Appendix A
A Crash Course in Probability

Fig. A.1 shows a spinner with a perimeter length of 1:

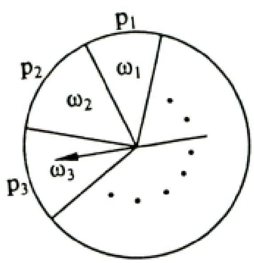


Figure A.1. Selecting a probability

Each segment ω_i has an arc length of p_i, such that

$$p_i \geq 0, \sum_{i \geq 1} p_i = 1. \tag{A.1}$$

Spinning the spinner once is a *one stage random experiment*, where ω_i denotes the *outcome* that the spinner comes to rest in sector i. The set

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$$

of all the possible outcomes is the *sample space*.

Spinning the spinner n times is an *n-stage experiment*, the outcomes being n-words from the alphabet Ω. Suppose that in n spins, ω_i is obtained F_i times. Then F_i is the *absolute frequency* and $f_i = F_i/n$ is the *relative frequency* of ω_i.

The symmetry of the circle implies that if the relative frequency of ω_i always approaches the same value, that value has to be the proportion of the perimeter assigned to ω_i, i.e. p_i. Indeed, Las Vegas records of various spinning devices confirm that for large n, $f_i \approx p_i$, although, as n increases, f_i's approach towards p_i is both slow and erratic. We call p_i the *probability* of the outcome ω_i

We should mention that it is a profound problem to come up with a logical argument that explains why an average over many repetitions of a random experiment should approach a limit; it could conceivably fluctuate aimlessly forever.

We call every subset A of Ω an *event*. Suppose a spin results in $\omega \in A$ then we say that *the event A has occurred*. For every event the frequency of occurrences of A is the sum of the frequencies of the outcomes in A. Hence the probability of the event A is

$$P(A) = \text{ sum of the probabilities of all outcomes in A.} \tag{A.2}$$

If $|\Omega| = n$ and $p_i = 1/n$ for all i (i.e. the outcomes are equally likely) then we have

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{\# of elements in A}}{\text{\# of elements in }\Omega} = \frac{\text{\# of favorable cases}}{\text{\# of possible cases}}. \tag{A.3}$$

You might think that a more natural description of the spinner would be to use the angles as the set of outcomes. However, that model actually requires a more complicated theory.

We want to study *discrete probability*, which is the theory of random experiments in which every possible outcome has a positive probability.

Usually we are given a random process in which we perform a series of random experiments. This process can be pictured as a *directed graph* (see Fig. A.2), with vertices called *states*. Conceptually, there is a spinner located inside each state which determines the *transition probabilities* for moving to the neighboring states.

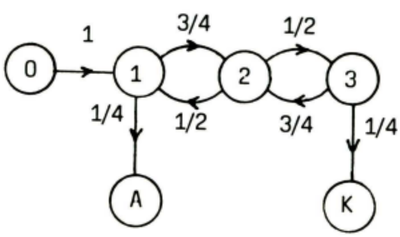


Figure A.2. A graph of transition probabilities

A "particle" starts in some state, say state 0, and moves from state to state, according to the transition probabilities. For example, the spinner in state 3 (see Fig. A.3), decides whether to move back to state 2 or to state K. The series of state transitions is called a *random walk*. For instance, one of the infinitely many possible paths is 0121232123K.

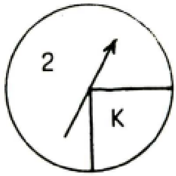


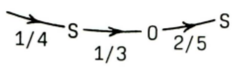
Figure A.3. Deciding between '2' and 'K'

With this terminology we can say:

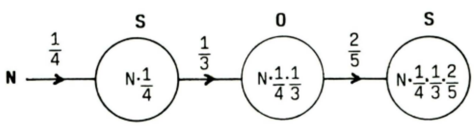
Discrete probability studies random walks on graphs.

To see how path probabilities are obtained from transition probabilities, consider the following example. From each box in Fig. A.4a we select a letter at random from left to right, and write down the letters in the order they were drawn. The probability of the outcome being "SOS" in depicted in Fig. A.4b.

