Chapter 0

PROBABILITY TOOLS AND TECHNIQUES

0.1 Probabilities and Conditional Probabilities

The theory of probability is a mathematical theory to analyze experiments with multiple outcomes where one does not know a priori which outcome will actually occur. Such experiments are usually called *random experiments.* A natural and accepted way to model such phenomena is to associate a number called *probability* to each possible outcome. These numbers are supposed to reflect the chances of occurrence of the different outcomes. How these numbers are arrived at (more specifically, the numerical value of these probabilities) is not the major concern in developing a mathematical model. It must however be noted that in practical applications of probability models, these numerical values would matter in determining how close the model is to reality. Before we go to the axiomatic definition of probability, here are a few simple and familiar examples.

Example 1: The simplest example of a random experiment is tossing a coin. Here there are two possible outcomes: either the coin lands Head up *or* Tail up. The two possibilities can conveniently be denoted by H and T respectively. A mathematical model would then associate two numbers p and q which will denote the probabilities of H and T respectively. At this point let us agree on the following convention. First, we want the chances to be non-negative numbers and second, we want the chances of all possible outcomes to add up to one. Instead of trying to justify this, let us note that this is consistent with one's intuition of 'chances'. In the absence of a priori knowledge, one is inclined to believe that p and q should be equal, which according to the above convention forces $p = q = \frac{1}{2}$.

The above model can be thought of as an abstraction of any dichotomous experiment, that is, an experiment with two possible outcomes. For example, consider a machine manufacturing bolts where each bolt produced by the

A. Goswami and B. V. Rao, *A Course in Applied Stochastic Processes*

machine has a chance of being defective. Here again we have two outcomes: defective and non-defective. We can still label them as H and T. Of course, in this case $p = q = \frac{1}{2}$ does not appear realistic because any reasonable machine is expected to produce a much larger proportion of non-defective items than defective items.

Example 2: Consider a usual six-faced die with faces numbered 1 through 6. If it is rolled once, any one of the six faces may show up. So there are six outcomes which could be denoted by the numbers 1 through 6. If nothing else is known, it seems intuitively clear that each of these outcomes should have probability 1/6.

Example 3: Pick up a name at random from the telephone directory and consider the first letter. It can be any one of the 26 letters of the alphabet. At the same time, not all the letters are equally likely to appear. For example, one certainly does not expect the letter X to occur as frequently as B. Thus it would not be reasonable to attribute equal probabilities to all the outcomes.

All the above examples show that a random experiment consists of two ingredients: first, the set of possible outcomes, to be called the *sample space* — denoted by Ω, and second, an assignment of probabilities to the various outcomes. Of course, in all the above examples, the set Ω is only a finite set, that is, $\Omega = \{\omega_1, \ldots, \omega_n\}$. In this case probability assignment means assigning non-negative numbers $p_1, \ldots, p_n$ adding up to unity, where the number p_i denotes the probability of the outcome ω_i. We write $P(\{\omega_i\}) = p_i$. Often we will be interested not in individual outcomes but with a certain collection of outcomes. For example, in rolling of a die we may ask: what is the probability that an even-numbered face shows up? In the context of a name being selected from the telephone directory we may ask: what are the chances that the letter is a vowel? These are called *events*. In general an event is any subset of the sample space. The probability of an event A is defined by

$$P(A) = \sum_{\omega \in A} P(\{\omega\})$$

where $P(\{\omega\})$ denotes the probability of the outcome ω.

Example 4: Suppose we roll a die twice. The sample space is

$$\Omega \quad = \quad \{(i,j); 1 \leq i \leq 6; 1 \leq j \leq 6\}$$

We assign equal probabilities to all the 36 outcomes, that is, for any $\omega \in \Omega$, $P(\{\omega\}) = 1/36$. If A is the event described by "first face is even", then A consists of $\{(i,j) : i = 2,4,6; 1 \leq j \leq 6\}$ and $P(A) = 1/2$. If A is described by "sum of the two faces is 5" then A consists of $\{(1,4),(2,3),(3,2),(4,1)\}$ and $P(A) = 1/9$.

As the above example shows, if, in general, we have a finite sample space with all outcomes equally likely, then for any event A, $P(A) = |A|/|\Omega|$ where, for any set B, $|B|$ denotes the number of elements of the set B. In these situations, probability computations become a combinatorial exercise.

In any case, equally likely or not, one can easily verify that probabilities of events satisfy the following properties:

1. $0 \leq P(A) \leq 1$, $P(\Omega) = 1$.
2. $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$.
   In particular, $P(A^c) = 1 - P(A)$.

So far we have restricted ourselves only to finite sample spaces but the same idea as described in the paragraph following Example 3 applies also to situations where Ω is countably infinite. With $\Omega = \{\omega_1, \omega_2, \ldots\}$ and non-negative numbers $p_1, p_2, \ldots$, adding to unity, one can define $P(A) = \sum_{\omega_i \in A} p_i$ for $A \subset \Omega$, as probability of the event A. One needs only to notice that the sum appearing in the definition of $P(A)$ may now be an infinite series. But with usual caution as necessary while dealing with infinite sums, one can show that the above properties hold and one has moreover,

3. $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$ if $A_i \cap A_j = \emptyset$ for $i \neq j$.

We now give a formal definition of probability.

Definition: Let Ω be a countable set. A *probability* on Ω is a function P defined on all subsets of Ω satisfying the following conditions.
(0) $P(\emptyset) = 0$ and $P(\Omega) = 1$
(1) $P(\cup_i A_i) = \sum_i P(A_i)$ if $A_i \cap A_j = \emptyset$ for $i \neq j$.

The next few exercises list some standard properties that are easy consequences of the definition.

Exercise 1: Let P be a probability on Ω. Then
(a) $0 \leq P(A) \leq 1$; $P(A^c) = 1 - P(A)$; if $A \subset B$ then $P(A) \leq P(B)$.
(b) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. More generally,

$$P(\bigcup_1^n A_i) = S_1 - S_2 + S_3 - \cdots \tag{1}$$

where S_i denotes the sum of probabilities of i-fold intersections.
(c) If $A_n \uparrow A$ then $P(A_n) \uparrow P(A)$. If $A_n \downarrow A$ then $P(A_n) \downarrow P(A)$.

Exercise 2: For a sequence (B_n) of events, one defines

$$\limsup_n B_n = \bigcap_n \bigcup_{k \geq n} B_k.$$

Show that $\limsup_n B_n$ is the event that B_n occurs for infinitely many n (sometimes described as the events B_n occurring *infinitely often*). Show that if $\sum P(B_n) < \infty$, then $P(\limsup_n B_n) = 0$. This is called *(the first) Borel-Cantelli Lemma.*

Exercise 3: Suppose that p is a non-negative function on Ω such that $\sum_\omega p(\omega) = 1$. Then $P(A) = \sum_{\omega \in A} p(\omega)$ defines a probability on Ω.

From now on, by a random experiment, we mean a pair (Ω, P) where Ω is a non-empty countable set and P is a probability on Ω. The number $P(A)$ represents the probability that the event A will occur when the random experiment is performed. Of course, if the experiment is really performed and we know the exact outcome, there is no need for probabilities. Probability of an event is really an assessment of the chance of occurrence of the event irrespective of whether the experiment is actually conducted and we know the outcome or not. However, sometimes we may have a situation where a random experiment is performed and some partial information is available to us about the outcome and we are to assess the chances of an event taking this additional information into account. It is intuitively clear that we should modify probability assignments of events in the presence of this additional information.

Consider the example of rolling a die twice with all outcomes being equally likely. The probability that the first face is 3 is already known to be $1/6$. But suppose now we have the additional information that the sum of the two faces is 5. This information already tells us that the outcome must be among $(1,4)$, $(2,3)$, $(3,2)$ and $(4,1)$, so that the chance of first face being 3 is now $1/4$. Such probabilities are called *conditional probabilities*. More precisely, if A is the event that the first face is 3 and B is the event that the sum of the two faces is 5, then the unconditional probability of A is $1/6$ whereas the conditional probability of A given that B has occured is $1/4$. This later probability is denoted $P(A \mid B)$. Here is the general definition.

Definition: Let (Ω, P) be a random experiment and let $B \subset \Omega$ be an event with $P(B) > 0$. Then for any event A, the *conditional probability* of A given the event B is defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{2}$$

In the equally likely case (as in the earlier example) this reduces to

$$P(A \mid B) = \frac{|A \cap B|}{|B|}.$$

The following can be easily verified:

Theorem 0.1:

1. *Fix B and let $P_B(A) = P(A \mid B)$, then P_B is a probability on Ω.*
2. *$P(A \cap B \mid C) = P(A \mid B \cap C)\, P(B \mid C)$. More generally,*

$$P(A_1 \cap \cdots \cap A_n \mid A_{n+1}) = \prod_{j=1}^{n} P(A_j \mid A_{j+1} \cap \cdots \cap A_{n+1}).$$

3. If $B_1, \ldots, B_n$ is a partition of Ω then for any event A

$$P(A) = \sum P(A \mid B_i)\, P(B_i).$$

More generally,

$$P(A \mid C) = \sum P(A \mid B_i \cap C)\, P(B_i \mid C).$$

4. If $B_1, \ldots, B_n$ is a partition of Ω then for any event A

$$P(B_i \mid A) = \frac{P(A \mid B_i)\, P(B_i)}{\sum_j P(A \mid B_j)\, P(B_j)}.$$

Exercise 4: $P(A \mid B) \leq P(A)$ if and only if $P(B \mid A) \leq P(B)$. In particular, $P(A \mid B) = P(A)$ if an only if $P(B \mid A) = P(B)$.

