you are given an infinite plane with parallel horizontal lines, with 1 centimeter distance between any two closest ones. You drop a needle of length one on this plane randomly. How many intersections do you expect to get with this parallel lines.

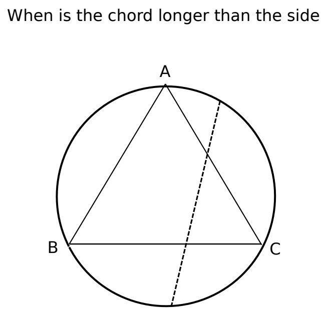

Fig. 1. Chord and the equilateral triangle problem.

This seems like a more difficult question (the needle can be dropped at any angle and any position). Though, a careful formulation and some integration can potentially give us a solution. We will develop a theory which will make this problem much easier to deal with.

Exercise 4. A cancer test is accurate with probability 95%. That means, if a person has cancer, the test will output YES 95% of the time. Similarly, if the person does not have cancer, the test will output YES 5% of the time. Suppose, I went for this test and result came out to be a YES. What is the probability that I have cancer?

It takes some time to realize what is being asked in this question. Many qualified professionals (doctors) predict that I have 95% chance of having cancer, which is incorrect!! Actually, there is not enough information in the question. You need to know the fraction of people who have cancer in the general population. We can assume that this fraction is pretty small, that will give a pretty small probability of me having cancer.

Exercise 5. There was a survey done in a hospital having just two kind of patients, Dengue and Malaria. They found that the probability of a person having Malaria is pretty high as compared to the probability of having Malaria if the patient already had Dengue. What do you think might be the reason?

This does not even look like a probability question. You might guess that the Malaria virus does not let Dengue virus grow and some other such biological reason. It turns out that such a negative correlation has to be expected purely because of probability and has nothing to do with the diseases. A hint, the oddity arises because we are only concentrating on people having at least one disease.

Exercise 6. A survey was conducted in different towns (cities, villages) of England about the percentage of people having diabetes. It turns out that the best five towns (lowest fraction of diabetes) were all small villages. What could be the reason?

You might again give credit to cleaner air, better lifestyle etc. for this statistics. Though, let us look at the survey results from one more angle.

Exercise 7. A survey was conducted in different towns (cities, villages) of England about the percentage of people having diabetes. It turns out that the worst five towns (largest fraction of diabetes) were all small villages. What could be the reason?

Given this statistics alone, you might have blamed it on poor health facilites, bad diet etc. We will see that both this observations can be explained with some knowledge of probability theory. How about these two seemingly unrelated questions which do not even talk about chance or probability?

- A family of subsets of $\{1, 2, \cdots, n\}$ is called an anti-chain if no element of the family contain another element of the family. How many anti-chains can be there?
- Prove that for every $B = \{b_1, b_2, \cdots, b_n\}$ (set of non-zero integers) contain a subset A (of size $\geq \frac{1}{3}n$) which is sum-free (no two elements of A sum up to an element of A).

The theory of *probabilistic methods* will answer these questions using simple ideas from probability.

These long list of questions might have convinved you that the concepts learnt in high school are not sufficient to tackle many problems on probability. This course will focus on concepts of probability theory to take your knowledge one step further.

On a lighter note, you might know this quote,

> *"The 50-50-90 rule: anytime you have a 50-50 chance of getting something right, there's a 90% probability you will get it wrong."*
> – Andy Rooney

We will learn concepts so that you have better than 50% chance of getting something right in most of the cases. In cases when you have 50% chance, you get it right with higher probability.

Outline of the course:

The course will start with basic formalism of probability theory; simultaneously, we will revise topics learnt before in high school. Next, the concepts of random variables, expectation, moments, distributions will be introduced. We will move to conditional probability and concentration inequalites, topics of immense importance in computer science.

Close to finish, we will look at statistics, as it is used widely in machine learning today. We will end the course by looking at Probabilistic methods, a pretty successful field giving existential proofs (in contrast to constructive proofs) in many diverse fields.

Our main focus will be about the role of probability in computer science and mathematics. The probability theory considered will mostly deal with sample spaces and sets which are discrete. Though, the basics of continuous probability spaces will also be covered.

2 Basics of probability

In probability, our main focus is to compute the chance of an *outcome* in a certain experiment. The experiment could be tossing a coin, throwing a dice or picking a random number. We might be interested in different kind of outcomes, e.g., getting head, getting a prime number on the top of the dice or picking a number bigger than 3/4.