(a) Selecting SOS



(b) Probabilities for selecting SOS



(c) Transition probabilities for SOS

Figure A.4. SOS probabilities

Suppose we repeat the experiment a large number, N, times. Fig. A.4c shows the computation of the approximate numbers of paths through the two intermediate states, to the final state. It tells us that the probability of "SOS" is

$$P(SOS) = \frac{1}{4}\frac{1}{3}\frac{2}{5} = \frac{1}{30}.$$

A general formulation of what we have just observed is:

P1. The probability of a path is equal to the product of the probabilities along the path.

This is the *first path rule*.

A second example illustrates another rule.

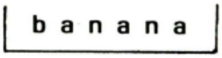


Figure A.5. A banana box

From the box in Fig. A.5 we draw two letters at random without replacement. What is the probability that the second letter is 'a'?

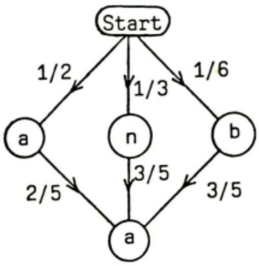


Figure A.6. Probability that the second letter is 'a'

Tracing the three paths in Fig. A.6 which produce 'a' in the second draw results in the equation

$$P(a \text{ at second draw}) = \frac{1}{2}\frac{2}{5} + \frac{1}{3}\frac{3}{5} + \frac{1}{6}\frac{3}{5} = \frac{1}{2}.$$

We can now formulate a *second path rule*:

P2. The probability of getting from S to F is the sum of the probabilities of all the paths leading from S to F.

This is a reformulation of Equ. (A.2) into the language of paths. We are thinking of all the possible outcomes as all the ordered pairs of letters that can be drawn, and A = {aa, ba, na}.

One objective of the theory of probability is to compute the probabilities of complicated events associated with random experiments from the known probabilities of the outcomes of a single experiment. For this purpose a simplified concept of probability suffices. The only thing we need to add is that the probabilities

p_i of the outcomes ω_i of a single experiment are non-negative real numbers satisfying Equ. (A.1), and that the probabilities of compound events are given by the boxed rules P1 and P2.

The fact that individual outcomes are unpredictable, while long-term averages approach limits, is a subtle and mysterious fact. Fortunately, in order to compute probabilities we only need the simple consequences which we stated above! Mathematicians were very pleased when it became clear that the theory could be derived from axioms as simple as these. However, we ought to add that this relative-frequency interpretation of probability lets us easily grasp many aspects of probability that are a lot less easy to formulate or prove axiomatically.

Next we look at a few examples that bring up additional concepts.

A.1 A Bold Gamble

You have 1 dollar but you desperately need 5 dollars. So you go into a casino where you can win as much as you bet with probability p and lose with probability $q = 1 - p$. You decide to use the *bold strategy*: at each stage you stake as much money as possible in order to minimize the distance to your goal. In this scenario, what is your probability of going broke?

Fig. A.7 shows the graph of the game. As a convenience, the state labels correspond to the amount of money you currently possess. Therefore, you start at state 1 and stake all your money. With probability p you move to state 2, and with probability q you are ruined. In state 2 you stake all your money again, but in state 4 you stake just one dollar (since you're only one dollar from your goal), and in state 3 two dollars.

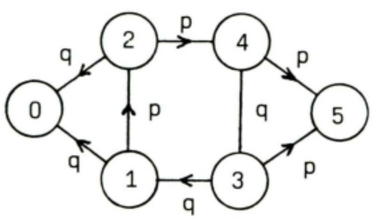


Figure A.7. Probability graph for the *bold gamble*

The probability B of going broke is the sum of the probabilities along all the paths leading to state 0 (i.e. the 'no money' state). You can lose directly, or after one, two, ... cycles around the graph. Thus $B = q + pq + p^2q^2(q + pq) + p^4q^4(q + pq) + ...$ This is a geometric series which sums to

$$B = \frac{q + pq}{1 - p^2q^2}.$$

A.2 Random Variables and Expectation

Often a random experiment produces a *random variable* X. This means that the possible outcome can come from a range of real numbers $x_1, x_2, ...$ whose probabilities of being chosen are $p_1, p_2, ...$ You might like to interpret the x_i as gains in a roulette game (Fig. A.8).