Let us return to the example of rolling a die twice. Let, as earlier, A be the event that the first face is 3 and B be the event that the sum of the two faces is 5. Then $P(A \mid B) = 1/4 > 1/6 = P(A)$. So here the additional information has the effect of increasing the chances of A. On the other hand if we consider the event C that the sum is 11, then clearly $P(A \mid C) = 0$, that is, the additional information reduces the chances of A (to indeed zero!). Does it always happen this way? That is, will additional information always change the chances one way or other? The answer is NO. For example if D is the event that the sum is 7, then $P(A \mid D) = 1/6 = P(A)$. That is, the probability of A remains unchanged even if we are told that D has occured. This situation is described by saying that A is *independent* of D. Here is the precise definition.

Definition: Two events A and B are said to be *independent* if $P(A \cap B) = P(A)P(B)$.

Of course when one of the two events, say, B has positive probability then A and B are independent is the same as saying $P(A \mid B) = P(A)$.

Exercise 5: If A, B are independent, then A^c, B are independent ; A, B^c are independent ; A^c, B^c are independent.

Definition: Events A_1, A_2, ..., A_n are said to be *independent* if for any $1 \leq i_1 < i_2 < \cdots < i_k \leq n$

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}). \tag{3}$$

Exercise 6: Let A_1, A_2, ..., A_n be independent.
(i) If for each i, B_i denotes one of the events A_i or A_i^c then B_1, B_2, ..., B_n are independent.
(ii) If $1 \leq j < n$, $\cap_{1 \leq j} B_i$ is independent of $\cap_{i>j} B_i$. $\cup_{1 \leq j} B_i$ is independent of $\cap_{i>j} B_i$. $\cup_{1 \leq j} B_i$ is independent of $\cup_{i>j} B_i$. Here B_i are as in (i).

The assertions in (ii) above are merely special cases of a more general phenomenon: if $1 \leq j < n$ and C is an event "constructed" out of $A_1, \ldots, A_j$ and

D is an event constructed out of $A_{j+1}, \ldots, A_n$, then C and D are independent events. This is intuitively clear, but a formal proof requires more machinery than what is available at this level.

Often random experiments can be thought of as composed of simpler random experiments in the sense explained below. If you toss a coin twice you can describe the outcomes of the experiment by $\Omega = \{HH, HT, TH, TT\}$. Notice that $\Omega = \{H, T\} \times \{H, T\}$, that is, Ω is the two-fold product of a single toss experiment. More generally, the sample space for 10 tosses of a coin (or a toss of 10 coins) can be thought of as the ten-fold product of $\{H, T\}$. But what is important is that not only the sample space can be thought of as a product, but the probabilities of the outcomes can also be thought of as products. Here is the general method.

Let (Ω_i, P_i), for $1 \leq i \leq n$, be random experiments. Put

$$\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n = \{(\omega_1, \ldots, \omega_n) : \forall i, \quad \omega_i \in \Omega_i\}.$$

For $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$, put $P(\{\omega\}) = P_1(\{\omega_1\}) \times \cdots \times P_n(\{\omega_n\})$. One can now define $P(A)$ for any $A \subset \Omega$, thus getting a probability P on Ω.

Exercise 7: If $A = A_1 \times A_2 \times \cdots \times A_n$ then $P(A) = \prod_{i=1}^{n} P_i(A_i)$. Conclude that if $\widetilde{A_i} \subset \Omega$ is the set of all points in Ω whose i-th coordinate is in A_i then $\widetilde{A_1}, \ldots, \widetilde{A_n}$ are independent.

The exercise above really means that events that depend on different coordinates are independent. This, of course, is a consequence of the way the probability P has been defined on Ω. It is clearly possible to construct other probabilities P on Ω, such that $P(\widetilde{A_i}) = P_i(A_i)$ for all i, but independence fails. One can easily see that 10 tosses of a coin with all outcomes equally likely is the same as the ten-fold product of single toss of coin with $P(H) = P(T) = 1/2$.

If $\Omega_1 = \Omega_2 = \cdots = \Omega_n$, then we write $\Omega = \Omega_1^n$. If further $P_1 = P_2 = \cdots = P_n$, then we write $P = P_1^n$. (Ω_1^n, P_1^n) represents n independent repetitions of the experiment (Ω_1, P_1).

0.2 Random Variables and Distributions

In the context of random experiments, the actual outcomes may often be quite abstract. For example, if you toss a coin 10 times, outcomes will be 10-tuples of H's and T's. Often one is interested not in the exact outcome per se but some numerical value associated with each outcome. For example, in case of 10 tosses of a coin, one may be interested in the number of times heads showed up *or* in the number of times a tail was immediately followed by a head. Such numerical values associated with outcomes are what are called *random variables.* This section is devoted to a study of random variables and their distributions.

0.2.1 Distribution of a Random Variable

Definition: A *random variable* is a real-valued function defined on the sample space Ω.

It is customary to denote random variables by X, Y, Z etc. For example, in 10 tosses of a coin, let X denote the total number of heads and Y denote the number of times a tail is immediately followed by a head. Then for the outcome $\omega = HTTHTTTHHH$, $X(\omega) = 5$ and $Y(\omega) = 2$, while for another outcome $\omega' = THHTHTHHTH$, $X(\omega') = 6$ and $Y(\omega') = 4$.

Given a random variable, we can ask what the possible values of the random variable are and the chances (probabilities) of it taking each of those values. This is what is called the *distribution* of the random variable. Since our sample space is countable, any random variable can only take countably many values.

Definition: Let X be a random variable on (Ω, P). Then by the *distribution* of X is meant the set of possible values $D = \{x_1, x_2, \ldots\}$ of the random variable X and the probabilities $\{p(x_1), p(x_2), \ldots\}$ where $p(x_i) = P(\omega : X(\omega) = x_i)$. The right side is often abbreviated as $P(X = x_i)$.

Of course, p can be extended to a function on R by setting $p(x) = P(X = x)$. However, for any $x \notin D$ we have $p(x) = 0$. This p is called the *probability mass function* (p.m.f.) of the random variable X.

Once we know the probability mass function of a random variable X, we can compute for any $A \subset R$, the probability $P(X \in A)$ by the formula

$$P(X \in A) = \sum_{x \in A} p(x).$$

Example 1: Consider n independent tosses of a coin. Assume that in each toss the probability of heads is p. Define X to be the total number of heads obtained. Clearly X is a random variable which can take any integer value from 0 to n. One might wonder: how do we get a random variable even before describing the sample space. We concede that we were jumping steps. So here is our sample space: $(\Omega, P) = (\Omega_1^n, P_1^n)$ where $\Omega_1 = \{H, T\}$; $P_1(H) = p$ and $P_1(T) = 1 - p$. The definition of the random variable X as a real-valued function on Ω should now be clear. It is also easy to verify that the probability mass function of X is given by

$$\begin{aligned} p(x) &= \binom{n}{x} p^x (1-p)^{n-x} && \text{for } x \in \{0, 1, \ldots, n\} \\ p(x) &= 0 && \text{for } x \notin \{0, 1, 2, \ldots, n\} \end{aligned}$$

This random variable is called the *Binomial* random variable with parameters n and p, in short, a $B(n, p)$ random variable. We write $X \sim B(n, p)$ for this. The distribution is called the *Binomial distribution.*

Almost all the information about the random variable X is contained in its distribution (or its p.m.f.) — the underlying sample space or the precise definition of X as a function on the sample space is of no additional importance.

Therefore it is often customary to describe random variables simply by their distributions without any reference to any underlying sample space.

Example 2: Fix a number p with $0 < p < 1$. A random variable X is said to have $G(p)$ distribution — *geometric distribution with parameter* p — if X takes value x with probability $p(1-p)^x$ for $x \in \{0, 1, \ldots\}$. In other words, X has p.m.f.

$$p(x) = p(1-p)^x \quad \text{for} \quad x \in \{0, 1, \ldots\}$$

It is to be understood here and elsewhere that $p(x) = 0$ for all other x. Suppose you have a coin with chance of heads p. If the coin is tossed repeatedly until a head shows up, then the number of tails preceeding the head has this geometric distribution.

Example 3: Here is a generalization of the above example. Again we have a coin for which the chance of a head in each toss is p. Fix an integer $m \geq 1$. Toss the coin until a total of m heads show up. (What is the sample space?) The random variable X is the total number of tails obtained. Clearly X takes values $x = 0, 1, 2, \ldots$ as earlier. A simple combinatorial argument shows that $P(X = x) = \binom{x+m-1}{x-1}(1-p)^x p^m$. This random variable is called a *negative binomial* random variable with parameters (m, p) — in short, $NB(m, p)$ random variable — and the distribution is called the negative binomial distribution (why?). Clearly when $m = 1$, we get the geometric random variable of Example 2.

