EECS 70 Discrete Mathematics and Probability Theory
Spring 2014 Anant Sahai **Note 9**

Probability Theory: Brief Intro

Probability[1] theory is one of the great conceptual achievements of the 20th century. In statistical physics it provides a technique for understanding the aggregate behavior of large numbers of particles, without analyzing their individual behavior. In quantum mechanics, it is an integral part of the fabric of the theory. Probabilistic methods form the backbone of many recent achievements in algorithms (from the theory of computation), are foundational for understanding the nature of information in both communication and cryptograph, and play a critical role in estimation/learning and signal processing more generally. When combined with ideas from linear-algebra and optimization (studied in EECS 127), probabilistic perspectives on signals form the foundation of modern machine learning. This perspective, when further combined with ideas from control, is also the driving force behind contemporary approaches[2] to artificial intelligence. Probability ideas are also critical in understanding economics and finance. And of course, statistical techniques based on probability theory are absolutely fundamental to experimental science from the physical sciences through biology and even in the social sciences.

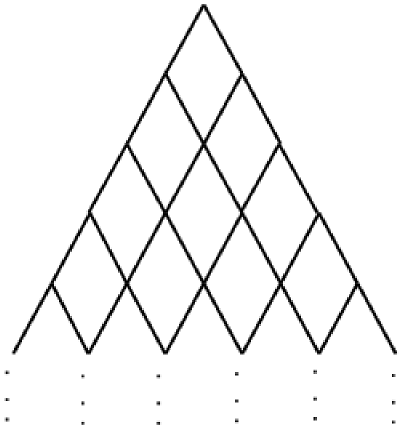
One of the basic themes in probability theory is the emergence of almost deterministic behavior from probabilistic phenomena. For example, if we flip a fair coin, we believe that the underlying frequency of heads and tails should be equal. When we flip it 10,000 times, we are pretty certain in expecting between 4900 and 5100 heads. A random fluctuation around the true frequency will be present, but it will be relatively small. This phenomenon (referred to colloquially as a law of large numbers) is responsible for the stability of the physical world we live in, whose building blocks - elementary particles, atoms, molecules - individually behave in many ways like coin flips, but in aggregate present a solid front to us, much like the proportion of heads in the example above.

Coin Flipping

Let's look at the coin flipping example in more detail. To picture coin flipping, consider the simple apparatus below called the Galton-Watson board.

[1] Seriously, while here in 70 we can give you a taste for probability and certain basic tools. You really are strongly encouraged to take EECS126 to get a more in-depth understanding, especially of the connection between Probability and Linear Algebra which we cannot cover in this course because Linear Algebra is not a prereq. In 126, you will see things like how Probability is intimately connected to things like Google's PageRank, Speech Recognition, Planning for Robots, various topics in networking, etc.

[2] The older, more old-timey "classic CS" approach of rule-based "expert systems" for artificial intelligence has been almost completely replaced by the more modern "EE+Stat" approach that combines large amounts of data with linear-algebraic and comm-theoretic optimization algorithms for graphical-model signal processing.

