EECS 70 Discrete Mathematics and Probability Theory
Spring 2014 Anant Sahai

Note 10

Introduction to Basic Discrete Probability

In the last note we considered the probabilistic experiment where we flipped a fair coin $10,000$ times and counted the number of Hs. We asked "what is the chance that we get between 4900 and 5100 Hs".

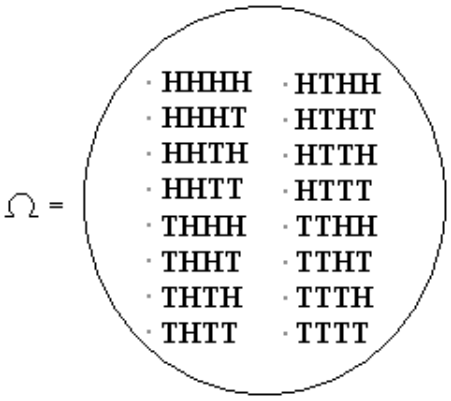
At a more basic level, we asked "what was the chance that we got a head?"

In this note we will begin to formalize these notions for an arbitrary probabilistic experiment. We will start by introducing the space of all possible outcomes of the experiment, called a sample space. Exactly one of these will be realized. In our current model, each element of the sample space is assigned a probability - which tells us how likely it is to occur when we actually perform the experiment. The mathematical formalism we introduce might take you some time to get used to. But you should remember that ultimately it is just a precise way to say what we mean when we describe a probabilistic experiment like flipping a coin n times.

Random Experiments

The outcome of the random experiment is called a *sample point*. The *sample space*, often denoted by Ω, is the set of all possible outcomes.

An example of such an experiment is tossing a coin 4 times. $HTHT$ is an example of a sample point and the sample space has 16 elements:



How do we determine the chance of each particular outcome, such as $HHTT$, of our experiment? In order to do this, we need to define the probability for each sample point, as we will do below.

Probability Spaces

A probability space is a sample space Ω, together with a probability $\Pr[\omega]$ for each sample point ω, such that

- $0 \leq \Pr[\omega] \leq 1$ for all $\omega \in \Omega$.
- $\sum_{\omega \in \Omega} \Pr[\omega] = 1$, i.e., the sum of the probabilities of all outcomes is 1.

The easiest, and intuitively paradigmatic, case is that of uniform probability.

Uniform Probability

The easiest way to assign probabilities to sample points is uniformly: if $|\Omega| = N$, then $\Pr[x] = \frac{1}{N}\ \forall x \in \Omega$. For example, if we toss a fair coin 4 times, each of the 16 sample points (as pictured above) is assigned probability $\frac{1}{16}$. We will see examples of non-uniform probability distributions soon.

After performing an experiment, we are often interested in knowing whether an event occurred. For example, the might be interested in the event that there were "exactly 2 H's in four tosses of the coin". How do we formally define the concept of an event in terms of the sample space Ω? Here is a beautiful answer. We will identify the event "exactly 2 H's in four tosses of the coin" with the subset consisting of those outcomes in which there are exactly two H's:
$\{HHTT, HTHT, HTTH, THHT, THTH, TTHH\} \subseteq \Omega$. Now we turn this around and say that formally an event A is just a subset of the sample space, $A \subseteq \Omega$.

Another way to think about it is to associate "properties" with experimental outcomes. An event of interest is about one such property[1]. The set A is then just the set of those outcomes that have that property. Following the same example as above, the property is "having exactly 2 Hs" where the outcomes are strings of length 4.

How should we define the probability of an event A? Naturally, we should just *add up* the probabilities of the sample points in A. For uniform probability, this is the same as asking what is the frequency of the relevant property among possible outcomes.

Formally, for any event $A \subseteq \Omega$, we define the probability of A to be

$$\Pr[A] = \sum_{\omega \in A} \Pr[\omega].$$

Thus the probability of getting exactly two H's in four coin tosses can be calculated using this definition as follows. A consists of all sequences that have exactly two H's, and so $|A| = 6$. For this example, there are $2^4 = 16$ possible outcomes for flipping four coins. Thus, each sample point $\omega \in A$ has probability $\frac{1}{16}$; and, as we saw above, there are six sample points in A, giving us $\Pr[A] = 6 \cdot \frac{1}{16} = \frac{3}{8}$.

For the special case when the probability is uniform,

$$\Pr[A] = \frac{|A|}{|\Omega|}$$

So probability is just a kind of generalized proportion and is exactly proportion in the uniform case.