Example 4: Fix integers N, $n < N$ and $N_1 < N$. A random variable X is said to be $Hyp(N, N_1; n)$ — *hypergeometric with parameters* N, N_1 *and* n — if it takes value x with probability

$$p(x) = \binom{N_1}{x}\binom{N - N_1}{n - x} \Big/ \binom{N}{n}.$$

Of course you have to interpret $\binom{a}{b} = 0$ unless b is an integer with $0 \leq b \leq a$. This arises if you have a bunch of N items of which N_1 are good, the remaining are defective and you select a random sample of size n without replacement. The random variable in question is the number of good items in the sample.

Example 5: Fix a number $\lambda > 0$. A random variable X is said to be $P(\lambda)$, written $X \sim P(\lambda)$ — *Poisson with parameter* λ — if it takes value x with probability $e^{-\lambda}\lambda^x/x!$ for $x = 0, 1, 2, \ldots$. This random variable arises as a limiting case of the number of heads when you toss a coin a large number of times and the chance of heads in each toss is very small. For details see Section 0.3.

Example 6: Roll a fair die twice and let X be the sum of the two numbers obtained. Then X takes values

$$2, \quad 3, \ldots, \quad 7, \quad 8, \ldots, \quad 12$$

with probabilities given respectively by

$$1/36, 2/36, \ldots, 6/36, 5/36, \ldots, 1/36.$$

Suppose that a fair coin is tossed ten times and X is the number of heads. Clearly X can take any one of the values $0, 1, 2, \ldots, 10$ with different probabilities, the actual value depending on the outcome of the ten tosses. But if we were to choose one "representative value" of X without knowing the actual outcome, what would be a good candidate? One possibility is to consider the most probable value, which in this case is 5. However a commonly used and mathematically more tractable quantity is what is known as the *expected value*. As the next definition shows, this is weighted average of the possible values.

Definition: Let X be a random variable with set of values D and p.m.f. $p(x)$ for $x \in D$. If $\sum_{x \in D} |x|\, p(x) < \infty$ (automatically true if D is finite), then X is said to have a finite expectation and the *expected value* of X is defined to be

$$E(X) = \sum_{x \in D} x\, p(x)\,. \tag{4}$$

Thus, expected value of a random variable, when it exists, is the weighted average of its values, weighted by their probabilities. Expected value or *expectation* is also called the *mean value* or the *mean*.

If X is a random variable and $g : R \longrightarrow R$ is a function then clearly $g(X)$ is again a random variable. It is not difficult to check that $g(X)$ has finite expectation iff $\sum_{x \in D} |g(x)|\, p(x) < \infty$ and in that case $E(g(X)) = \sum_{x \in D} g(x)\, p(x)$. This is a very useful formula because we can compute $E(g(X))$ straight from the p.m.f. of X rather than having to go to the p.m.f. of the random variable $g(X)$.

Definition: A random variable X with $E(X^m)$ finite is said to have a *finite m-th moment*, given by $E(X^m)$. For X with a finite second moment, the *variance* of X, denoted $V(X)$, is defined by

$$V(X) = E[(X - EX)^2]\,. \tag{5}$$

The quantity $V(X)$ measures the spread of the distribution of X. For example, $V(X) = 0$ iff the distribution of X is concentrated at one point (that is, X is a constant random variable).

Indicator random variables as defined below form a very simple, yet useful, class of random variables.

Definition: For any event A the *Indicator random variable* of the event is defined as

$$I_A(\omega) = \begin{cases} 1 & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \end{cases}\,.$$

Clearly the expectation of I_A is $P(A)$.

Exercise 1: If the m-th moment is finite, then so is the n-th, for any $n < m$.

Exercise 2. For a random variable X on a probability space (Ω, P), $E(X)$ exists iff $\sum_{\omega \in \Omega} |X(\omega)|\, P(\{\omega\}) < \infty$ and in that case $E(X) = \sum_{\omega \in \Omega} X(\omega)\, P(\{\omega\})$.

Exercise 3. If $P(X = c) = 1$ for some real number c, then $E(X) = c$.

Exercise 4. If X is a random variable with finite expectation, then $|E(X)| \leq E(|X|)$.

Exercise 5. If X and Y are two random variables defined on the same space and having finite expectations, then

(a) $X \leq Y$ implies $E(X) \leq E(Y)$.

(b) $E(aX + bY) = a\,E(X) + b\,E(Y)$ for any two real numbers a and b. In particular, $E(aX + b) = a\,E(X) + b$.

Exercise 6. If $P(X \geq 0) = 1$ and $E(X) = 0$, then $P(X = 0) = 1$. More generally, if $P(X \geq Y) = 1$ and $E(X) = E(Y)$, then $P(X = Y) = 1$.

Exercise 7. If X_n and X are non-negative random variables defined on the same space and $X_n \uparrow X$, then $E(X_n) \uparrow E(X)$. In case X has infinite expectation, this should be read as $E(X_n) \uparrow \infty$. This is known as Lebesgue's *Monotone Convergence Theorem.*

Exercise 8. Supose that X_n and X are random variables defined on the same space such that $X_n \to X$. Suppose also that there is a random variable Y with finite expectation such that $|X_n| \leq Y$ for all n, that is, all the random variables X_n are dominated in modulus by the random variable Y. Then $E|X_n - X| \longrightarrow 0$. In particular $E(X_n) \to E(X)$. This is called Lebesgue's *Dominated Convergence Theorem.*

Exercise 9. If X and Y are two random variables on the same space such that $E(X \cdot I_A) \geq E(Y \cdot I_A)$ for every event A then $P(X \geq Y) = 1$. In particular, $E(XI_A) = E(YI_A)$ for every A if and only if $P(X = Y) = 1$.

Exercise 10. $V(X) = E(X^2) - (EX)^2$.

Exercise 11. $V(X) = 0$ iff $P(X = c) = 1$ for some constant c.

Exercise 12. $V(aX + b) = a^2V(X)$.

Exercise 13. $V(I_A) = P(A)[1 - P(A)]$.

Exercise 14. If X has finite variance and $E(X) = \mu$, then $E(X - a)^2 \geq V(X)$ for every real a. Thus $E(X - a)^2$ is minimized when $a = \mu$.

Exercise 15. For each of the random variables in Examples 1 through 6, find its expected value and variance.

0.2.2 Joint Distributions

Suppose that X and Y are two random variables defined on the same space. As mentioned earlier, probabilities of events concerning the random variable X (respectively, Y) can be computed from the distribution of X (respectively, of Y). However we may often be interested in probabilities of events that concern both X and Y. For example, 'what is $P(X = Y)$?' or 'what is $P(X + Y = 7)$?' etc. For such probabilities individual distributions of X and Y alone would not suffice. We need to know what is called the *joint distribution* of X and Y.

Definition: Let X and Y be two random variables defined on the same space.

Let D_X and D_Y denote the set of possible values of the random variables X and Y respectively. The set of possible values of the pair (X, Y) are clearly contained in $D_X \times D_Y$. The *joint distribution* of (X, Y) is given by the *joint probability mass function* defined as $p(x, y) = P(X = x, Y = y)$ for $(x, y) \in D_X \times D_Y$.

Consider, for example, tossing a coin 15 times, with the chance of a head in each toss being p. Let X be the number of heads in the first ten tosses and Y be the number of heads in the last ten tosses. Clearly both X and Y are $B(10, p)$ random variables. Here $D_X = D_Y = \{0, 1, \ldots, 10\}$. The joint distribution of (X, Y) would be given by the mass function p on $D_X \times D_Y$. For example, $p(10, 10) = p^{15}$. In general,

$$p(m, n) = \sum_{k=0}^{5} \binom{5}{m-k}\binom{5}{k}\binom{5}{n-k} p^{m+n-k}(1-p)^{15+k-m-n}$$

with the usual convention that $\binom{a}{b} = 0$ unless b is integer with $0 \leq b \leq a$.

From the joint p.m.f. of (X, Y), the individual (marginal) p.m.f. of X and Y can be obtained as follows:

$$p_1(x) = P(X = x) = \sum_{y \in D_Y} p(x, y) \quad \text{for} \quad x \in D_X$$

$$p_2(y) = P(Y = y) = \sum_{x \in D_X} p(x, y) \quad \text{for} \quad y \in D_Y$$

In an analogous way the joint distribution of n random variables (defined on the same space) is given by their joint p.m.f.

$$p(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

Example 7: Consider an n faced die with p_1, p_2, ..., p_n denoting the probabilities of different faces in a single throw. Roll the die r times and let X_i be the number of times face i shows up. The joint p.m.f. of $(X_1, X_2, \ldots, X_n)$ is given by

$$p(x_1, x_2, \ldots, x_n) = \frac{r!}{x_1! x_2! \cdots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n}$$

for $x_1, x_2, \ldots, x_n$ non-negative integers adding to r. This distribution is called the *multinomial distribution with parameters* $(r; p_1, p_2, \ldots, p_n)$.

Definition: For a pair of random variables X and Y defined on the same space, the *covariance* between X and Y is defined as

$$Cov(X, Y) = E[(X - EX)(Y - EY)]. \tag{6}$$

Further, $E(X^m Y^n)$, for positive integers m and n, are called the various *cross-product moments* of the pair (X, Y).

Exercise 16. $Cov(X, Y) = E(XY) - E(X)E(Y)$.

Exercise 17. $Cov(\sum_i a_i X_i, \sum_j b_j Y_j) = \sum_i \sum_j a_i b_j Cov(X_i, Y_j)$.

Exercise 18. $Cov(X, X) = V(X)$.