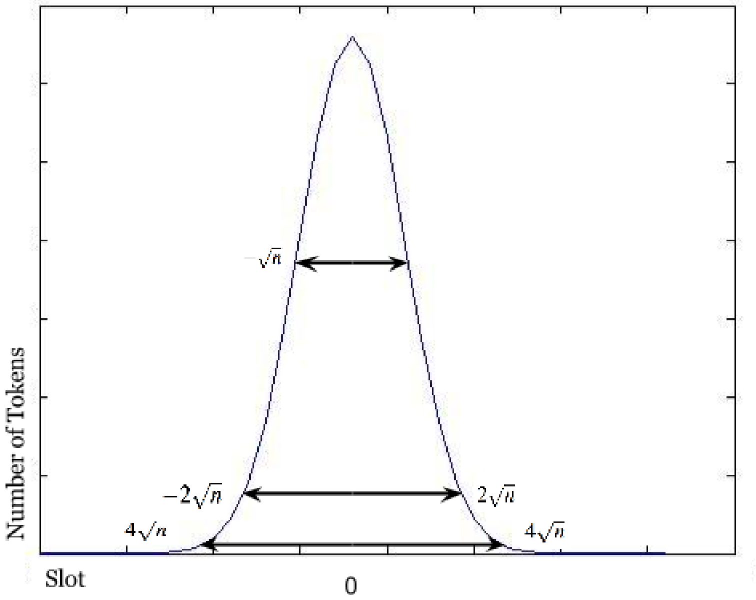


Imagine we drop a token in the slot at the top of the board. As the token falls, at each peg it chooses randomly with 50-50 chance whether to go right or left. If we were to watch a single token drop, it would end up in one of the slots at the bottom of the board, chosen in some random way. What happens if we watch N tokens drop? How many tokens does each of the bottom slots hold?

Before we answer the question, let us observe that the Galton-Watson board is just a way of visualizing a sequence of coin flips. To see this, imagine that each time the token encounters a peg, it flips a coin to decide whether to go left or right. If the coin is heads, the token goes right, and if the coin is tails, the token goes left.

If the height of the Galton-Watson board is n, then there are exactly n coin flips as the token makes its way down. Moreover, if we think of heads as +1 and tails as -1, then the slot that the token ends up in is just the sum of the choices at each round, which ranges from $-n$ to n. So we can label each slot by a number between $-n$ and n in this way, with the leftmost slot labelled $-n$ and the right most slot is labelled n. Notice that the slot number also represents the number of tails subtracted from the number of heads.

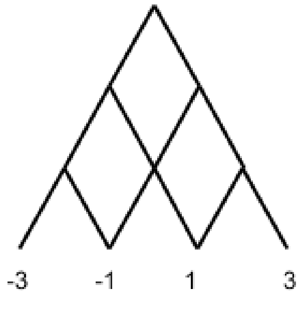
Now let's get back to our question - what happens if we watch N tokens drop? Here is a figure showing the answer for a board of height n:



As you can see, when $n = 10,000$, we expect almost all the tokens to fall in slots between -200 and 200. If we go back to the coin flipping setting, we can see that we should expect the number of tails subtracted from the number of heads to be between -200 and 200; equivalently, the number of heads is expected to be between 4900 and 5100.

What is the significance of the range 4900 to 5100? Note that the size of that interval 5100-4900 = 200 is the twice the square root of $n = 10,000$. This is no coincidence. In general if we flip a coin n times, the probability of landing in a slot between $-2\sqrt{n}$ and $2\sqrt{n}$ is around 95% (we're outside less than 1 time out of every 20). And the probability of landing outside of slots between $-4\sqrt{n}$ and $4\sqrt{n}$ is quite tiny, less than one-hundredth of a percent — actually less than once out of 15,000 times! This rapid decline in probability is evident in the picture above - as you move further away from the center of the picture, the number of tokens in the corresponding slots decreases rapidly.

Why does the probability fall so quickly? To understand this, we can turn back to the Galton-Watson branching process. First consider a smaller example, where $n = 3$:



There are eight total paths from the start slot to one of the final slots. There is only one path leading to the slot labelled -3 (the token must choose the left branch at each step). The same applies for the right most slot. However, there are 3 paths to each middle slot (labelled -1 and 1).

As n grows, the total number of paths that the token can take increases exponentially. In fact, since there are 2 choices for each coin flip, there are $2 \times 2 \times \cdots \times 2 = 2^n$ choices of paths for the token if the board has height n. Even for relatively small values of n like 50 or 100, 2^n is already astronomically large. Moreover, as the number of paths grows, the discrepancy between the number of paths leading to each slot is magnified.

For example, for a token to land in either the left most or right most slot, it would have to choose one out of 2^n paths! The vast majority of the paths end up at one of the slots near the middle. This discrepancy between the likelihood of landing in a central spot rather than a spot closer to one of the two ends is responsible for the almost deterministic behavior described above; this is why we can say that if we flip a coin 10,000 times, we expect the number of heads to be between 4900 and 5100.

We will make this sort of calculation rigorous in later notes, but you will see it for yourself experimentally in the homework.

