

NOTES ON ELEMENTARY PROBABILITY

KARL PETERSEN

1. PROBABILITY SPACES

Probability theory is an attempt to work mathematically with the relative uncertainties of random events. In order to get started, we do not attempt to estimate the probability of occurrence of any event but instead assume that somehow these have already been arrived at and so are given to us in advance. These data are assembled in the form of a *probability space* (X, \mathcal{B}, P) , which consists of

- (1) a set X , sometimes called the *sample space*, which is thought of as the set of all possible *states* of some system, or as the set of all possible *outcomes* of some experiment;
- (2) a family \mathcal{B} of subsets of X , which is thought of as the family of *observable events*; and
- (3) a function $P : \mathcal{B} \rightarrow [0, 1]$, which for each observable event $E \in \mathcal{B}$ gives the probability $P(E)$ of occurrence of that event.

While the set X of all possible outcomes is an arbitrary set, for several reasons, which we will not discuss at this moment, the set \mathcal{B} of observable events is not automatically assumed to consist of *all* subsets of X . (But if X is a *finite* set, then usually we do take \mathcal{B} to be the family of all subsets of X .)

We also assume that the family \mathcal{B} of observable events and the probability measure P satisfy a minimal list of properties which permit calculations of probabilities of combinations of events:

- (1) $P(X) = 1$
- (2) \mathcal{B} contains X and is closed under the set-theoretic operations of union, intersection, and complementation: if $E, F \in \mathcal{B}$, then $E \cup F \in \mathcal{B}$, $E \cap F \in \mathcal{B}$, and $E^c = X \setminus E \in \mathcal{B}$. (Recall that $E \cup F$ is the set of all elements of X that are either in E or in F , $E \cap F$ is the set of all elements of X that are in both E and F , and E^c is the set of all elements of X that are not in E .)

In fact, in order to permit even more calculations (but not *too* many) we suppose that also the union and intersection of *countably many* members of \mathcal{B} are still in \mathcal{B} .

- (3) If $E, F \in \mathcal{B}$ are disjoint, so that $E \cap F = \emptyset$, then $P(E \cup F) = P(E) + P(F)$. In fact, we assume that P is *countably additive*: if E_1, E_2, \dots are pairwise disjoint (so that $E_i \cap E_j = \emptyset$ if $i \neq j$), then

$$(1) \quad P(\cup_{i=1}^{\infty} E_i) = P(E_1 \cup E_2 \cup \dots) = P(E_1) + P(E_2) + \dots = \sum_{i=1}^{\infty} P(E_i).$$

Example 1.1. In some simple but still interesting and useful cases, X is a finite set such as $\{0, \dots, d-1\}$ and \mathcal{B} consists of all subsets of X . Then P is determined by specifying the value $p_i = P(\{i\})$ of each individual point i of X . For example, the single flip of a fair coin is modeled by letting $X = \{0, 1\}$, with 0 representing the outcome heads and 1 the outcome tails, and defining $P(0) = P(1) = 1/2$. Note that the probabilities of all subsets of X are then determined (in the case of the single coin flip, $P(X) = 1$ and $P(\emptyset) = 0$).

Exercise 1.1. Set up the natural probability space that describes the roll of a single fair die and find the probability that the outcome of any roll is a number greater than 2.

Exercise 1.2. When a pair of fair dice is rolled, what is the probability that the sum of the two numbers shown (on the upward faces) is even?

Exercise 1.3. In a certain lottery one gets to try to match (after paying an entry fee) a set of 6 different numbers that have been previously chosen from $\{1, \dots, 30\}$. What is the probability of winning?

Exercise 1.4. What is the probability that a number selected at random from $\{1, \dots, 100\}$ is divisible by both 3 and 7?

Exercise 1.5. A fair coin is flipped 10 times. What is the probability that heads comes up twice in a row?

Exercise 1.6. Ten fair coins are dropped on the floor. What is the probability that at least two of them show heads?

Exercise 1.7. A fair coin is flipped ten times. What is the probability that heads comes up at least twice?

Exercise 1.8. Show that if E and F are observable events in any probability space, then

$$(2) \quad P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

2. CONDITIONAL PROBABILITY

Let (X, \mathcal{B}, P) be a probability space and let $Y \in \mathcal{B}$ with $P(Y) > 0$. We can restrict our attention to Y , making it the set of possible states or outcomes for a probability space as follows:

- (1) The set of states is $Y \subset X$ with $P(Y) > 0$;
- (2) The family of observable events is defined to be
- (3)
$$\mathcal{B}_Y = \{E \cap Y : E \in \mathcal{B}\};$$
- (3) The probability measure P_Y is defined on \mathcal{B}_Y by
- (4)
$$P_Y(A) = \frac{P(A)}{P(Y)} \quad \text{for all } A \in \mathcal{B}_Y.$$

Forming the probability space (Y, \mathcal{B}_Y, P_Y) is called “conditioning on Y ”. It models the revision of probability assignments when the event Y is known to have occurred: we think of $P_Y(A)$ as the probability that A occurred, given that we already know that Y occurred.