Exercise 19. $Cov(X, a) = 0$ for any constant random variable a.

Exercise 20. $Cov(X, Y) \leq \sqrt{V(X)}\sqrt{V(Y)}$.

Exercise 21. $V(\sum_i X_i) = \sum_i V(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$. In particular, if $Cov(X_i, X_j) = 0$ for all $i \neq j$, then $V(\sum_i X_i) = \sum_i V(X_i)$.

0.2.3 Conditional Distributions and Conditional Expectations

Let X be a random variable. For any event A with $P(A) > 0$, the conditional distribution of X given A simply means the conditional probabilities for X taking various values given the event A. Thus the conditional distribution is given by the (conditional) p.m.f. $p(x \mid A) = P(X = x \mid A)$. It is, of course, immediate that this is indeed a probability mass function. The conditional expectation and conditional variance of X given A are just the expectation and variance of this conditional distribution. Clearly all the properties listed in Exercises 4,5,7,8 and 10 through 14 can be formulated and shown to hold with conditional expectation and conditional variance.

Next, let X and Y be two random variables defined on the same space. For y with $P(Y = y) > 0$, we can talk about the *conditional distribution* of X given $Y = y$. This is given by the conditional mass function

$$p(x \mid y) = P(X = x \mid Y = y) = \frac{p(x, y)}{p_2(y)} \,. \tag{7}$$

Here p is the joint p.m.f. of (X, Y) and p_2 is the (marginal) p.m.f. of Y. It is clear that for each y with $p_2(y) > 0$, the function $p(\cdot \mid y)$ is a probability mass function — called the *conditional p.m.f.* of X given $Y = y$.

If X has finite expectation, then the *conditional expectation* of X given $Y = y$ is defined to be

$$E(X \mid Y = y) = \sum x\, p(x \mid y) \,. \tag{8}$$

The assumption of finite expectation ensures the convergence of the right hand side of Equation (8). Thus, the conditional expectation of X given $Y = y$ is just the expectation of X under the conditional distribution given $Y = y$. Clearly $E(X \mid Y = y)$ is a function of y, say $\phi(y)$. The random variable $\phi(Y)$ is denoted by $E(X \mid Y)$. We do this because, in many contexts it is convenient to think of the conditional expectation itself as a random variable. One can similarly define the conditional distribution of Y given $X = x$ and also $E(Y \mid X)$.

It may be noted that if Y is a constant random variable, say, $Y \equiv c$, then the conditional distribution as well as the conditional expectation of X given $Y = c$ reduce to the unconditional distribution and unconditional expectation

of X. The following facts on conditional expectation are easy to verify, and left as exercises.

Exercise 22. $E(E(X \mid Y)) = E(X)$.

Exercise 23. If X has finite expectation and if g is a function such that $Xg(Y)$ also has finite expectation, then show that $E(X\, g(Y) \mid Y) = E(X \mid Y)g(Y)$.

Exercise 24. $E(X - g(Y))^2 \geq E(X - E(X \mid Y))^2$ for any X and g such that X^2 and $(g(Y))^2$ have finite expectations. (Exercise 14 in 0.2.1 is easily seen to be a special case of the above.)

Exercise 25. For any function g such that $g(X)$ has finite expectation, $E(g(X) \mid Y = y) = \sum g(x)\, p(x \mid y)$.

Exercise 26. $|E(X \mid Y)| \leq E(|X| \mid Y)$.

The above notions of conditional distribution and conditional expectation naturally extend to the case of more then two random variables. To be precise, if $X_1, X_2, \ldots, X_n$ are random variables on the same space, one can, in a natural way, talk about the conditional joint distribution of k of these random variables given the others. For instance, the conditional joint distribution of $(X_1, \ldots, X_k)$, given $X_{k+1} = x_{k+1}, \ldots, X_n = x_n$ is defined by

$$p(x_1, \ldots, x_k \mid x_{k+1}, \ldots, x_n) = \frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(X_{k+1} = x_{k+1}, \ldots, X_n = x_n)}\,.$$

provided, of course, $P(X_{k+1} = x_{k+1}, \ldots, X_n = x_n) > 0$, and for each such $(x_{k+1}, \ldots, x_n)$, the function $p(\cdot \mid x_{k+1}, \ldots, x_n)$ is a p.m.f. — called the *conditional joint p.m.f.* of $(X_1, \cdots, X_k)$, given $X_{k+1} = x_{k+1}, \ldots, X_n = x_n$.

If g is a k-variable function such that $Y = g(X_1, \ldots, X_k)$ has finite expectation, then the conditional expectation of Y given $X_{k+1} = x_{k+1}, \ldots, X_n = x_n$, has a natural definition, namely

$$\begin{aligned} &E(Y \mid X_{k+1} = x_{k+1}, \ldots, X_n = x_n) \\ &\qquad = \sum_{x_1, \ldots, x_k} g(x_1, \ldots, x_k)\, p(x_1, \ldots, x_k \mid x_{k+1}, \ldots, x_n)\,. \end{aligned}$$

In particular, one can talk about the conditional expectation of X_1 given $X_2, \ldots, X_n$ or conditional expectation of $X_1^2 + X_2^2$ given X_3, X_5, and so on.

Exercise 27. $E(E(X \mid Y, Z) \mid Y) = E(X \mid Y)$. More generally

$$E(E(X \mid X_1, \ldots, X_n) \mid X_1, \ldots, X_{n-1}) = E(X \mid X_1, \ldots, X_{n-1})\,.$$

Here $E(Y \mid X_{k+1}, \ldots, X_n)$ denotes the random variable $\phi(X_{k+1}, \ldots, X_n)$ where ϕ is the $(n-k)$-variable function defined by

$$\phi(x_{k+1}, \ldots, x_n) = E(Y \mid X_{k+1} = x_{k+1}, \ldots, X_n = x_n)\,.$$

If these things look a little abstract there is no cause for alarm. Simply try to understand the meaning of the conditional expectation of X_1 given

$X_2, \ldots, X_n$ or the conditional expectation of $X_1^2 + X_2^2$ given X_3 and X_4. Here is a useful exercise left to be proved by the reader. This is often referred to as the *smoothing property* of conditional expectation.

$$E(E(g(X,Y)\,|\,Z,W)\,|\,Z) = E(g(X,Y)\,|\,Z)\,.$$

Or more generally, if $U = g(X_1, \ldots, X_m)$ then

$$E(E(U\,|\,Y_1, \ldots, Y_n)\,|\,Y_1, \ldots, Y_{n-1}) = E(U\,|\,Y_1, \ldots, Y_{n-1})\,. \tag{9}$$

Indeed, one may think of (9) as equivalent to Exercise 27 above. What this says is the following. In order to get the conditional expectation of a random variable given $Y_1, Y_2, \ldots, Y_{n-1}$, one may first calculate its conditional expectation given $Y_1, Y_2, \ldots, Y_n$ and then take the conditional expectation of this random variable given $Y_1, Y_2, \ldots, Y_{n-1}$. Here is an application.

Example 8: Toss a fair coin a Poisson number of times. Find the conditional expectation of the time of occurrence of the first Head, given the total number of Heads. More precisely, let N be a random variable having the Poisson distribution with parameter λ. Suppose that a fair coin is tossed N times. Let X be the number of Heads obtained and T be the time of occurrence of the first Head. In case there are no Heads, T is defined to be one plus the number of tosses, that is to say, $T = 1 + N$ in case $X = 0$. Of course, if $N = 0$, then $X = 0$ automatically so that $T = 1$. We want $E(T\,|\,X = x)$ for each $x \geq 0$.

The plan is the following. We first compute $E(T\,|\,X, N)$ and then compute its conditional expectation given X. By the smoothing propertry this will be the same as $E(T\,|\,X)$.

For integers $0 \leq x \leq n$, let $f(n,x) = E(T\,|\,N = n, X = x)$. In case $x = 0$, by our convention made above, $f(n,0) = 1 + n$ clearly. For $1 \leq x \leq n$, $f(n,x)$ is simply the expected waiting time till the first head, given that n tosses of a fair coin has resulted in a total of x heads. For the sake of completeness, we set $f(n,x) = 0$ (or any other value, for that matter) for $x > n \geq 0$. We now proceed to obtain a recurrence relation among the $f(n,x)$. For $1 \leq x \leq n$, we obtain, by conditioning on the outcome of the first toss,

$$f(n,x) = \alpha + \beta\,,$$

where

$$\alpha = E(T\,|\,x \text{ heads in } n \text{ tosses, first is heads}) \cdot P(\text{first heads}\,|\,N = n, X = x)\,,$$

$$\beta = E(T\,|\,x \text{ heads in } n \text{ tosses, first is tails}) \cdot P(\text{first tails}\,|\,N = n, X = x)\,.$$

A routine calulation now shows that

$$\alpha = \frac{\binom{n-1}{x-1}}{\binom{n}{x}} = \frac{x}{n} \quad \text{and}$$

$$\beta = [1 + f(n-1,x)]\,\frac{\binom{n-1}{x}}{\binom{n}{x}} = \frac{n-x}{n}\,[1 + f(n-1,x)]\,,$$

giving us the recurrence relation

$$f(n,x) = 1 + \frac{n-x}{n} f(n-1,x)\,.$$

Since $f(x,x) = 1$, we get by induction on n, that for $n \geq x$,

$$f(n,x) = \frac{n+1}{x+1}\,.$$

(Try to directly compute the conditional expectation $E(T \,|\, N = n, X = x)$.) Thus $E(T|X,N) = (N+1)/(X+1)$. To calculate the conditional expectation of this given $X = x$ we calculate the conditional distribution of N given $X = x$. Clearly, $P(N < x \,|\, X = x) = 0$ and for $n \geq x$,

$$P(N = n \,|\, X = x) = e^{-\lambda/2} (\lambda/2)^{n-x} \frac{1}{(n-x)!}$$

As a consequence for $x \geq 1$,

$$E(T \,|\, X = x) = E[\,(N+1)/(X+1) \;|\; X = x] = \frac{x + \frac{\lambda}{2} + 1}{x+1} = 1 + \frac{\lambda}{2(x+1)}$$

Even though given $X = 0$, T equals $1 + N$ and $E(N) = \lambda$, it does not mean that $E(T \,|\, X = 0) = 1 + \lambda$; indeed $E(T \,|\, X = 0) = 1 + \frac{\lambda}{2}$ (why?).