Rolling Dice Example

The next random experiment we will discuss consists of rolling two dice. In this experiment, $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$. The probability space is uniform, i.e. all of the sample points have the *same* probability, which

[1] If you have noticed a connection here between sets and propositions, this is no coincidence.

must be $\frac{1}{|\Omega|}$. In this case, $|\Omega| = 36$, so each sample point has probability $\frac{1}{36}$. In such circumstances, the probability of any event A is clearly just

$$\Pr[A] = \frac{\text{\# of sample points in } A}{\text{\# of sample points in } \Omega} = \frac{|A|}{|\Omega|}.$$

So for uniform spaces, computing probabilities reduces to *counting* sample points!

Now consider two events: the event A that the sum of the dice is at least 10 and the event B that there is at least one 6. By writing out the number of sample points in each event, we can determine the number of sample points in each event; $|A| = 6$ and $|B| = 11$. By the observation above, it follows that $\Pr[A] = \frac{6}{36} = \frac{1}{6}$ and $\Pr[B] = \frac{11}{36}$.

Nonuniform Probability

The general formalism above works for any assignment of probability "weights" on outcomes, not just uniform probability.

However, for beginners, it is useful to think about nonuniform probability as being a view of a story that is fundamentally uniform underneath. For example, back to the fair coin toss example, the uniform perspective says that the experimental outcome is the string of tosses. However, someone could also look at the same experiment and instead say that the outcome is the number of heads. In that case, the sample space just has five possibilities $\{0,1,2,3,4\}$ and the probability is not uniform. We have $\Pr[0] = \Pr[4] = \frac{1}{16}$, $\Pr[1] = \Pr[3] = \frac{4}{16} = \frac{1}{4}$, and as we saw above $\Pr[2] = \frac{6}{16} = \frac{3}{8}$. Notice that these numbers sum up to one.

When all of the probabilities involved are rational[2] numbers and the set of outcomes is finite, it is always possible to come up with such a uniform back-story for nonuniform probabilities.

The Monty Hall Problem

In an (in)famous 1970s game show hosted by one Monty Hall, a contestant was shown three doors; behind one of the doors was a prize, and behind the other two were goats. The contestant (who prefers cars to goats) picks a door (but doesn't open it). Then Hall's assistant (Carol), opens one of the other two doors, revealing a goat (since Carol knows where the prize is, and there are two goats, she can always do this). The contestant is then given the option of sticking with his current door, or switching to the other unopened one. He wins the prize if and only if his final chosen door is the correct one. The question, of course, is: Does the contestant have a better chance of winning if he switches doors?

Intuitively, it may seem obvious that since there are only two remaining doors after the host opens one, they must have equal probability. So you may be tempted to jump to the conclusion that it should not matter whether or not the contestant stays or switches.

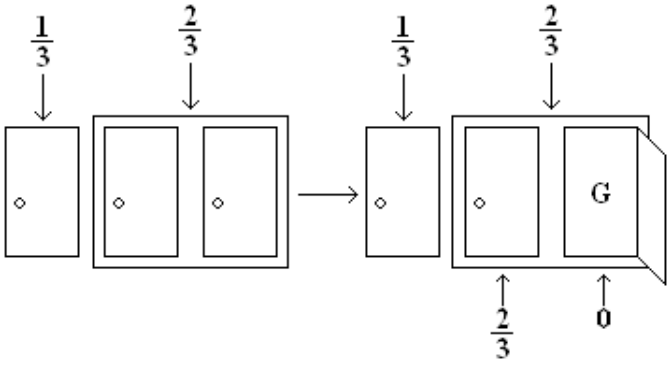
Yet there are other people whose intuition cries out that the contestant is better off switching. So who's correct?

As a matter of fact, the contestant has a better chance of picking the car if he uses the switching strategy. How can you convince yourself that this is true? One way you can do this is by doing a rigorous analysis. You would start by writing out the sample space, and then assign probabilities to each sample point. Finally you would calculate the probability of the event that the contestant wins under the sticking strategy. This

[2] In the irrational case or when the set of outcomes is infinite, it is still possible to come up with a uniform back-story. But that back-story will not be discrete. Instead, we will have to use continuous probability as we will see in a later lecture note.

is an excellent exercise if you wish to make sure you understand the formalism of probability theory we introduced above.

Let us instead give a more intuitive pictorial argument. Initially when the contestant chooses the door, he has a $\frac{1}{3}$ chance of picking the car. This must mean that the other doors combined have a $\frac{2}{3}$ chance of winning. But after Carol opens a door with a goat behind it, how do the probabilities change? Well, everyone knows that there is a goat behind one of the doors that the contestant did not pick. So no matter whether the contestant is winning or not, Carol is always able to open one of the other doors to reveal a goat. This means that the contestant still has a $\frac{1}{3}$ chance of winning. Also the door that Carol opened has no chance of winning. What about the last door? It must have a $\frac{2}{3}$ chance of containing the car, and so the contestant has a higher chance of winning if he or she switches doors. This argument can be summed up nicely in the following picture:



We will be able to formalize this intuitive argument once we cover conditional probability.