Example 2.1. When a fair die is rolled, the probability of an even number coming up is $1/2$. What is the probability that an even number came up if we are told that the number showing is greater than 3? Then out of the three possible outcomes in $Y = \{4, 5, 6\}$, two are even, so the answer is $2/3$.

Definition 2.1. For any (observable) $Y \subset X$ with $P(Y) > 0$ and any (observable) $E \subset X$ we define the *conditional probability of E given Y* to be

$$(5) \quad P(E|Y) = \frac{P(E \cap Y)}{P(Y)} = P_Y(E).$$

Exercise 2.1. A fair coin is flipped three times. What is the probability of at least one head? Given that the first flip was tails, what is the probability of at least one head?

Exercise 2.2. From a group of two men and three women a set of three representatives is to be chosen. Each member is equally likely to be selected. Given that the set includes at least one member of each sex, what is the probability that there are more men than women in it?

Definition 2.2. The observable events A and B in a probability space (X, \mathcal{B}, P) are said to be *independent* in case

$$(6) \quad P(A \cap B) = P(A)P(B).$$

Notice that in case one of the events has positive probability, say $P(B) > 0$, then A and B are independent if and only if

$$(7) \quad P(A|B) = P(A);$$

that is, knowing that B has occurred does not change the probability that A has occurred.

Example 2.2. A fair coin is flipped twice. What is the probability that heads occurs on the second flip, given that it occurs on the first flip?

We model the two flips of the coin by bit strings of length two, writing 0 for heads and 1 for tails on each of the two flips. If Y is the set of outcomes which have heads on the first flip, and A is the set that has heads on the second flip, then

$$\begin{aligned} X &= \{00, 01, 10, 11\}, \\ Y &= \{00, 01\}, \quad \text{and} \\ A &= \{10, 00\}, \end{aligned}$$

so that $A \cap Y = \{00\}$ includes exactly one of the two elements of Y . Since each of the four outcomes in X is equally likely,

$$P(A|Y) = \frac{P(A \cap Y)}{P(Y)} = \frac{|A \cap Y|}{|Y|} = \frac{1}{2} = P(A).$$

Thus we see that A and Y are independent.

This example indicates that the definition of independence in probability theory reflects our intuitive notion of events whose occurrences do not influence one another. If repeated flips of a fair coin are modeled by a probability space consisting of bit strings of length n , all being equally likely, then an event whose occurrence is determined by a certain range of coordinates is independent of any other event that is determined by a disjoint range of coordinates.

Example 2.3. A fair coin is flipped four times. Let A be the event that we obtain a head on the second flip and B be the event that among the first, third, and fourth flips we obtain at least two heads. Then A and B are independent.

Exercise 2.3. Show that the events A and B described in the preceding example will be independent whether or not the coin being flipped is fair.

Exercise 2.4. Show that events A and B described in the preceding example will be independent even if the probability of heads could be different on each flip.

Exercise 2.5. When a pair of fair dice is rolled, is the probability of the sum of the numbers shown being even independent of it being greater than six?

3. BAYES' THEOREM

Looking at the definition of conditional probability kind of backwards leads very easily to a simple formula that is highly useful in practice and has profound implications for the foundations of probability theory (frequentists, subjectivists, etc.). We use the notation from [1], in which C is an event, thought of as a *cause*, such as the presence of a disease, and I is another event, thought of as the existence of certain *information*. The formula can be interpreted as telling us how to revise our original estimate $P(C)$ that the cause C is present if we are given the information I .

Theorem 3.1 (Bayes' Theorem). *Let (X, \mathcal{B}, P) be a probability space and let $C, I \in \mathcal{B}$ with $P(I) > 0$. Then*

$$(8) \quad P(C|I) = P(C) \frac{P(I|C)}{P(I)}.$$

Proof. We just use the definitions of the conditional probabilities:

$$(9) \quad P(C|I) = \frac{P(C \cap I)}{P(I)}, \quad P(I|C) = \frac{P(I \cap C)}{P(C)}$$

and the fact that $C \cap I = I \cap C$. □

Example 3.1. We discuss the example in [1, p. 77] in this notation. C is the event that a patient has cancer, and $P(C)$ is taken to be .01, the incidence of cancer in the general population for this example taken to be 1 in 100. I is the event that the patient tests positive on a certain test for this disease. The test is said to be 99% accurate, which we take to mean that the probability of error is less than .01, in the sense that $P(I|C^c) < .01$ and $P(I^c|C) < .01$.

Then $P(I|C) \approx 1$, and

$$(10) \quad P(I) = P(I|C)P(C) + P(I|C^c)P(C^c) \approx .01 + (.01)(.99) \approx .02.$$

Applying Bayes' Theorem,

$$(11) \quad P(C|I) = P(C) \frac{P(I|C)}{P(I)} \approx (.01) \frac{1}{.02} \approx .5.$$

The surprising conclusion is that even with such an apparently accurate test, if someone tests positive for this cancer there is only a 50% chance that he actually has the disease.