Bias Estimation

So far we have been considering the case where the coin we flip has equal probability of landing on heads or tails. What happens when it is biased- when the probability of heads is $p = \frac{3}{4}$ rather than $\frac{1}{2}$? In this case, if we flip the coin $n = 10,000$ times, we expect to see 7500 heads and 2500 tails. Once again, if we repeat this experiment N times, we expect almost all the experiments to end up with between 7400 and 7600 heads.

This behavior suggests that if we did not know the bias p to begin with, we could simply divide the number of observed number of heads by the total number of heads, say $\hat{p} = 7470/10000$ to get an estimate for p.

If we could ensure that such an estimate is accurate, it would be useful in various settings. For example, we can use it in election polling. Let's say we would like to estimate the fraction of the population who supports one of two candidates. Polling a random sample of the population is analogous to flipping a coin with bias p, where p is the fraction of the population supporting the candidate in the question. Given the polling results, we can use the above technique to estimate p.

Can we ensure that our estimate $\hat{p}$ is close to p? The fact that the actual number of heads we observe is almost certain to be within a very narrow range means that we can be pretty confident that our estimate is very close to the actual bias p. This is one of the fundamental principles of statistics.

There are two parameters which tell us how good our estimate $\hat{p}$ is. The first is the error of our estimate, $|p - \hat{p}|$. For example, maybe we would like this error to be smaller than some bound ε. i.e. we are looking for an estimate $\hat{p}$ such that $|p - \hat{p}| \leq \varepsilon$.

The second parameter corresponds to our confidence that our estimate is within an error of ε. This is necessary, because when estimating p, we can never be completely sure that our estimate $\hat{p}$ is within ε of p. For example, even if $p = \frac{1}{2}$, it might be the case that all our 10,000 coin flips come up heads. In this case, our estimate $\hat{p}$ would be 1. Of course the chance of this is smaller than anything we could imagine ($\frac{1}{2^{10,000}}$ to be exact), but it is still possible. The confidence parameter addresses this possibility. For example, we might say that we are 98% confident that our estimate $\hat{p}$ is within $\varepsilon = .01$ of the actual bias p. We can introduce another parameter $0 \leq \delta \leq 1$ for our confidence. So if we are 98% confident, we will say that $\delta = .02$, so that our confidence is $1 - \delta$ as a fraction or $100(1 - \delta)$ as a percentage.

Both of these parameters - ε and the confidence percentage- depend on the number of trials. Another way of saying this is that if we wish to achieve a certain error ε, say .1, and a certain confidence, say 95% or $\delta = .05$, then there is a value for n the number of trials, in this case $n = 500$, that will guarantee that with confidence at least 95% our estimate will be within .1 of the actual bias. More generally, as you will see proven in later lectures, to obtain error ε and confidence δ (where δ is between 0 and 1, so our confidence percentage is 100(1-δ)), $n = \frac{1}{4\varepsilon^2\delta}$ trials would be sufficient.

Outline

Over the course of the next half of the course, we will develop the tools needed to understand the questions discussed above. This will be done in two phases. The first will be a rapid survey. We'll start with this "coin toss" model of randomness with probability as the bias of a coin. We'll begin with a rapid introduction to the basic idea of probability as reflecting underlying frequencies in a population. As we'll see, this is a kind of "accounting" perspective that doesn't concern itself with quantifying random fluctuations. Even so, it helps us understand certain very basic questions of inference. After that rapid introduction into basic probability, we'll take a purely empirical perspective on the laws of large numbers, restricted to coin tosses. This will help us get a practical handle on random fluctuations and how to reason about them in engineering contexts, but we won't prove anything about them yet.

Once we've done this brief overview of both "frequency" and "fluctuation," we'll take a more systematic approach with the aim of understanding why things behave the way that they do as well as building towards appropriate generalizations from the coin-tossing case. This will proceed through counting (where a new kind of proof will also be introduced — the combinatorial proof), the introduction of random variables, the concept of expectation, and so on. These will allow us to formalize previous concepts and answer more difficult questions, such as estimating the bias of a coin.