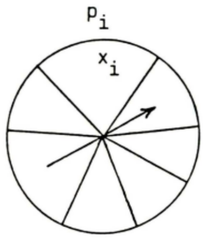


Figure A.8. Probability as roulette

Suppose you spin the spinner in Fig. A.8 N times, where N is large. We can compute the average gain as follows. We will win x_i approximately Np_i times, so the total gain will be approximately

$$Np_1x_1 + Np_2x_2 + Np_3x_3 + ...$$

If we play long enough, the average gain per game is going to be close to $1/N$-th of this:

$$E(X) = x_1p_1 + x_2p_2 + x_3p_3 + ... = \sum_{i\geq 1} x_ip_i. \qquad \text{(A.4)}$$

The quantity E(X) defined by Equ. (A.4) is called the *expectation* or *expected value* of X. Note that long term averages, and the complicated concept of playing "long enough", only appear when we have to explain the meaning of $E(X)$.

Suppose that besides X, another random variable X' with values $x'_1, x'_2, ...$ and corresponding probabilities $p'_1, p'_2, ...$ is defined. If a is any numerical constant then the definition of expectation as a long-term average tells us that the expectations of the random variables aX and $X + X'$ are:

$$E(aX) = aE(X) \qquad \text{(A.5)}$$

and

$$E(X + X') = E(X) + E(X'). \qquad \text{(A.6)}$$

Equ. (A.5) is an immediate consequence of Equ. (A.4). To derive Equ. (A.6) we need to set up a notation for an experiment in which one value is obtained for each of the two random variables X, X'. (The same approach still applies if the experiment involves two values from the same random variable.)

An outcome of the two-variable experiment is that we get a value x_i for X and and a value x'_j for X'. Let p_{ij} denote the probability of this outcome. In general these quantities can not be computed from $p_1, p_2, ...$ and $p'_1, p'_2, ...$

For example, suppose we have an box with 4 balls, two with '0' written on them and two with '1', and X and X' are obtained by drawing 2 balls without replacement. If we assume that each ball is equally likely to be the first and each of the remaining ones is equally likely to be the second, then we get

$$p_1 = p_2 = p'_1 = p'_2 = \frac{1}{2} \quad \text{and} \quad p_{11} = p_{22} = \frac{1}{6}, \quad p_{12} = p_{21} = \frac{1}{3}.$$

The relationship between the individual probabilities and the joint probabilities is the following. The event $X = x_i$ occurs if we have any one of the outcomes $X = x_i$ and $X' = x'_1$ or x'_2 or x'_3, ... Hence by Equ. (A.2):

$$p_{i1} + p_{i2} + ... = p_i \text{ and similarly} \tag{A.7}$$

$$p'_{1j} + p'_{2j} + ... = p'_j. \tag{A.8}$$

Applying Equ. (A.4) to $X + X'$ we get

$$E(X + X') = \sum_{i,j}(x_i + x'_j)p_{ij}.$$

If we separate out the terms involving the x_i and apply Equ. (A.7) we get $E(X)$, and the remaining terms combine to give $E(X')$.

A.3 More on Random Walks

Usually a graph includes a set A of *absorbing states* – states that once you enter, you cannot leave. The other states are in the set I of *interior* states. We are interested in two problems.

(1) If we start in some state i, what is the probability p_i of reaching some subset W of A?

(2) Starting in a state i, what is the expected time (# of transitions), E_i, to reach A?

For $i \in W$ we have $p_i = 1$, and for $i \in L = A \backslash W$ (the complement of W in A) we have $p_i = 0$. If we are in state $i \in I$, let p_{ik} denote the probability that the next move will be to state k. so, we have

$$p_i = \sum p_{ik} p_k, \tag{A.9}$$

where the sum is over all the neighbors k of i. Indeed, we move from i to k with probability p_{ik}, and from there we reach W with probability p_k. For all $i \in A$ we have $E_i = 0$, and for $i \in I$ we get

$$E_i = 1 + \sum p_{ik} E_k \quad \text{(The sum is over all neighbors k of i.)} \tag{A.10}$$

We have made one step to k with probability p_{ik}, and from there the expected time to absorption is E_k.