Often Bayes' Theorem is stated in a form in which there are several possible causes C_1, C_2, \dots which might lead to a result I with $P(I) > 0$. If we assume that the observable events C_1, C_2, \dots form a *partition* of the probability space X , so that they are pairwise disjoint and their union is all of X , then

$$(12) \quad P(I) = P(I|C_1)P(C_1) + P(I|C_2)P(C_2) + \dots,$$

and Equation (8) says that for each i ,

$$(13) \quad P(C_i|I) = P(C_i) \frac{P(I|C_i)}{P(I|C_1)P(C_1) + P(I|C_2)P(C_2) + \dots}.$$

This formula applies for any finite number of observable events C_i as well as for a *countably infinite* number of them.

Exercise 3.1. Suppose we want to use a set of medical tests to look for the presence of one of two diseases. Denote by S the event that the test gives a positive result and by D_i the event that a patient has disease $i = 1, 2$. Suppose we know the incidences of the two diseases in the population:

$$(14) \quad P(D_1) = .07, \quad P(D_2) = .05, \quad P(D_1 \cap D_2) = .01.$$

From studies of many patients over the years it has also been learned that

$$(15) \quad \begin{aligned} P(S|D_1) &= .9, & P(S|D_2) &= .8, \\ P(S|(D_1 \cup D_2)^c) &= .05, & P(S|D_1 \cap D_2) &= .99. \end{aligned}$$

(a) Form a partition of the underlying probability space X that will help to analyze this situation.

(b) Find the probability that a patient has disease 1 if the battery of tests turns up positive.

(c) Find the probability that a patient has disease 1 but not disease 2 if the battery of tests turns up positive.

4. BERNOULLI TRIALS

In Section 2 we came across independent repeated trials of an experiment, such as flipping a coin or rolling a die. Such a sequence is conveniently represented by a probability space whose elements are strings on a finite alphabet. Equivalently, if a single run of the experiment is modeled by a probability space (D, \mathcal{B}, P) , then n independent repetitions of the experiment are modeled by the Cartesian product of D with itself n times, with the probability measure formed by a product of P with itself n times.

We now state this more precisely. Let (D, \mathcal{B}, P) be a probability space with $D = \{0, \dots, d-1\}$, \mathcal{B} = the family of all subsets of D , $P(i) = p_i > 0$ for $i = 0, \dots, d-1$.

Denote by $D^{(n)}$ the Cartesian product of D with itself n times. Thus $D^{(n)}$ consists of all ordered n -tuples (x_1, \dots, x_n) with each $x_i \in D, i = 1, \dots, n$. If we omit the commas and parentheses, we can think of each element of $D^{(n)}$ as a *string of length n on the alphabet D* .

Example 4.1. If $D = \{0, 1\}$ and $n = 3$, then

$$D^{(3)} = \{000, 001, 010, 011, 100, 101, 110, 111\},$$

the set of all bit strings of length 3.

We now define the set of observables in $D^{(n)}$ to be $\mathcal{B}^{(n)}$ = the family of all subsets of $D^{(n)}$. The probability measure $P^{(n)}$ on $D^{(n)}$ is determined by

$$(16) \quad P^{(n)}(x_1 x_2 \dots x_n) = P(x_1)P(x_2) \cdots P(x_n)$$

for each $x_1 x_2 \dots x_n \in D^{(n)}$.

This definition of $P^{(n)}$ in terms of products of probabilities seen in the different coordinates (or entries) of a string guarantees the independence of two events that are determined by disjoint ranges of coordinates. Note that this holds true even if the strings of length n are not all equally likely.

Exercise 4.1. A coin whose probability of heads is p , with $0 < p < 1/2$, is flipped three times. Write out the probabilities of all the possible outcomes. If A is the event that the second flip produces heads, and B is the event that either the first or third flip produces tails, find $P^{(3)}(A \cap B)$ and $P^{(3)}(A)P^{(3)}(B)$.

Let $D = \{0, 1\}$ and $P(0) = p \in (0, 1), P(1) = 1 - p$. Construct as above the probability space $(D^{(n)}, \mathcal{B}^{(n)}, P^{(n)})$ representing n independent repetitions of the experiment (D, \mathcal{B}, P) . The *binomial distribution* gives the probability for each $k = 0, 1, \dots, n$ of the set of strings of length n that contain exactly k 0's. recall that $C(n, k)$ denotes the binomial coefficient $n!/(k!(n-k)!)$, the number of k -element subsets of a set with n elements.

Proposition 4.1. *Let $(D^{(n)}, \mathcal{B}^{(n)}, P^{(n)})$ be as described above. Then for each $k = 0, 1, \dots, n$,*

$$(17) \quad P^{(n)}\{x_1 \dots x_n \in D^{(n)} : x_i = 0 \text{ for } k \text{ choices of } i = 1, \dots, n\} = C(n, k)p^k(1-p)^{n-k}.$$

Proof. For each subset S of $\{1, \dots, n\}$, let

$$E(S) = \{x \in D^{(n)} : x_i = 0 \text{ if and only if } i \in S\}.$$

Note that if S_1 and S_2 are different subsets of $\{1, \dots, n\}$, then $E(S_1)$ and $E(S_2)$ are disjoint.