0.2.4 Independence

Definition: Random variables $X_1, X_2, \ldots, X_n$ are said to be *independent* if for any $x_1, x_2, \ldots, x_n$,

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)\,. \tag{10}$$

Thus independence requires that the joint p.m.f. is just the product of the marginal probability mass functions. Moreover (10) is clearly equivalent to saying that for sets $B_1, B_2, \ldots, B_n$, the events $\{(X_i \in B_i), 1 \leq i \leq n\}$ are independent. Also, independence of $X_1, X_2, \ldots, X_n$ clearly implies independence of $X_{j_1}, X_{j_2}, \ldots, X_{j_m}$ for any $1 \leq j_1 < j_2 < \cdots < j_m \leq n$. With some work, one can also show the following. Let $1 \leq i_1 < i_2 < \cdots < i_{k-1} \leq n$ and consider the random variables $Y_1, Y_2, \cdots, Y_k$ defined as $Y_1 = g_1(X_1, X_2, \ldots, X_{i_1})\,, Y_2 = g_2(X_{i_1+1}, \ldots, X_{i_2})\,, \ldots,\ Y_k = g_k(X_{i_{k-1}+1}, \ldots, X_n)$, for functions $g_1, g_2, \ldots, g_k$. Then independence of $X_1, X_2, \ldots, X_n$ implies that of $Y_1, Y_2, \ldots, Y_k$. Here are some more consequences of the definition of independence that the reader should work out.

Exercise 28. If $X_1, X_2, \ldots, X_n$ are independent, then the conditional joint distribution of any subset of them, given the others, is the same as the unconditional joint distribution.

Exercise 29. A constant random variable is independent of any random variable. Moreover, a random variable is independent of itself if and only if it is a

constant random variable.

Exercise 30. If $X_1, X_2, \ldots, X_n$ are independent random variables with finite expectations, then the product $\Pi_{i=1}^n X_i$ also has finite expectation and $E(\Pi_{i=1}^n X_i) = \prod_{i=1}^n E(X_i)$.

Exercise 31. If X and Y are independent with finite expectations, then $Cov(X,Y) = 0$. In particular, if X and Y have finite variances, then $V(X+Y) = V(X) + V(Y)$.

Exercise 32. Give an example of random variables X and Y such that $Cov(X,Y) = 0$, but X and Y are not independent.

Exercise 33. Suppose that X and Y are independent random variables and suppose that g is a function such that $Z = g(X,Y)$ has finite expectation, then $E(Z \,|\, Y = y) = E(g(X,y))$. More generally, if $X_1, X_2, \ldots, X_n$ are independent random variables and g is a function such that $Z = g(X_1, X_2, \ldots, X_n)$ has finite expectation, then

$$E(Z \,|\, X_{k+1} = x_{k+1}, \ldots, X_n = x_n) = E(g(X_1, \cdots, X_k, x_{k+1}, \ldots, x_n)) .$$

Exercise 34. In fifteen tosses of a fair coin, let X_1 be the number of heads in the first three tosses, X_2 be the number of tails in the next six tosses, and X_3 be the number of heads minus the number of tails in the last six tosses. Show that X_1, X_2 , X_3 are independent. Find $E(X_1 X_2 X_3)$.

0.3 Generating Functions

Let $(a_k)_{k\geq 0}$ be a sequence of numbers with $0 \leq a_k \leq 1$ for all k. Then clearly for any $t \in (-1,1)$ the series $\sum_{k=0}^{\infty} a_k t^k$ converges absolutely. The function $A(t) = \sum_{k=0}^{\infty} a_k t^k$ defined for $t \in (-1,1)$ is called the *generating function* of the sequence $(a_k)_{k\geq 0}$. By the uniqueness of the Taylor expansion, the function $A(t)$ determines the sequence (a_k) completely. Indeed, the function $A(t)$ is infinitely differentiable on $(-1,1)$ and $a_k = A^{(k)}(0)/k!$ where $A^{(k)}(0)$ is the k-th derivative of the function $A(t)$ at $t = 0$. Moreover as $t \uparrow 1$, $A(t)$ also increases and the limit $\lim_{t\uparrow 1} A(t)$ is finite iff $\sum a_k$ converges. In fact $\lim_{t\uparrow 1} A(t) = \sum a_k$. We denote this limit by $A(1)$. It should however be noted that in case $\sum a_k$ does not converge, then $A(t)$ increases to ∞. In this latter case also, we say that the limit $A(1) = \lim_{t\uparrow 1} A(t)$ exists and equals infinity. It is known from calculus that the derivative of the function $A(t)$ also has a power series expansion in the interval $(-1,1)$ given by $A'(t) = \sum_{k=1}^{\infty} k a_k t^{k-1}$. In fact, one can similarly get power series expansions for higher order derivatives. Once again as $t \uparrow 1$, $A'(t)$ has a finite limit iff $\sum ka_k$ converges and $\lim_{t\uparrow 1} A'(t) = \sum_{k\geq 1} ka_k$. This equality remains valid even if the right-hand side does not converge. We denote this limit by $A'(1)$. In general we will always use the notation $A^{(k)}(1)$ for the limit $\lim_{t\uparrow 1} A^{(k)}(t)$, finite or not.

By the *convolution* of two sequences $(a_k)_{k\geq 0}$ and $(b_k)_{k\geq 0}$ is meant the new sequence $c_k = (a * b)_k$ defined by $c_k = \sum_{l=0}^{k} a_l b_{k-l}$. It is easy to see that the generating function of the convolution of two sequences equals the product of the corresponding generating functions. That is, $C(t) = A(t)B(t)$.

A particularly interesting case arises when the sequence $(a_k)_{k\geq 0}$ is the probability mass function of a non-negative integer-valued random variable X. In that case, $A(t)$ is denoted by $\varphi_X(t)$ and is called the *probability generating function (p.g.f.)* or *generating function (g.f.)* of X. From our earlier discussion, it follows that the distribution of a non-negative integer valued random variable is completely determined by its p.g.f. Indeed, for such a random variable X, $P(X = k) = \varphi_X^{(k)}(0)/k!$. Clearly $\varphi_X(1) = 1$. Also

$$\varphi'_X(1) = \lim_{t\uparrow 1} \varphi'_X(t) = E(X)\,, \tag{11}$$

whether this expectation is finite or not. It is left as an exercise to show that, in case X has finite variance,

$$V(X) = \varphi''_X(1) + \varphi'_X(1) - [\varphi'_X(1)]^2 \tag{12}$$

Exercise 1. (i) If $X \sim B(n,p)$, then show that its p.g.f is $\varphi_X(t) = (1-p+pt)^n$. (ii) If $X \sim P(\lambda)$, then show that $\varphi_X(t) = e^{-\lambda(1-t)}$. (iii) If $X \sim NB(k,p)$, then show that $\varphi_X(t) = p^k(1-qs)^{-k}$.

Exercise 2. If X and Y are independent non-negative integer valued random variables, then show that $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.

Exercise 3. Show that (i) the sum of independent $B(n,p)$ and $B(m,p)$ random variables is $B(n+m,p)$; (ii) the sum of two independent Poisson random variables is again Poisson; (iii) the sum of independent $NB(k,p)$ and $NB(l,p)$ random variables is again NB.

Exercise 4. Let $X_1, X_2, \ldots$ and N be non-negative integer-valued random variables. Suppose that, for every $k \geq 1$, the $(k+1)$ random variables $X_1, X_2, \ldots, X_k$ and N are independent. Suppose further that the X_i have a common distribution with p.g.f. $\psi(t)$. Define $Z = \sum_{i\leq N} X_i$, with the convention that if $N = 0$, then this sum is zero. Show that the p.g.f. of Z is $\varphi_N(\psi(t))$, where φ_N is the p.g.f. of N. In particular, show that if each $X_i \sim B(1,p)$ and $N \sim P(\lambda)$, then $Z \sim P(\lambda p)$.

Exercise 5. Let $\varphi(s)$ be the p.g.f. of a random variable X. Let $q_k = P(X > k)$ for $k \geq 0$. Then the function $Q(s) = \dfrac{1-P(s)}{1-s}$ is the generating function of the sequence $(q_k)_{k\geq 0}$.