Expectations can be infinite. For instance, in a symmetric random walk on a line, starting at O, the expected number of steps to return to O is infinite. The system of equations $E_i = 1 + \frac{1}{2}E_{i-1} + \frac{1}{2}E_{i+1}$ has no solution, unless we consider $E_i = +\infty$ a solution. This result models how, as time goes on, the particle wanders over ever longer stretches of the line. We can show that it's almost certain that the particle will eventually return to the origin or, equivalently, to any vertex it has previously visited. However, the average time it takes to return to a particular place will increase without limit as more and more new vertices are visited.

A.4 Five Practice Problems

In this section, we gain some practice in representing a problem using a graph and in applying Equs. (A.9) and (A.10).

A.4.1 Problem 1. What is the expected time m to get from state 0 to 1 in Fig. A.9?

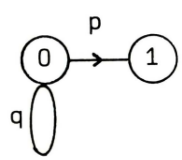


Figure A.9. Getting from '0' to '1'

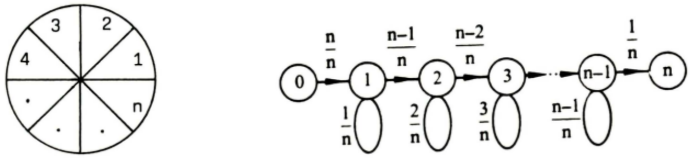
The expected time to get from state 1 to 1 is 0. So we get

$$m = 1 + qm, \quad q = 1 - p, \text{ and } m = \frac{1}{p}. \tag{A.11}$$

A.4.2 Problem 2. Spin the spinner in Fig. A.10a until all n outcomes have occurred. What is the expected number of spins?

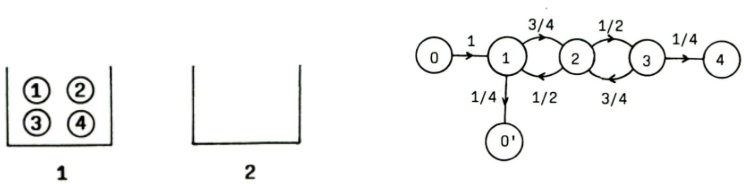
Consider Fig. A.10b: we are in state i if we have collected i outcomes. By Equ. (A.11) the expected time to get from i to $i+1$ is $n/(n-i)$. Due to Equs. (A.5) and (A.6), the expected time to get from state 0 to n is

$$E(n) = \sum_{i=1}^{n} \frac{n}{(n-i)} = n(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}).$$



(a) n outcomes (b) Transition probabilities

Figure A.10. Getting all n outcomes



(a) Two boxes (b) Transition probabilities for moving a ball

Figure A.11. Moving balls between boxes

A.4.3 Problem 3. Fig. A.11a shows two boxes labeled 1 and 2, with box 1 initially containing four balls. We repeatedly randomly select a ball from any box and move it to the other box.

We'll say that the system is in state i if the second box contains i balls, and will stop if either we reenter state 0 or state 4 is reached.

The random process can be represented by Fig. A.11b. Let p_0 be the probability of being absorbed in state 4 if we start in state 0. The instantiated versions of Equ. (A.9) are

$$p_0 = p_1, \quad p_1 = \frac{3}{4}p_2, \quad p_2 = \frac{1}{2}p_1 + \frac{1}{2}p_3, \quad p_3 = \frac{3}{4}p_2 + \frac{1}{4}$$

Solving these we get $p_0 = \frac{3}{8}$.

Let E_i be the expected time until we stop after starting in state i. Equ. (A.10) is instantiated as

$$E_0 = 1 + E_1; \quad E_1 = 1 + \frac{3}{4}E_2; \quad E_2 = 1 + \frac{1}{2}E_1 + \frac{1}{2}E_3; \quad E_3 = 1 + \frac{3}{4}E_2.$$

This system of equations sets $E_0 = 8$.