Fix $k = 0, 1, \dots, n$. There are $C(n, k)$ subsets of $\{1, \dots, n\}$ which have exactly k elements, and for each such subset S we have

$$P^{(n)}(E(S)) = p^k(1-p)^{n-k}.$$

Adding up the probabilities of these disjoint sets gives the result. \square

Exercise 4.2. For the situation in Exercise 4.1 and each $k = 0, 1, 2, 3$, list the elements of $A_k =$ the event that exactly k heads occur. Also calculate the probability of each A_k .

Representing repetitions of an experiment with finitely many possible outcomes by strings on a finite alphabet draws an obvious connection with the modeling of information transfer or acquisition. A single experiment can be viewed as reading a single symbol, which is thought of as the outcome of the experiment. We can imagine strings (or experimental runs) of arbitrary lengths, and in fact even of infinite length. For example, we can consider the space of one-sided infinite bit strings

$$(18) \quad \Omega^+ = \{x_0x_1x_2 \cdots : \text{each } x_i = 0 \text{ or } 1\},$$

as well as the space of two-sided infinite bit strings

$$(19) \quad \Omega = \{\dots x_{-1}x_0x_1 \cdots : \text{each } x_i = 0 \text{ or } 1\}.$$

Given p with $0 < p < 1$, we can again define a probability measure for many events in either of these spaces: for example,

$$(20) \quad P_p^{(\infty)}\{x : x_2 = 0, x_6 = 1, x_7 = 1\} = p(1-p)(1-p).$$

A set such as the one above, determined by specifying the entries in a finite number of places in a string, is called a *cylinder set*. Let us define the probability of each cylinder set in accord with the idea that 0's and 1's are coming independently, with probabilities p and $1-p$, respectively. Thus, if $0 \leq i_1 < i_2 < \dots < i_r$, each $a_1, \dots, a_r = 0$ or 1 , and s of the a_j 's are 0, let

$$(21) \quad P_p^{(\infty)}\{x \in \Omega^+ : x_{i_1} = a_1, \dots, x_{i_r} = a_r\} = p^s(1-p)^{r-s}.$$

It takes some effort (which we will not expend at this moment) to see that this definition does not lead to any contradictions, and that there is a unique extension of $P_p^{(\infty)}$ so as to be defined on a family $\mathcal{B}^{(\infty)}$ which contains all the cylinder sets and is closed under complementation, countable unions, and countable intersections.

Definition 4.1. If D is an arbitrary finite set, we denote by $\Omega^+(D)$ the set of one-sided infinite strings $x_0x_1x_2\dots$ with entries from the alphabet D , and we denote by $\Omega(D)$ the set of two-sided infinite strings with entries from D . We abbreviate $\Omega^+ = \Omega^+(\{0, 1\})$ and $\Omega = \Omega(\{0, 1\})$.

With each of these sequence spaces we deal always with a fixed family \mathcal{B} of observable events which contains the cylinder sets and is closed under countable unions, countable intersections, and complementation.

The spaces $\Omega^+(D)$ and $\Omega(D)$ are useful models of *information sources*, especially when combined with a family of observables \mathcal{B} which contains all cylinder sets and with a probability measure P defined on \mathcal{B} . (We are dropping the extra superscripts on \mathcal{B} and P in order to simplify the notation.) Given a string $a = a_0\dots a_{r-1}$ on the symbols of the alphabet D and a time $n \geq 0$, the probability that the source emits the string at time n is given by the probability of the cylinder set $\{x : x_n = a_0, x_{n+1} = a_1, \dots, x_{n+r-1} = a_{r-1}\}$.

Requiring that countable unions and intersections of observable events be observable allows us to consider quite interesting and complicated events, including various combinations of infinite sequences of events.

Example 4.2. In the space Ω^+ constructed above, with the probability measure $P_p^{(\infty)}$, let us see that the set of (one-sided) infinite strings which contain infinitely many 0's has probability 1. For this purpose

we assume (as can be proved rigorously) that the probability space $(\Omega^+, \mathcal{B}^{(\infty)}, P_p^{(\infty)})$ does indeed satisfy the properties set out axiomatically at the beginning of these notes. Let

$$A = \{x \in \Omega^+ : x_i = 0 \text{ for infinitely many } i\}.$$

We aim to show that $P^{(\infty)}(A^c) = 0$ ($A^c = \Omega^+ \setminus A$ = the complement of A), and hence $P^{(\infty)}(A) = 1$.

For each $n = 0, 1, 2, \dots$ let

$$B_n = \{x \in \Omega^+ : x_n = 0 \text{ but } x_i = 1 \text{ for all } i > n\},$$

and let B_{-1} consist of the single string 1111.... Then the sets B_n are pairwise disjoint and their union is A^c .

By countable additivity,

$$P_p^{(\infty)}\left(\bigcup_{n=-1}^{\infty} B_n\right) = \sum_{n=-1}^{\infty} P_p^{(\infty)}(B_n),$$

so it is enough to show that

$$P_p^{(\infty)}(B_n) = 0 \quad \text{for all } n.$$

Fix any $n = -1, 0, 1, 2, \dots$. For each $r = 1, 2, \dots$,

$$B_n \subset Z_{n+1, n+r} = \{x \in \Omega^+ : x_{n+1} = x_{n+2} = \dots = x_{n+r} = 1\},$$

and

$$P_p^{(\infty)}(Z_{n+1, n+r}) = (1-p)^r.$$

Since $0 < 1-p < 1$, we have $(1-p)^r \rightarrow 0$ as $r \rightarrow \infty$, so $P_p^{(\infty)}(B_n) = 0$ for each n .