Suppose that $X_n \sim B(n,p_n)$, and denote np_n by λ_n. Then the p.g.f. of X_n is

$$\varphi_n(t) = (1-p_n+p_nt)^n = (1+p_n(t-1))^n = \left[1+\frac{\lambda_n}{n}(t-1)\right]^n .$$

If we assume that $\lambda_n \to \lambda$ then clearly

$$\varphi_n(t) \to \varphi(t) = e^{\lambda(t-1)}$$

which is the p.g.f. of $P(\lambda)$ random variable. From this it looks plausible that the distribution of X_n converges to a $P(\lambda)$ distribution. That is, for $k \geq 0$,

$$P(X_n = k) \to e^{-\lambda}\lambda^k/k!$$

This is indeed true and is actually a consequence of the next theorem.

Theorem: *For each $n \geq 1$, let φ_n be the generating function of a sequence of numbers $(a_{n,k})_{k\geq 0}$. In order that $\lim_{n\to\infty} a_{n,k} = a_k$ exists for each k, it is necessary and sufficient that $\lim_{n\to\infty} \varphi_n(s) = \varphi(s)$ exists for each $s \in (0,1)$. In that case, φ is actually the generating function of the sequence $(a_k)_{k\geq 0}$.*

Remark: It may be noted that even when the φ_n are p.g.f. of a sequence of random variables, the limit function φ need not be a p.g.f. — that is, even if $\sum_k a_{n,k} = 1$ for each n, the sequence a_k may not be a probability distribution (consider $\varphi_n(s) = s^n$). Of course $\sum a_k \leq 1$ will always hold.

Proof of Theorem: Let $\varphi_n(s) = \sum_k a_{n,k}s^k$. First assume that for each k, $a_{n,k} \to a_k$ as $n \to \infty$. Clearly $0 \leq a_k \leq 1$ for each k. Let $\varphi(s)$ be the generating function of the sequence (a_k). Fix $s \in (0,1)$ and $\epsilon > 0$ be given. Choose k_0 large enough so that $s^{k_0} < \frac{1}{4}\epsilon(1-s)$. Since $\lim_n a_{n,k} = a_k$ for each k, we choose n_0 so that for $n \geq n_0$, $|a_{n,k} - a_k| \leq \frac{\epsilon}{2k_0}$. Then

$$|\varphi_n(s) - \varphi(s)| \leq \sum_{k=1}^{k_0-1} |a_{n,k} - a_k|s^k + \sum_{k\geq k_0} |a_{n,k} - a_k|s^k \,.$$

By choice of k_0, the second term is smaller than $\epsilon/2$ and, for all $n \geq n_0$, the first term is smaller than $\epsilon/2$. Thus $|\varphi_n(s) - \varphi(s)| \leq \epsilon$ for all $n \geq n_0$, showing that $\varphi_n(s) \to \varphi(s)$ for each $s \in (0,1)$.

Conversely, suppose that $\varphi_n(s) \to \varphi(s)$ for each s with $0 < s < 1$. Clearly $0 \leq \varphi(s) \leq 1$ and $\varphi(s)$ is non-decreasing in s. In particular $\lim_{s\downarrow 0} \varphi(s) = a_0$ (say) exists. Further,

$$|a_{n,0} - a_0| \leq |a_{n,0} - \varphi_n(s)| + |\varphi_n(s) - \varphi(s)| + |\varphi(s) - a_0| \,, \tag{13}$$

and

$$|a_{n,0} - \varphi_n(s)| = \sum_1^\infty a_{n,k}s^k \leq \frac{s}{1-s} \,.$$

Therefore, given $\epsilon > 0$, we can choose s close enough to zero so that the first and third terms of the right side of (13) are each less than $\epsilon/3$. Now choose n large enough so that the second term is smaller than $\epsilon/3$. Thus we conclude that $a_{n,0} \to a_0$. Now note that

$$\frac{\varphi_n(s) - a_{n,0}}{s} \to \frac{\varphi(s) - a_0}{s} \quad \text{for } 0 < s < 1 \,.$$

It is easy to see that $g_n(s) = \dfrac{\varphi_n(s) - a_{n,0}}{s}$ is the generating function of the sequence $(a_{n,k+1})_{k\geq 0}$ so that by the same argument as above we can conclude that

$$\lim_{s\downarrow 0} \frac{\varphi(s) - a_0}{s} = a_1, \quad \text{say}$$

exists and moreover $\lim_n a_{n,1}$ exists and equals a_1. One can use induction to show that for each k, $\lim_{n\to\infty} a_{n,k} = a_k$ (say) exists.

Referring now to the *only if* part of the theorem we conclude that $\varphi_n(s)$ must converge, for each $s \in (0,1)$, to the generating function of the sequence (a_k), which therefore has to be the function $\varphi(s)$. This completes the proof of the theorem. ∎

The concept of a generating function as discussed above extends naturally to higher dimensions. We will briefly outline the definition and basic facts. Also for the sake of simplicity we confine ourselves to the case of multivariate probability generating functions.

Let $X_1, X_2, \ldots, X_d$ be random variables, defined on the same space, each taking non-negative integer values. Let their joint probability mass function be $p(k_1, k_2, \ldots, k_d)$. The *joint probability generating function (joint p.g.f.)* of $(X_1, X_2, \ldots, X_d)$ is the function φ defined on $[-1,1]^d$ defined by

$$\varphi(t_1, \ldots, t_d) = E(t_1^{X_1} \cdots t_d^{X_d}) = \sum_{k_1,\ldots,k_d} p(k_1, \ldots, k_d) t_1^{k_1} \cdots t_d^{k_d}. \tag{14}$$

It is not difficult to see that the series above converges absolutely. The function φ can also be shown to have partial derivatives of all orders and

$$p(k_1, \ldots, k_d) = \frac{1}{k_1! \cdots k_d!} \varphi^{(k_1,\ldots,k_d)}(0, \ldots, 0), \tag{15}$$

where $\varphi^{(k_1,\ldots,k_d)}$ denotes $D_1^{k_1} \cdots D_d^{k_d} \varphi$ with the usual notation that for $i = 1, \ldots, d$ and $k \geq 0$, D_i^k is the k-th order partial derivative with respect to the i-th variable. Thus for example, with $d = 3$,

$$\varphi^{(1,2,1)}(0,0,0) = \frac{\partial}{\partial t_1} \frac{\partial^2}{\partial t_2^2} \frac{\partial}{\partial t_3} \varphi(t_1, t_2, t_3)|_{(t_1=0, t_2=0, t_3=0)}.$$

Equation (15) shows that, as in the case of one dimension, the joint distribution of $(X_1, \ldots, X_d)$ is completely determined by the joint p.g.f. φ. Note that $\varphi(1, \ldots, 1) = 1$ by definition. One can also find all the moments, including cross-product moments of $(X_1, X_2, \ldots, X_d)$ from φ. For example,

$$E(X_1^2 X_2) = \varphi^{(2,1,0,\ldots,0)}(1, \ldots, 1) + \varphi^{(1,1,0,\ldots,0)}(1, \ldots, 1).$$

Also for any i, $1 \leq i \leq d$, $\varphi(t_1, \ldots, t_{i-1}, 1, t_{i+1}, \ldots, t_d)$ is precisely the joint p.g.f. of the random variables $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$.

In case $X_1, X_2, \ldots, X_d$ are independent with p.g.f.s $\varphi_1, \varphi_2, \ldots, \varphi_d$ respectively, then the joint p.g.f. of $(X_1, \ldots, X_d)$ is easily seen to be

$$\varphi(t_1, t_2, \ldots, t_d) = \varphi_1(t_1)\varphi_2(t_2)\cdots\varphi_d(t_d)\,. \tag{16}$$

In fact the condition (16) is also sufficient for independence. More generally, one can factor φ, the joint p.g.f of $(X_1, \ldots, X_d)$, as

$$\varphi(t_1, \ldots, t_d) = \widetilde{\varphi}(t_1, \ldots, t_i)\,\overline{\varphi}(t_{i+1}, \ldots, t_d)$$

if and only if $(X_1, \ldots, X_i)$ is independent of $(X_{i+1}, \ldots, X_d)$. Moreover, the functions $\widetilde{\varphi}$ and $\overline{\varphi}$ in the above factorization are the joint p.g.f.s of $(X_1, \ldots, X_i)$ and $(X_{i+1}, \ldots, X_d)$ respectively except possibly for some multiplicative constants. For example, the functions $3\widetilde{\varphi}$ and $\overline{\varphi}/3$ would also give a factorization.

The continuity theorem proved for one dimension has the following multivariate analogue.

Theorem: *For each $n \geq 1$, let φ_n be the joint p.g.f. of $(X_1^n, \ldots, X_d^n)$. In order that $\lim_{n\to\infty} P(X_1^n = k_1, \ldots, X_d^n = k_d)$ exists for all d-tuples $(k_1, \ldots, k_d)$ it is necesary and sufficient that for all $(t_1, \ldots, t_d) \in (0,1)^d$, the limit*

$$\lim_{n\to\infty} \varphi_n(t_1, \ldots, t_d) = \varphi(t_1, \ldots, t_d), \quad \textit{(say)}$$

exists. In this case, φ is actually the function

$$\varphi(t_1, \ldots, t_d) = \sum_{k_1,\ldots,k_d} a(k_1, \ldots, k_d) t_1^{k_1} \cdots t_d^{k_d},$$

where

$$a(k_1, \ldots, k_d) = \lim_{n\to\infty} P(X_1^n = k_1, \ldots, X_d^n = k_d)\,.$$

Barring complications arising out of d-dimensional variables, the idea of proof is no different from the one dimensional case. We omit the proof. In general the limit function $\varphi = \lim \varphi_n$ need not be a joint p.g.f.