For these situation, we had one of the most basic rule of probability: the probability of an event is the ratio of favourable outcomes to the total outcomes. Though, words like *experiment*, *outcome* and *favourable* are only loosely defined. Our first task is to give a mathematical foundation to these words.

2.1 Sample space and events

The first observation is, the physical implementation of an experiment (how did we toss a coin or threw a dice) is not important. Mathematically, we are only interested in the list of outcomes and their corresponding probabilites. Let us begin trying to model an experiment's result for the sake of probability theory.

The set of all possible *distinct* outcomes for an experiment is known as the *sample space*; mathematically, it is a set and it is generally denoted by symbol Ω.

Exercise 8. What is the sample space for a coin toss, sequence of coin toss, throwing of a dice and picking a random number.

For a given sample space, we might be interested in a certain outcome or a subset of outcomes from the sample space. A subset of the sample space is known as an *event*. Our task is to model the probability of different events.

We will be studying probability theory in the context of computer science. Hence, *with very high probability*, our sample sets will be discrete. In this case, we can define *probability distribution function* easily.

A probability distribution function for a sample space Ω is a map $P : \Omega \to \mathbb{R}$, s.t.,

- $P(\omega) \in [0,1]$ for all $\omega \in \Omega$,
- $\sum_{\omega \in \Omega} P(\omega) = 1$.

The probability of an event (a subset of Ω) can be naturally defined as,

$$P(S) = \sum_{\omega \in S} P(\omega).$$

The sample space and the probability distribution function completely models probability for an experiment. Whenever we deal with a question in probability, our first aim should be to identify the sample space, and if possible the probability distribution function.

Exercise 9. Suppose Amitabh (from Sholay) tosses a coin twice and is interested in finding the probability that both coins come out to be head. If coin comes head with 10% chance, then what is the sample space and the probability distribution function for this experiment?

Let's take an example which is of interest in real life (as opposed to mathematical life). Your cousin tells you to that she has cards numbered from 1 to 1000. She will pick a card at random and if it is divisible by 2 or 5 she will pay you 100 rupees. Otherwise you will pay her 200 rupees.

Should you accept the bet. If you want to make a bet, how much money can you pay her?

Let's model the situation as a probability distribution function. Define the sample space as $\Omega = \{1, 2, \cdots, 1000\}$, set of all possible card numbers. The set of all events will be the set of all subsets $\mathcal{F} = 2^{\Omega}$.

We will assume that the card is picked uniformly at random, that is, the probability of obtaining a particular number in the range 1 to 1000 is 1/1000. This defines a probability distribution function for all $S \in \mathcal{F}$,

$$P(S) = \frac{|S|}{1000}.$$

Observe that we need to find the size of the set of numbers divisible by 2 or 5 and lie between 1 and 1000.

Exercise 10. Show that the numbers divisible by 2 or 5 between 1 and 1000 is 600.

The probability of you winning the game is $600/1000 = 3/5$. So *odds* of you winning are $3 : 2$ worse than $2 : 1$. So you should not accept the bet. But the bet will be favorable to you if you pay her less than 150 rupees.

2.2 Union of events

What can you say about the probability of the event $A \cup B$? Clearly, if A and B are disjoint,

$$P(A \cup B) = P(A) + P(B).$$

If they are not disjoint, we need to subtract the probability of the intersection (which was counted twice),

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

Similarly, most of the rules of set theory directly give us results about probability.

Exercise 11. Prove the following.

- $P(A^C) = 1 - P(A)$.
- $P((A \cup B)^C) = P(A^C) + P(B^C) - P(A^C \cup B^C)$.

The *union rule* above can be generalized to multiple sets. For example, if there are three sets, our first approximation of $P(A \cup B \cup C)$ would be,

$$P(A \cup B \cup C) \approx P(A) + P(B) + P(C).$$

Though, this counts the probability of elements in the intersection twice, a better approximation would be,

$$P(A \cup B \cup C) \approx P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A).$$

You can guess that the only elements to worry for, whose probabability might not be correctly counted, are the ones in the intersection of all three. This gives the final formula,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(C \cap A) + P(A \cap B \cap C).$$

This should remind you of *inclusion-exclusion principle.* A similar logic gives us the formula for the probability of the union of n sets,

$$P(\bigcup_{i \in [n]} A_i) = \sum_{S \subseteq [n], S \neq \varphi} (-1)^{|S|+1} P(\bigcap_{i \in S} A_i).$$

Let us see an application to this formula. Suppose we have n letters and n corresponding envelopes. If you place each letter randomly in an envelope, what is the probability that no letter goes into the correct envelope?