If A is an observable event in any probability space which has probability 1, then we say that A occurs *almost surely*, or *with probability 1*. If some property holds for all points $x \in D$ in a set of probability 1, then we say that the property holds *almost everywhere*.

Exercise 4.3. In the probability space $(\Omega^+, \mathcal{B}^{(\infty)}, P_p^{(\infty)})$ constructed above, find the probability of the set of infinite strings of 0's and 1's which never have two 1's in a row. (*Hint:* For each $n = 0, 1, 2, \dots$ consider $B_n = \{x \in \Omega^+ : x_{2n}x_{2n+1} \neq 11\}$.)

5. MARKOV CHAINS

Symbols in strings or outcomes of repeated experiments are not always completely independent of one another—frequently there are relations, interactions, or dependencies among the entries in various coordinates. In English text, the probabilities of letters depend heavily on letters near them: h is much more likely to follow t than to follow f . Some phenomena can show very long-range order, even infinite memory. Markov chains model processes with only short-range memory, in which the probability of what symbol comes next depends only on a fixed number of the immediately preceding symbols. In the simplest case, 1-step Markov chains, the probability of what comes next depends only on the immediately preceding symbol. The outcome of any repetition of the experiment depends only on the outcome of the immediately preceding one and not on any before that.

The precise definition of a Markov chain on a finite state space, or alphabet, $D = \{0, 1, \dots, d-1\}$ is as follows. The sample space is the set Σ^+ of all one-sided (could be also two-sided) infinite sequences $x = x_0x_1\dots$ with entries from the alphabet D . The family of observable events again contains all the cylinder sets. The probability measure M is determined by two pieces of data:

- (1) a *probability vector* $p = (p_0, \dots, p_{d-1})$, with each $p_i \geq 0$ and $p_0 + \dots + p_{d-1} = 1$, giving the *initial distribution* for the chain;
- (2) a matrix $P = (P_{ij})$ giving the *transition probabilities* between each pair of states $i, j \in D$. It is assumed that each $P_{ij} \geq 0$ and that for each i we have $P_{i1} + P_{i2} + \dots + P_{i,d-1} = 1$. Such a P is called a *stochastic matrix*.

Now the probability of each basic cylinder set determined by fixing the first n entries at values $a_0, \dots, a_{n-1} \in D$ is defined to be

$$(22) \quad M\{x \in \Sigma^+ : x_0 = a_0, \dots, x_{n-1} = a_{n-1}\} = p_{a_0} P_{a_0 a_1} P_{a_1 a_2} \dots P_{a_{n-2}, a_{n-1}}.$$

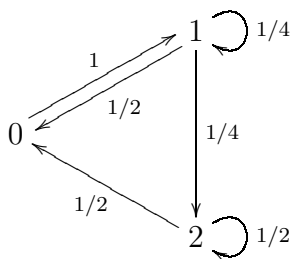
The idea here is simple. The initial symbol of a string, at coordinate 0, is selected with probability determined by the initial distribution p : symbol i has probability p_i of appearing, for each $i = 0, 1, \dots, d-1$. Then given that symbol, a_0 , the probability of transitioning to any other symbol is determined by the entries in the matrix P , specifically the entries in row a_0 : the probability that a_1 comes next, given that we just saw a_0 is $P_{a_0 a_1}$. And so on. The condition that the matrix P have

row sums 1 tells us that we are sure to be able to add *some* symbol each time.

The 1-step memory property can be expressed as follows. For any choice of symbols a_0, \dots, a_n ,

$$M\{x \in \Sigma^+ : x_n = a_n | x_0 = a_0, \dots, x_{n-1} = a_{n-1}\} = \\ M\{x \in \Sigma^+ : x_n = a_n | x_{n-1} = a_{n-1}\}.$$

Finite-state Markov chains are conveniently visualized in terms of random paths on directed graphs.



Here the states are 0, 1, 2 and the transition probabilities between states are the labels on the arrows. Thus the stochastic transition matrix is

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \end{pmatrix}.$$

If we specified an initial distribution $p = (1/6, 1/2, 1/3)$ listing the initial probabilities of the states 0, 1, 2, respectively, then the probabilities of strings starting at the initial coordinate would be calculated as in this example:

$$M\{x \in \Sigma^+ : x_0 = 1, x_2 = 1, x_3 = 0\} = p_1 P_{11} P_{10} = \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{16}.$$

Exercise 5.1. For the example above, with p and P as given, find the probabilities of all the positive-probability strings of length 3.

Recall that the vector $p = (p_0, \dots, p_{d-1})$ gives the initial distribution: the probability that at time 0 the system is in state $j \in \{0, \dots, d-1\}$ is p_j . So what is the probability that the system is in state j at time 1?