Exercise 6. Show that the p.g.f. of the d-dimensional multinomial distribution with parameters $n, p_1, p_2, \ldots, p_d$ is $(p_1 t_1 + \cdots + p_d t_d)^n$.

Exercise 7. If for each $n \geq 1$, $(X_{n,1}, \ldots, X_{n,d})$ is multinomial with parameters $(n, p_{n1}, \ldots, p_{nd})$ and if $np_{ni} \to \lambda_i$ for $1 \leq i \leq d-1$, then show that $(X_{n,1}, \ldots, X_{n,d-1})$ has a limiting distribution as $n \to \infty$ and find the limiting distribution.

0.4 Continuous Random Variables

So far we have considered random variables with values in a finite or a countably infinite set. But in many applications it is necessary to go beyond that. For

example, consider picking a point at random from the interval $(0, 1]$. Here by picking a point at random we mean that any point "is as likely" to be picked as any other. The selected point X would then represent a random variable whose possible value can be any real number in $(0, 1]$. How do we describe the distribution of such a random variable? First of all, since any point is as likely to be picked as any other point, $P(X = x)$ should be the same for *all* x. Noting that there are infinitely many points x, one can easily argue that $P(X = x) = 0$ for all $x \in [0, 1]$. Thus, if we wanted to define the probability mass function p of the random variable X, the only candidate would be $p(x) = 0$ for all x. Certainly the distribution of the random variable X cannot be captured by such a function.

So, instead of prescribing probabilities of events through probabilities of individual outcomes that constitute an event, one may hope to prescribe probabilities of all events at one go. In other words, one may think of directly specifying $P(X \in A)$ for various subsets $A \subset [0, 1]$. But clearly, that is a tall task! However, there is a general theory — known as *measure theory* — which says that it is sufficient to specify $P(X \in A)$ only for intervals $A \subset [0, 1]$ which, in turn, uniquely determine $P(X \in A)$ for a large class of sets A, known as *measurable sets.* One may still wonder what if we want $P(X \in A)$ for a non-measurable set A. However, there is no real need to worry! The class of measurable sets is really huge — almost any set A one is likely to come across for the purpose of computing $P(X \in A)$ is going to be a measurable set. Having said all these let us add that mere recognition and acceptance of this fact will do for the rest of this book. We do not make any explicit use of measure theory.

Continuing with our example and again noting that the point is selected at random, one can easily deduce for any $0 \leq a < b \leq 1$, we must have $P(a < X < b) = b - a$. In fact, the above is just a consequence of the fact that $P(X \in I)$ equals $P(X \in J)$ whenever I and J are intervals of same length.

Of course, for any random variable X, prescribing the probabilities $P(X \in A)$ for intervals A and hence prescribing the distribution of X could also be done by simply specifying the function

$$F(x) = P(X \leq x) \tag{17}$$

for all $x \in R$. This function F is called the *probability distribution function* of the random variable X and has the following properties:

(i) $0 \leq F(x) \leq 1$ for all x and $F(x) \leq F(y)$ whenever $x \leq y$,
(ii) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$, and
(iii) F is right-continuous, that is, $\lim_{y \downarrow x} F(y) = F(x)$.

It may be noted that $\lim_{y \uparrow x} F(y) = P(X < x)$, so that $P(X = x) = F(x) - \lim_{y \uparrow x} F(y)$. From all these it should be clear that $F(x)$ determines $P(X \in A)$ for every interval A (and hence for all measurable sets A).

In the example of picking a point at random, the corresponding distribution function is

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases} .$$

A *continuous random variable* is one whose distribution function is continuous. From the properties of F listed above, it follows that a random variable X is continuous if and only if $P(X = x) = 0$ for all $x \in R$. It is in this sense that continuous random variables are diametrically opposite to discrete random variables.

0.4.1 Probability Density Function

One special class of continuous random variables are those for which the distribution function is given by

$$F(x) = \int_{-\infty}^{x} f(y)\, dy \tag{18}$$

where f is a non-negative funcion with $\int_{-\infty}^{\infty} f(y)\, dy = 1$. Such a function f is called a *probability density function* (p.d.f., in short). Probabilities involving X can be calculated from its density function by the formula $P(X \in A) = \int_A f(y)\, dy$. Such probability distributions are called *absolutely continuous* distributions and the corresponding random variable is also called absolutely continuous. It may be noted that probability density function of a distribution (or, of a random variable) is not unique. (Changing the value of f at a finite number of points would not change the integrals appearing in (18) and therefore, would give the same F!)

Unlike probability mass function, the probability density function does not represent any probability. However, it has the approximate interpretation

$$P(X \in (x, x + \delta x)) \sim f(x)\delta x\,.$$

This should explain why f is called the density function as opposed to mass function of the discrete case. For a random variable X with density function f the expected value is defined by the formula

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx\,, \tag{19}$$

provided the integral exists. We allow the integral to equal $+\infty$ or $-\infty$. But there is a caveat! Two infinities cannot be added unless they have the same sign. We define, more generally,

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx\,, \tag{20}$$

provided the integral exists. The expected value so defined can be shown to satisfy all the properties that were proved to be true for discrete random variables. As in the discrete case, the m-th moment of X is defined to be $E(X^m)$ and the variance is defined as $V(X) = E(X^2) - (EX)^2$.

Exercise 1. Fix numbers $a < b$. Let f be the function which is $1/(b-a)$ for points in the interval (a, b) and zero for points outside the interval. Show that this is a probability density function. Calculate the corresponding distribution function. This is called the *Uniform distribution* on (a, b), denoted $\mathcal{U}(a, b)$ and a random variable with this distribution is called a $\mathcal{U}(a, b)$ random variable. Find the expected value and variance of such a random variable.

Exercise 2. Fix any number $\lambda > 0$. Consider the function f which is zero for negative numbers and is $\lambda \exp(-\lambda x)$ for non-negative numbers x. Show that this is a probability density function. Calculate the corresponding distribution function. This is called the *Exponential distribution with parameter* λ, written $\mathcal{E}xp(\lambda)$. For a $\mathcal{E}xp(\lambda)$ random variable X, find (i) $P(X > 10.25)$, (ii) $P((X - 3)^2 > 1)$. Also find $E(X)$, $V(X)$ and $E(e^{tX})$ for $t \in R$.

Exercise 3. Fix any real number μ and any strictly positive number σ. Let

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

for $-\infty < x < +\infty$. This is a probability density function (not easy to show this fact). Corresponding distribution is called the *Normal distribution* with parameters μ and σ^2, written $\mathcal{N}(\mu, \sigma^2)$. The distribution function cannot be calculated explicitly. Show that a $\mathcal{N}(\mu, \sigma^2)$ random variable X has mean μ and variance σ^2. Also show that $E(e^{tX}) = \exp[\mu t + \frac{1}{2}\sigma^2 t^2]$ for $t \in R$.

In case $\mu = 0$, $\sigma = 1$ in Exercise 3 above, the distribution is called *Standard Normal Distribution.* In this case, the distribution function is usually denoted by $\Phi(x)$ and the density function is denoted by $\phi(x)$.

Exercise 4. Repeat Exercises 4–14 of Section 0.2.1, assuming that all the random variables are absolutely continuous.

0.4.2 Joint Density Function

For two continuous random variables X and Y defined on the same space, we may be interested in probabilities of events that concern both X and Y. For computing such probabilities, knowing the individual density functions of X and Y alone would not suffice. We need to know what is called the *joint density function* of X and Y.

Definition: Let X and Y be two random variables defined on the same space. The pair (X, Y) is said to have a *joint density function* $f(x, y)$ if f is a non-negative function such that for any $x, y \in R$,

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v)\, du dv.$$

Clearly such an f satisfies $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u,v)\,du\,dv = 1$. Probabilities involving the pair (X,Y) can be computed from the formula $P((X,Y)\in A) = \int\int_A f(x,y)dx\,dy$.

From the joint density of (X,Y) the individual (marginal) densities of X and Y can be recovered as follows:

$$f_1(x) = \int f(x,y)dy, \quad f_2(y) = \int f(x,y)dx.$$

In an analogous way the joint density of n random variables (defined on the same space) is defined to be a non-negative function f of n variables such that

$$P(X_1 \le x_1, \ldots, X_n \le x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(u_1, \ldots, u_n)\,du_1 \cdots du_n.$$

Here also the individual density of each X_i can be obtained from the joint density f by a formula analogous to the bivariate case. An important point to note in this connection is that the existence of a joint density for $(X_1, \ldots, X_n)$ implies that each X_i has a density; however the converse is not true. For example, if X has $\mathcal{U}(0,1)$ distribution and $Y = X$, then both X and Y have densities, but the pair (X,Y) does not have a joint density (why?).

For $(X_1, \ldots, X_n)$ with joint density f, the expected value of any function of $(X_1, \ldots, X_n)$ can be computed by the formula,

$$E(g(X_1, \ldots, X_n)) = \int \cdots \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) dx_1 \cdots dx_n,$$

provided, of course, that the integral exists.