Exercise 12. What is the sample space and the probability distribution function?

Define A_i to be the event that letter i does not go into its corresponding envelope.

Exercise 13. What is the probability, in terms of $P(A_i)$, of no letter going to the correct envelope?

Some thought shows that we are interested in the quantity,

$$P\left(\left(\bigcup_{i \in [n]} A_i\right)^C\right) = 1 - P(\bigcup_{i \in [n]} A_i).$$

We already know the expression for the right hand side,

$$1 - P(\bigcup_{i \in [n]} A_i) = \sum_{S \subseteq [n]} (-1)^{|S|} P(\bigcap_{i \in S} A_i).$$

Exercise 14. What is the probability of $P(\bigcap_{i \in S} A_i)$?

If all letters in S go to their place, we need to only arrange $n - |S|$ places. The probability should be $\frac{(n-|S|)!}{n!}$. So, the probability that no letters goes to its correct place is,

$$\sum_{S \subseteq [n]} (-1)^{|S|} \frac{(n - |S|)!}{n!} = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \frac{(n-k)!}{n!}.$$

Simplifying the expression, we get $1 - 1/1! + 1/2! - 1/3! + \cdots$, approaches $1/e$ as n tends to infinity.

Using the concepts of sample space and probability distribution function, we have modelled probability/chance/odds in an experiment. To summarize, say, we perform an experiment and are interested in the probability of various events in the experiment. The set of outcomes of the experiment will be called the sample space Ω. Any subset of Ω is an event. The probability distribution function specifies probability for any such event.

As noted above, this is an easier way to define probability distribution function. This can lead to trouble in a non-discrete sample space. Also, not all subsets of the sample space need to be interesting and it might be a tedious task to define probability on every subset.

3 Formal definition of a probability distribution function

In the easier case, we were able to define probability for every element of 2^{Ω} (any subset of the sample space). In certain situations, we might only be interested in certain set of events. Though, for the probability function to make sense, if A, B are events, then A^C and $A \cup B$ should also be events.

This intuition gives rise to the concept of a *sigma-algebra.* A collection of *subsets* $\mathcal{F}$ of the sample space Ω is called a sigma-algebra, if,

1. Ω is in $\mathcal{F}$.
2. Complement of a set in $\mathcal{F}$ is in $\mathcal{F}$.
3. Countable unions of sets in $\mathcal{F}$ is in $\mathcal{F}$.

Exercise 15. Show that $\mathcal{F}$ is closed under countable intersection.

With all these definitions, we are ready to define probability function.

A function $P : \mathcal{F} \rightarrow [0, 1]$ is called a probability distribution function (or just probability distribution), if it satisfies

1. $P(\Omega) = 1$,
2. If A, B are disjoint then $P(A \cup B) = P(A) + P(B)$.

Exercise 16. Show that the second rule above implies the corresponding property for countable union. Why does it stop for countable union?

Exercise 17. Prove that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

This general model of probability needs a sample space as before, then defines set of interesting events as a sigma-algebra. The sigma-algebra should be a subset of 2^{Ω} and satisfy the three conditions mentioned above.

A probability distribution function assigns probability to every interesting event (element of sigma-algebra) in a consistent way. By consistent, we mean that the probability function satisfies two conditions above.

Exercise 18. What is the sigma-algebra for our easier definition of probability distribution function?

We will mostly worry about these cases, when sigma-algebra is the whole power set of Ω.

Consider another example of a family. What is the probability that in a family with 5 kids, there are more girls than boys?

Use g for a girl and b for a boy. Then, the sample space Ω is the set of all possible strings of g, b with length 5. Again, the *sigma*-algebra will be the set of all possible subsets of Ω.

All possible strings are equally likely. Hence, we are interested in number of strings with length 5, which have more g's than b's. There are 32 possibilities, you can check that 16 of them have more girls than boys. So, the probability is $1/2$.