Well, the event that the system is in state j at time 1, namely $\{x \in \Sigma^+ : x_1 = j\}$, is the union of d disjoint sets defined by the different

possible values of x_0 :

$$(23) \quad \{x \in \Sigma^+ : x_1 = j\} = \bigcup_{i=0}^{d-1} \{x \in \Sigma^+ : x_0 = i, x_1 = j\}.$$

Since the i 'th one of these sets has probability $p_i P_{ij}$, we have

$$(24) \quad M\{x \in \Sigma^+ : x_1 = j\} = \sum_{i=0}^{d-1} p_i P_{ij}.$$

So we have determined the distribution $p^{(1)}$ of the chain at time 1. The equations

$$(25) \quad p_j^{(1)} = \sum_{i=0}^{d-1} p_i P_{ij} \quad \text{for } j = 0, \dots, d-1$$

are abbreviated, using multiplication of vectors by matrices, by

$$(26) \quad p^{(1)} = pP.$$

Similarly, the distribution at time 2 is given by

$$(27) \quad p^{(2)} = p^{(1)}P = pP^2,$$

where P^2 is the square of the matrix P according to matrix multiplication. And so on: the probability that at any time $n = 0, 1, 2, \dots$ the chain is in state $j = 0, \dots, d-1$ is $(pP^n)_j$, namely, the j 'th entry of the vector obtained by multiplying the initial distribution vector p on the right n times by the stochastic transition matrix P .

Here's a *quick definition of matrix multiplication*. Suppose that A is a matrix with m rows and n columns ($m, n \geq 1$; if either equals 1, A is a (row or column) vector). Suppose that B is a matrix with n rows and p columns. Then AB is defined as a matrix with m rows and p columns. The entry in the i 'th row and j 'th column of the product AB is formed by using the i 'th row of A and the j 'th column of B : take the sum of the products of the entries in the i 'th row of A (there are n of them) with the entries in the j 'th column of B (there are also n of these)—this is the “dot product” or “scalar product” of the i 'th row of A with the j 'th column of B :

$$(28) \quad (AB)_{ij} = \sum_{k=1}^n A_{ik} B_{kj}, \quad \text{for } i = 1, \dots, m; j = 1, \dots, p.$$

Note that here we have numbered entries starting with 1 rather than with 0. (This is how Matlab usually does it).

Markov chains have many applications in physics, biology, psychology (learning theory), and even sociology. Here is a nonrealistic indication of possible applications.

Exercise 5.2. Suppose that a certain study divides women into three groups according to their level of education: completed college, completed high school but not college, or did not complete high school. Suppose that data are accumulated showing that the daughter of a college-educated mother has a probability .7 of also completing college, probability .2 of only making it through high school, and probability .1 of not finishing high school; the daughter of a mother who only finished high school has probabilities .5, .3, and .2, respectively, of finishing college, high school only, or neither; and the daughter of a mother who did not finish high school has corresponding probabilities .3, .4, and .3.

(a) We start with a population in which 30% of women finished college, 50% finished high school but not college, and 20% did not finish high school. What is the probability that a granddaughter of one of these women who never finished high school will make it through college?

(b) Suppose that the initial distribution among the different groups is (.5857, .2571, .1571). What will be the distribution in the next generation? The one after that? The one after that?

Remark 5.1. Under some not too stringent hypotheses, the powers P^k of the stochastic transition matrix P of a Markov chain will converge to a matrix Q all of whose rows are equal to the same vector q , which then satisfies $qQ = q$ and is called the *stable distribution* for the Markov chain. You can try this out easily on Matlab by starting with various stochastic matrices P and squaring repeatedly.

6. SPACE MEAN AND TIME MEAN

Definition 6.1. A *random variable* on a probability space (X, \mathcal{B}, P) is a function $f : X \rightarrow \mathbb{R}$ such that for each interval (a, b) of real numbers, the event $\{x \in X : f(x) \in (a, b)\}$ is an observable event. More briefly,

$$(29) \quad f^{-1}(a, b) \in \mathcal{B} \quad \text{for all } a, b \in \mathbb{R}.$$

This definition seeks to capture the idea of making measurements on a random system, without getting tangled in talk about numbers fluctuating in unpredictable ways.

Example 6.1. In an experiment of rolling two dice, a natural sample space is $X = \{(i, j) : i, j = 1, \dots, 6\}$. We take \mathcal{B} = the family of all subsets of X and assume that all 36 outcomes are equally likely. One important random variable on this probability space is the sum of the numbers rolled:

$$s(i, j) = i + j \quad \text{for all } (i, j) \in X.$$

Example 6.2. If X is the set of bit strings of length 7, \mathcal{B} = all subsets of X , and all strings are equally likely, we could consider the random variable

$$s(x) = x_0 + \dots + x_6 = \text{number of 1's in } x.$$

In the following definitions let (X, \mathcal{B}, P) be a probability space.

Definition 6.2. A *partition* of X is a family $\{A_1, \dots, A_n\}$ of observable subsets of X (each $A_i \in \mathcal{B}$) which are pairwise disjoint and whose union is X . The sets A_i are called the *cells* of the partition.

Definition 6.3. A *simple random variable* on X is a random variable $f : X \rightarrow \mathbb{R}$ for which there is a partition $\{A_1, \dots, A_n\}$ of X such that f is constant on each cell A_i of the partition: there are $c_1, \dots, c_n \in \mathbb{R}$ such that $f(x) = c_i$ for all $x \in A_i, i = 1, \dots, n$.