In the meantime, to approach this problem formally, first determine the sample space and the probability space.

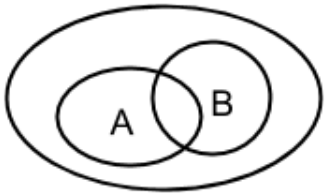
Then formalize the event we have described above (as a subset of the sample space), and compute the probability of the event.

Conditional Probability — The Basics

The intuition behind conditional probability is simple. It is about how to properly "update" your uncertainty based on pieces of knowledge. In the beginning, you don't know what the outcome is of the random experiment. Anything in Ω is possible. If you want to guess whether or not the outcome will have a certain feature A, you consider it like getting heads in a biased coin toss with probability of heads equal to $\Pr[A]$.

Now, somebody (who knows what outcome has happened) tells you that the outcome has a certain feature, which may or may not be related to the feature you are interested in. Equivalently, using the language of events, this person tells you that an event B has ocurred. At this point, how do you update your uncertainty? We would like to say that at this point, the probability is $\Pr[A|B]$.

How should we compute $\Pr[A|B]$? Well, since event B is guaranteed to happen, we need to look not at the whole sample space Ω, but at the smaller sample space consisting only of the sample points in B. In terms of the picture below, we are no longer looking at the large oval, but only the oval labeled B:



What should the probabilities of these sample points be? If they all simply inherit their probabilities from Ω, then the sum of these probabilities will be $\sum_{\omega \in B} \Pr[\omega] = \Pr[B]$, which in general is less than 1. So we need to *scale* the probability of each sample point by $\frac{1}{\Pr[B]}$. I.e., for each sample point $\omega \in B$, the new probability becomes

$$\Pr[\omega|B] = \frac{\Pr[\omega]}{\Pr[B]}.$$

Now it is clear how to compute $\Pr[A|B]$: namely, we just sum up these scaled probabilities over all sample points that lie in both A and B:

$$\Pr[A|B] = \sum_{\omega \in A \cap B} \Pr[\omega|B] = \sum_{\omega \in A \cap B} \frac{\Pr[\omega]}{\Pr[B]} = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

Definition (conditional probability): For events A, B in the same probability space, such that $\Pr[B] > 0$, the conditional probability of A given B is

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

This is particularly intuitive in cases of uniform probability. Then, the division by $\Pr[B]$ is just adjusting for the new size of the universe of items.

Notice that this definition only makes sense when $\Pr[B]$ is nonzero. When we are talking about discrete probability spaces, this turns out not to be a problem in practice. Zero probability events are not interesting. However, when we start talking about continuous probability spaces, figuring out how to condition on a zero probability event will turn out to be both important and a bit subtle (we'll have to take limits).

Bayesian Inference

Now that we've introduced the notion of conditional probability, we can see how it is used in real world settings. Conditional probability is at the heart of a subject called *Bayesian inference*, used extensively in fields such as modern machine learning, and communications and signal processing more generally. In this interpretation, $\Pr[A]$ can be thought of as a *prior* probability: our assessment of the likelihood of an event of interest A *before* making an observation. It reflects our prior knowledge. $\Pr[A|B]$ can be interpreted as the *posterior* probability of A after the observation that B has definitely ocurred. It reflects our new knowledge.

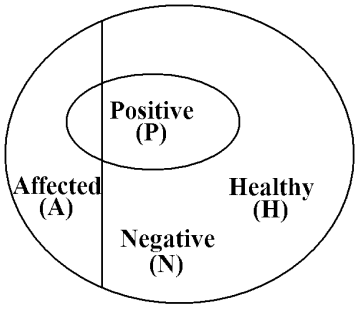
Here is an example of where we can apply such a technique. A pharmaceutical company is marketing a new test for a certain medical disorder. According to clinical trials[3], the test has the following properties:

1. When applied to an affected person, the test comes up positive in 90% of cases, and negative in 10% (these are called "false negatives").

[3] How such probabilities come to be known or estimated at all is a more subtle problem that we will ignore for now.

2. When applied to a healthy person, the test comes up negative in 80% of cases, and positive in 20% (these are called "false positives").

Suppose that the incidence of the disorder in the US population is 5%; this is our prior knowledge. When a random person is tested and the test comes up positive, how can we update this probability? (Note that this is presumably *not* the same as the simple probability that a random person has the disorder, which is just $\frac{1}{20}$.) The implicit probability space here is the entire US population with uniform probabilities.