The same result can be obtained directly by observing the symmetry between boys and girls. Their is another way to model the same situation. The sample space will stay the same, Ω is the set of all possible strings of g, b with length 5. The difference is, sigma-algebra is going to have only 4 elements $\{\emptyset, \Omega, A, A^C\}$.

Exercise 19. Show that for any A this is a sigma-algebra.

Choose A to be the subset of Ω which has more girls than boys. By symmetry, the probability of A and A^C is the same. So we get that there are more girls than boys with probability $1/2$.

What is the probability if there are 6 kids? If there are 6 kids then b's and g's could be equal. The number of such cases are $\frac{6!}{3!3!} = 20$. So the number of cases when girls are more than boys is $(64 - 20)/2 = 22$ and hence the probability is $22/64$.

3.1 Probability on a continuous sample space

Till now we have only dealt with discrete sample spaces. In real life, there are many situations when the associated sample space is naturally continuous (uncountable number of points). To take few examples,

- What is the probability of picking a number less than .5, if we pick a random number between 0 and 1?
- If we spin a wheel, what is the proability that the arrow rests in the first quadrant?
- If we break a stick of unit length at two points at random, what is the probability that the three sticks can form a triangle?

For some of these you might know the answer intuitively (1/2 for the first one). How should these experiments be put up in our formal framework? Start with one of the simplest experiment, picking a random number between 0 and 1. The obvious sample space is the set of points $[0, 1]$.

What should be the probability distribution function? It seems that all points should occur with equal probability by symmetry. Though, any non-zero value to them will make total probability greater than 1. In other words, every point should have probability 0!

Exercise 20. What could save us here?

Remember, at least intuitively, the probability of picking a number less than .5 seemed to be 1/2. Extending that, point would fall in an interval $[a, b]$ with probability $b - a$ (since length of $[0, 1]$ is 1). This seems to be correct, probability of getting to an interval should be proportional to its length. This way we can define probability of any disjoint union (countable) of intervals.

Exercise 21. What about other subsets of the power set of sample space?

The concept of sigma-algebra comes to our rescue. We DONT need to define probability of every subset. We can only define probability of all events in the sigma-algebra generated by intervals. When the probability for these intervals is proportional to their length, it is called a *uniform distribution.*

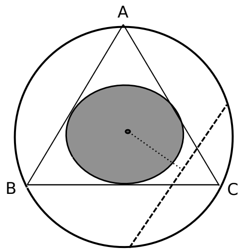

Fig. 2. Chord and the equilateral triangle problem.

Exercise 22. Do you remember the chord problem? Take a look at Fig. 1. What is the probability that a random chord is larger than the side of the equilateral triangle ABC?

There are multiple ways to fix the sample space. First one is to fix one end-point of the chord to be A and randomly pick the other point. Then, the chord is longer than the side if and only if the other endpoint falls between B and C. So, the probability should be 1/3.

Other sample space could be, when we pick the center of the chord uniformly in the circle. For chord to be bigger than the side, its center should fall in the shaded co-centric circle (Fig. 2). So, chord is larger with probability 1/4.

Another sample space could be to pick the point on a line from the center to the circumference, i.e., a chord can be defined uniquely by its distance from the center. Since the sides are half-way from the center (a little bit of trignometry), it will give us probability 1/2.

All of these are correct answers, the answer depend upon the sample space chosen. It reemphasizes the fact that we should be clear about the sample space before dealing with a problem on probability.

Some, not all, of you might have a doubt. There is a one to one correspondonce between these sample spaces, shouldn't we get the same answer. No, one to one correspondonce is not enough. Consider the probability that $4 \leq x^2 \leq 9$ given $0 \leq x \leq 6$.

Exercise 23. What is the probability if we pick a random number in $[0, 6]$? What is the probability if a random x^2 is picked in $[0, 36]$?

4 Assignment

Exercise 24. Read about Monty Hall problem.

Exercise 25. Where have you used probability in your life?

Exercise 26. Calculate the probability of getting two consecutive heads when you toss a coin 4 times.

Exercise 27. A *derangement* is a permutation of the elements of a set, such that, no element appears in its original position. If we pick a random permutation of m elements, what is the probability that we get a derangement?

Exercise 28. Suppose we break a unit stick at two random points. Find the probability that the broken parts can form a triangle.

Hint: Fix a point, find the probability that you get a triangle when other point is picked randomly, integrate.