Definition 6.4. Let f be a simple random variable as in Definition 6.3. We define the *space mean*, or *expected value*, or *expectation* of f to be

$$(30) \quad \mathbb{E}(f) = \sum_{i=1}^n c_i P(A_i).$$

Example 6.3. Let the probability space and random variable f be as in Example 6.1—the sum of the numbers showing. To compute the expected value of $f = s$, we partition the set of outcomes according to the value of the sum: let $A_j = s^{-1}(j), j = 2, \dots, 12$. Then we figure out the probability of each cell of the partition. Since all outcomes are assumed to be equally likely, the probability that $s(x) = i$ is the number of outcomes x that produce sum i , times the probability ($1/36$) of each outcome. Now the numbers of ways to roll 2, 3, \dots , 12, respectively, are seen by inspection to be 1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1. Multiply each value (2 through 12) of the random variable s by the probability that it takes that value ($1/36, 2/36, \dots, 1/36$) and add these up to get $\mathbb{E}(s) = 7$.

Thus 7 is the expected sum on a roll of a pair of dice. This is the *mean* or *average* sum. The expected value is not always the same as

the most probable value (if there is one)—called the *mode*—as the next example shows.

Exercise 6.1. Find the expected value of the random variable in Example 6.2.

Exercise 6.2. Suppose that the bit strings of length 7 in Example 6.2 are no longer equally likely but instead are given by the probability measure $P^{(7)}$ on $\{0, 1\}^{(7)}$ determined by $P(0) = 1/3$, $P(1) = 2/3$. Now what is the expected value of the number of 1's in a string chosen at random?

The expectation of a random variable f is its average value over the probability space X , taking into account that f may take values in some intervals with greater probability than in others. If the probability space modeled a game in which an entrant received a payoff of $f(x)$ dollars in case the random outcome were $x \in X$, the expectation $\mathbb{E}(f)$ would be considered a fair price to pay in order to play the game. (Gambling establishments charge a bit more than this, so that they will probably make a profit.)

We consider now a situation in which we make not just a single measurement f on a probability space (X, \mathcal{B}, P) but a *sequence* of measurements f_1, f_2, f_3, \dots . A sequence of random variables is called a *stochastic process*. If the system is in state $x \in X$, then we obtain a sequence of numbers $f_1(x), f_2(x), f_3(x), \dots$, and we think of $f_i(x)$ as the result of the observation that we make on the system at time $i = 1, 2, 3, \dots$.

It is natural to form the averages of these measurements:

$$(31) \quad A_n\{f_i\}(x) = \frac{1}{n} \sum_{k=1}^n f_k(x)$$

is the average of the first n measurements. If we have an *infinite* sequence f_1, f_2, f_3, \dots of measurements, we can try to see whether these averages settle down around a limiting value

$$(32) \quad A_\infty\{f_i\}(x) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f_k(x).$$

Such a limiting value may or may not exist—quite possibly the sequence of measurements will be wild and the averages will not converge to any limit.

We may look at the sequence of measurements and time averages in a different way: rather than imagining that we make a sequence of measurements on the system, we may imagine that we make the *same* measurement f on the system each time, but the system changes with time. This is the viewpoint of dynamical systems theory; in a sense the two viewpoints are equivalent.

Example 6.4. Consider the system of Bernoulli trials $(\Omega^+, \mathcal{B}^{(\infty)}, P_p^{(\infty)})$ described above: the space consists of one-sided infinite sequences of 0's and 1's, the bits arriving independently with $P(0) = p$ and $P(1) = 1-p$. We can “read” a sequence in two ways.

(1) For each $i = 0, 1, 2, \dots$, let $f_i(x) = x_i$. We make a different measurement at each instant, always reading off the bit that is one place more to the right than the previously viewed one.

(2) Define the *shift transformation* $\sigma : \Omega^+ \rightarrow \Omega^+$ by $\sigma(x_0x_1x_2\dots) = x_1x_2\dots$. This transformation lops off the first entry in each infinite bit string and shifts the remaining ones one place to the left. For each $i = 1, 2, \dots$, σ^i denotes the composition of σ with itself i times; thus σ^2 lops off the first two places while shifting the sequence two places to the left. On the set Ω of two-sided infinite sequences we can shift in both directions, so we can consider σ^i for $i \in \mathbb{Z}$.

Now let $f(x) = x_0$ for each $x \in \Omega^+$. Then the previous $f_i(x) = f(\sigma^i x)$ for all $i = 0, 1, 2, \dots$. In this realization, we just sit in one place, always observing the first entry in the bit string x as the string streams by toward the left.

This seems to be maybe a more relaxed way to make measurements. Besides that, the dynamical viewpoint has many other advantages. For example, many properties of the stochastic processes $\{f(\sigma^i x)\}$, can be deduced from study of the action of σ on $(\Omega^+, \mathcal{B}^{(\infty)}, P_p^{(\infty)})$ alone, independently of the particular choice of f .