The sample space here consists of all people in the US — denote their number by N (so $N \approx 250$ million). Let A be the event that a person chosen at random is affected, and B be the event that a person chosen at random tests positive. Now we can rewrite the information above:

- $\Pr[A] = 0.05$, (5% of the U.S. population is affected)
- $\Pr[B|A] = 0.9$ (90% of the affected people test positive)
- $\Pr[B|\overline{A}] = 0.2$ (20% of healthy people test positive)

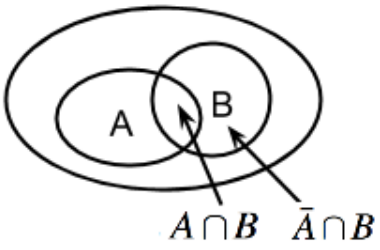
We want to calculate $\Pr[A|B]$. We can proceed as follows:

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} \tag{1}$$

We obtained the second equality above by applying the definition of conditional probability:

$$\Pr[B|A] = \frac{\Pr[A \cap B]}{\Pr[A]}$$

Now we need to compute $\Pr[B]$. This is the probability that a random person tests positive. To compute this, we can sum two values: the probability that a healthy person tests positive, $\Pr[\bar{A} \cap B]$ and the probability that an affected person tests positive, $\Pr[A \cap B]$. We can sum because the events $\bar{A} \cap B$ and $A \cap B$ do not intersect:



By again applying the definition of conditional probability we have:

$$\Pr[B] = \Pr[A \cap B] + \Pr[\overline{A} \cap B] = \Pr[B|A]\Pr[A] + \Pr[B|\overline{A}](1 - \Pr[A]) \tag{2}$$

Combining equations (1) and (2), we have expressed $\Pr[A|B]$ in terms of $\Pr[A], \Pr[B|A]$ and $\Pr[B|\overline{A}]$:

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B|A]\Pr[A] + \Pr[B|\overline{A}](1 - \Pr[A])} \tag{3}$$

By plugging in the values written above, we obtain $\Pr[A|B] = \frac{9}{47} \approx .19$.

Equation (3) is useful for many inference problems. We are given $\Pr[A]$, which is the (unconditional) probability that the event of interest A happens. We are given $\Pr[B|A]$ and $\Pr[B|\overline{A}]$, which quantify how noisy the observation is. (If $\Pr[B|A] = 1$ and $\Pr[B|\overline{A}] = 0$, for example, the observation is completely noiseless.) Now we want to calculate $\Pr[A|B]$, the probability that the event of interest happens given we made the observation. Equation (3) allows us to do just that.

Of course, equations (1), (2) and (3 are derived from the basic axioms of probability and the definition of conditional probability, and are therefore true with or without the above Bayesian inference interpretation. However, this interpretation is very useful when we apply probability theory to study inference problems.

Bayes' Rule and Total Probability Rule

Equations (1) and (2) are very useful in their own right. The first is called **Bayes' Rule** and the second is called the **Total Probability Rule**. Bayes' rule is useful when one wants to calculate $\Pr[A|B]$ but one is given $\Pr[B|A]$ instead, i.e. it allows us to "flip" things around.

The Total Probability rule is an application of the strategy of "dividing into cases" .There are two possibilities: either an event A happens or A does not happen. If A happens the probability that B happens is $\Pr[B|A]$. If A does not happen, the probability that B happens is $\Pr[B|\overline{A}]$. If we know or can easily calculate these two probabilities and also $\Pr[A]$, then the total probability rule yields the probability of event B.

Example: Tennis Match

You are about to play a tennis match against a randomly chosen opponent and you wish to calculate your probability of winning. You know your opponent will be one of two people, X or Y. If person X is chosen, you will win with probability .7. If person Y is chosen, you will win with probability .3. Your opponent is chosen by flipping a coin with bias .6 in favor of X.

Let's first determine which events we are interested in. Let A be the event that you win. Let B_1 be the event that person X is chosen, and let B_2 be the event that person Y is chosen. We wish to calculate $\Pr[A]$. Here is what we know so far:

- $\Pr[A|B_1] = 0.7$, (if person X is chosen, you win with probability .7)
- $\Pr[A|B_2] = 0.3$ (if person Y is chosen, you win with probability .3)
- $\Pr[B_1] = 0.6$ (person X is chosen with probability .6)

- $\Pr[B_2] = 0.4$ (person Y is chosen with probability .4)

By using the Total Probability rule, we have:

$$\Pr[A] = \Pr[A|B_1]\Pr[B_1] + \Pr[A|B_2]\Pr[B_2].$$

Now we can simply plug in the known values above to obtain $\Pr[A]$:

$$\Pr[A] = .7 \times .6 + .3 \times .4 = .54$$