A.4.4 Problem 4. Start with $x = 3$ coins. While $x > 0$, toss the coins and eliminate those that come up heads. What is the expected number of tosses until $x = 0$?

Let a, b, c be the expected number of tosses for $x = 3, 2, 1$. From Fig. A.12, we get $a = 1 + a/8 + 3b/8 + 3c/8$, $b = 1 + b/4 + c/2$, $c = 1 + c/2$, with the results $c = 2, b = 8/3, a = 22/7$.

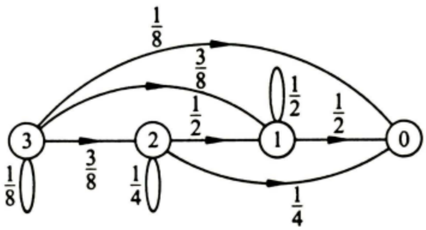


Figure A.12. Eliminating three coins

A.4.5 Problem 5. Let's reconsider the Three Pile Game from Ex. 2 of "Additional Exercises for Chapters 1-4". We have three piles of a, b, and c chips. A clock is ticking, and each second two different piles X, Y are chosen at random and a chip moved from X to Y. Let T be the time until one pile is empty. Devise an equation for the expected time $E(T) = f(a, b, c)$.

By simulation we can arrive at a guess

$$f(a,b,c) = \frac{3abc}{a+b+c}.$$

Theory lets us verify this equation. The six neighbors of the state (a, b, c) are $(a, b+1, c-1)$, $(a, b-1, c+1)$, $(a+1, b, c-1)$, $(a-1, b, c+1)$, $(a+1, b-1, c)$, and $(a-1, b+1, c)$. Equ. (A.10) becomes

$$f(a,b,c) = 1 + \frac{1}{6}\sum f(x,y,z) \text{ (sum over all neighbors of } (a,b,c)) \quad \text{(A.12)}$$

$$f(a,b,0) = f(a,0,c) = f(0,b,c) = 0 \quad \text{(boundary conditions).} \quad \text{(A.13)}$$

We can verify that the f function is a solution for these equations by substitution, but is it the only solution? Suppose that there's another solution $g(a, b, c)$. We consider $h(a, b, c) = f(a, b, c) - g(a, b, c)$ so that at an interior point (i.e. a point where a, b, c are all positive) this difference satisfies

$$h(a,b,c) = \frac{1}{6}\sum h(x,y,z).$$

In words: the value of h at an interior point is the average of the values at the six neighboring points.

Let $h_{\max}$ be the maximum of h. If $h = h_{\max}$ at an interior point then we must also have $h = h_{\max}$ at all the neighbors of that point. This means that we can move from point to point obtaining $h = h_{\max}$ at each one. However, since $h = 0$ on the boundary, then h must $= 0$ everywhere.

Alternatively, if $h = h_{\max}$ only occurs at the boundary then we still have $h_{\max} = 0$ but now we can only say that $h \leq 0$ everywhere else. However, we can apply the same argument to the minimum of h and get $h \geq 0$ everywhere, which implies that $h = 0$ everywhere. Hence, every solution g is equal to f.

A.5 More Information

An excellent modern introduction to probability and statistics is the text by Joseph K. Blitzstein and Jessica Hwang [**BH19**], developed from the Harvard course, 'Statistics 110' (`https://projects.iq.harvard.edu/stat110/`). Applications areas include genetics, medicine, computer science, and information theory. There's a free online (but not printable) version of the second edition at `http://probabilitybook.net`, and a series of 35 video lectures on Youtube at `https://www.youtube.com/watch?v=KbB0FjPg0mw&list=PL2SOU6wwxB0uwwH80KTQ6ht66KWxbzTIo`.

A superb classic presentation of discrete probability theory, for those who wish to immerse themselves in the subject, is William Feller's *Introduction to the Theory of Probability*, Vol. 1 [**Fel68**].