Exercise 6.3. In the example $(\Omega^+, \mathcal{B}^{(\infty)}, P_p^{(\infty)})$ just discussed, with $f(x) = x_0$ as above, do you think that the time average

$$A_\infty f(x) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(\sigma^k x)$$

will exist? (For all x ? For most x ?) If it were to exist usually, what should it be?

Exercise 6.4. Same as the preceding exercise, but with f replaced by

$$f(x) = \begin{cases} 1 & \text{if } x_0x_1 = 01 \\ 0 & \text{otherwise.} \end{cases}$$

7. STATIONARY AND ERGODIC INFORMATION SOURCES

We have already defined an information source. It consists of the set of one- or two-sided infinite strings $\Omega^+(D)$ or $\Omega(D)$ with entries from a finite alphabet D ; a family \mathcal{B} of subsets of the set of strings which contains all the cylinder sets and is closed under complementation and countable unions and intersections; and a probability measure P defined for all sets in \mathcal{B} . (For simplicity we continue to delete superscripts on \mathcal{B} and P .) We also have the shift transformation, defined on each of $\Omega^+(D)$ and $\Omega(D)$ by $(\sigma x)_i = x_{i+1}$ for all indices i . If $f(x) = x_0$, then observing $f(\sigma^k x)$ for $k = 0, 1, 2, \dots$ “reads” the sequence $x = x_0x_1x_2\dots$ as σ makes time go by.

Definition 7.1. An information source as above is called *stationary* if the probability measure P is shift-invariant: given any word $a = a_0a_1\dots a_{r-1}$ and any two indices n and m in the allowable range of indices (\mathbb{Z} for $\Omega(D)$, $\{0, 1, 2, \dots\}$ for $\Omega^+(D)$),

$$(33) \quad \begin{aligned} P\{x : x_n = a_0, x_{n+1} = a_1, \dots, x_{n+r-1} = a_{r-1}\} = \\ P\{x : x_m = a_0, x_{m+1} = a_1, \dots, x_{m+r-1} = a_{r-1}\}. \end{aligned}$$

The idea is that a stationary source emits its symbols, and in fact consecutive strings of symbols, according to a probability measure that does not change with time. The probability of seeing a string such as 001 is the same at time 3 as it is at time 3003. Such a source can be thought of as being in an equilibrium state—whatever mechanisms are driving it (which are probably random in some way) are not having their basic principles changing with time.

Example 7.1. The Bernoulli sources discussed above are stationary. This is clear from the definition of the probability of the cylinder set determined by any word as the product of the probabilities of the individual symbols in the word.

Example 7.2. Consider a Markov source as above determined by an initial distribution p and a stochastic transition matrix P . If p is in fact a stable distribution for P (see Remark 5.1),

$$pP = p,$$

then the Markov process, considered as an information source, is stationary.

Definition 7.2. A stationary information source as above is called *ergodic* if for every simple random variable f on the set of sequences, the time mean of f almost surely equals the space mean of f . More precisely, the set of sequences x for which

$$(34) \quad A_\infty f(x) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(\sigma^k x) = \mathbb{E}(f)$$

(in the sense that the limit exists and equals $\mathbb{E}(f)$) has probability 1.

In fact, it can be shown that in order to check whether or not a source is ergodic, it is enough to check the definition for random variables f which are the *characteristic functions* of cylinder sets. Given a word $a = a_0 a_1 a_2 \dots a_{r-1}$, define

$$(35) \quad f_a(x) = \begin{cases} 1 & \text{if } x_0 x_1 \dots x_{r-1} = a \\ 0 & \text{otherwise.} \end{cases}$$

Ergodicity is then seen to be equivalent to requiring that *in almost every sequence, every word appears with limiting frequency equal to the probability of any cylinder set defined by that word*. Here “almost every” sequence means a set of sequences which has probability one.

Example 7.3. The Bernoulli systems defined above are all ergodic. This is a strong version of Jakob Bernoulli’s Law of Large Numbers (1713).

What kinds of sources are *not* ergodic, you ask? It’s easiest to give examples if one knows that ergodicity is equivalent to a kind of indecomposability of the probability space of sequences.

Example 7.4. Let us consider an information source which puts out one-sided sequences on the alphabet $D = \{0, 1\}$. Let us suppose that the probability measure P governing the outputs is such that with probability $1/2$ we get a constant string of 0’s, otherwise we get a string of 0’s and 1’s coming independently with equal probabilities. If we consider the simple random variable f_0 , which gives a value of 1 if $x_0 = 0$ and otherwise gives a value 0, we see that on a set of probability $1/2$ the time mean of f_0 is 1, while on another set of probability $1/2$ it is $1/2$ (assuming the result stated in Example 7.3). Thus, no matter the value of $\mathbb{E}(f_0)$, we cannot possibly have $A_\infty f_0 = \mathbb{E}(f_0)$ almost surely.

Exercise 7.1. Calculate the space mean of the random variable f_0 in the preceding example.

Exercise 7.2. Calculate the space mean and time mean of the random variable f_1 in the preceding example (see Formula (35)).

REFERENCES

- [1] Hans Christian von Baeyer, *Information: The New Language of Science*, Phoenix, London, 2004.