For a pair (X,Y) with joint density f, the covariance between X and Y is defined by the same formula (21) of Section 0.2.2 and all the properties listed in Exercises 16–21 there remain valid.

0.4.3 Conditional Density

For a pair (X,Y) with joint density f the *conditional density* of X given $Y = y$ is defined to be

$$f_1(x \,|\, y) = \frac{f(x,y)}{f_2(y)} \quad \text{if} \quad f_2(y) > 0.$$

For y with $f_2(y) = 0$, one may define $f_1(x \,|\, y)$ to equal any density function, for example, one may put $f_1(x \,|\, y) = f_1(x)$. Here f_1 and f_2 are the marginal densities as defined in the previous section. One can easily check that $f_1(x \,|\, y)$ is a density (in x) for every y. The distribution given by this density is called the *conditional distribution* of X given $Y = y$. One can similarly define the conditional distribution of Y given $X = x$. It is also not difficult to extend this concept to the case of n random variables with a joint density.

For a random variable X with density f and for any event A with $P(A) > 0$, one can define the conditional density of X given A. However, there is no explicit formula for this density in general. One is only guaranteed of the existence of this density by a result known as *Radon-Nikodym Theorem* which is beyond the scope of this book. However, in the special case when $A = \{X \in B\}$ the conditional density of X given A is given by $f(x|A) = f(x)/P(A)$ if $x \in B$ and equals zero otherwise.

As in Section 0.2.3, expectation and variance of the conditional distribution are known as *conditional expectation* and *conditional variance* respectively. As before, it is sometimes convenient to think of conditional expectation itself as a random variable, denoted by $E(X|Y)$, which has the same interpretation as in Section 0.2.3. Also, all the properties in Exercises 4,5,7,8, 10–14 and 22–26 go through.

0.4.4 Independence

Definition: Random variables $(X_1, \ldots, X_n)$ with a joint density f are said to be *independent* if for any $x_1, x_2, \ldots, x_n$,

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n) .$$

Thus the random variables are independent if the joint density factors into product of the marginal probability densities $f_1, \ldots, f_n$. This can be shown to be equivalent to the condition that for sets $B_1, B_2, \ldots, B_n$, the n events $(X_i \in B_i)$, $1 \leq i \leq n$ are independent. Also independence of $X_1, X_2, \ldots, X_n$ clearly implies independence of $X_{j_1}, X_{j_2}, \ldots, X_{j_m}$ for $1 \leq j_1 < j_2 < \cdots < j_m \leq n$. With some work one can also show the following. Consider k indices $1 \leq i_1 < i_2 < \cdots < i_{k-1} \leq n$ and let $Y_1, Y_2, \ldots, Y_k$ be random variables defined as follows: $Y_1 = g_1(X_1, \ldots, X_{i_1}), Y_2 = g_2(X_{i_1+1}, \ldots, X_{i_2}), \cdots$, $Y_k = g_k(X_{i_{k-1}+1}, \ldots, X_n)$, for functions $g_1, g_2, \ldots, g_k$. Then independence of $X_1, X_2, \ldots, X_n$ implies that of $Y_1, Y_2, \ldots, Y_k$.

It is left as an exercise to verify that properties 28–33 of Section 0.2.4 remain valid here also.

0.5 Sequences of Random Variables

Let Y be a random variable with finite mean μ. Let $Y_1, Y_2, \ldots$ be independent observations on the variable Y, that is, for each n, the random variables $Y_1, \ldots, Y_n$ are independent each having the same distribution as Y. One says that $Y_1, Y_2, \ldots$ is a sequence of *independent and identically distributed*, abbreviated as i.i.d., random variables. Let X_n denote the average of the first n observations, that is, $X_n = (Y_1 + \cdots + Y_n)/n$. This X_n is also called the observed mean or the sample mean, based on n observations. An important question is : what happens to these observed means as n, the sample size, becomes large? A classical result in probability (known as the *law of large numbers*) is that

the observed means converge to the common population mean μ. It should be noted that the observed means X_n are random variables. Thus, one has to know what is meant by the convergence of a sequence of random variables. In this section, we discuss some of the various concepts of convergence that are used in probability.

In what follows, $(X_n)_{n\geq 1}$ will stand for a sequence of random variables defined on the same space.

Definition: We say that X_n *converges in probability* to a random variable X, and write $X_n \xrightarrow{P} X$, if for each $\epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.

That is, given any $\epsilon > 0$ however small, the chances that X_n deviates from X by more than ϵ become smaller and smaller as n gets larger and larger. However this should not be misinterpreted as X_n remaining close to X eventually for almost all sample points. One can construct an example where $X_n \xrightarrow{P} X$ but $X_n(\omega) \not\longrightarrow X(\omega)$ for any sample point ω! [see Exercise 5 below]. This motivates the next definition.

Definition: We say that X_n *converges with probability one* to X, and write $X_n \longrightarrow X$ w.p.1, if for all sample points ω, outside a set of zero probability, $X_n(\omega) \to X(\omega)$ as $n \to \infty$.

Another mode of convergence that will be useful for us, is the following.

Definition: Let $p \geq 1$ be a number. We say that X_n *converges in p-th mean* to X, written $X_n \xrightarrow{L_p} X$, if $E|X_n - X|^p \to 0$ as $n \to \infty$. This mode of convergence is also referred to as convergence in L_p.

We shall now see the relationships between these three modes of convergence. The last two modes of convergence are stronger than the first one. Indeed if $X_n \xrightarrow{p} X$ then for any $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^p > \epsilon^p) = E\left(I_{\{|X_n - X|^p > \epsilon^p\}}\right)$$

$$\leq E\left(\frac{|X_n - X|^p}{\epsilon^p} I_{\{|X_n - X|^p > \epsilon^p\}}\right) \leq \frac{E|X_n - X|^p}{\epsilon^p} \longrightarrow 0,$$

by hypothesis. Note that here ϵ is fixed and n becomes large.

Hidden in the above argument is the fact that for any random variable Z and any $\epsilon > 0$

$$P(|Z| \geq \epsilon) \leq \frac{E|Z|}{\epsilon}, \tag{21}$$

a useful inequality, known as *Markov's inequality*. This inequality can be easily proved using Exercise 5(a) of Section 0.2.1.

Next suppose that $X_n \longrightarrow X$ w.p.1; that is, there is a set A of probability zero such that for $\omega \notin A$, $X_n(\omega) \to X(\omega)$. Let $\epsilon > 0$. Then for any n,

$$\{|X_n(\omega) - X(\omega)| > \epsilon\} \quad \subset \quad \bigcup_{k>n} (|X_k(\omega) - X(\omega)| > \epsilon)$$

and the set on the right side decreases as n increases and the limiting set is contained in A (because for any ω in the limiting set $X_n(\omega) \not\longrightarrow X(\omega)$). Since $P(A) = 0$ it follows that

$$\lim_{n\to\infty} P(|X_n(\omega) - X(\omega)| > \epsilon) = 0\,.$$

Convergence with probability 1 and convergence in L_p are, in general, not comparable. However, here is a useful result.

$$\text{If } X_n \longrightarrow X \text{ w.p.1} \quad \text{and} \quad E(\sup_n |X_n|^p) < \infty, \qquad \text{then } X_n \xrightarrow{L_p} X\,.$$

Indeed one can replace $X_n \longrightarrow X$ w.p.1 by the weaker hypothesis $X_n \xrightarrow{P} X$. So we will assume only this. Denote the random variable $\sup_n |X_n|^p$ by Z. It is not difficult to see that $X_n \xrightarrow{P} X$ yields that $P(|X|^p \leq Z) = 1$. [Show that for any $\epsilon > 0$, $P(|X|^p > Z + \epsilon) = 0$]. Thus $|X_n - X|^p \leq 2^p Z$ Note that the hypothesis says that Z has finite expectation. Therefore given $\delta > 0$, we can choose $\lambda > 0$ so that $E(ZI_{(Z>\lambda)}) < 2^{-p}\delta/3$. We can also choose $\epsilon > 0$ so that $\epsilon^p < \delta/3$. Now choose n_0 such that for $n \geq n_0$ we have, $P(|X_n - X| \geq \epsilon) \leq \delta 2^{-p}/3\lambda$. Now for $n \geq n_0$,

$$\begin{aligned} E|X_n - X|^p &\leq E(|X_n - X|^p I_{|X_n - X| \leq \epsilon}) + E(|X_n - X|^p I_{|X_n - X| > \epsilon}) \\ &\leq \epsilon^p + 2^p E(Z I_{|X_n - X| > \epsilon}) \\ &\leq \epsilon^p + 2^p E(Z I_{Z \leq \lambda} I_{|X_n - X| > \epsilon}) + 2^p E(Z I_{Z > \lambda}) \\ &\leq \epsilon^p + 2^p \lambda P(|X_n - X| > \epsilon) + 2^p E(Z I_{Z > \lambda})\,. \end{aligned}$$

Each term on the right side is at most $\delta/3$, completing the proof.

The reader must have already realized that Lebesgue's Dominated Convergence Theorem as given in Exercise 8, Section 0.2.1, is just a special case of the above.

Exercise 1. If $X_n \xrightarrow{P} X$, then show that $X_n^{19} \xrightarrow{P} X^{19}$. More generally, if f is a continuous function, then $f(X_n) \xrightarrow{P} f(X)$. What if, convergence in probability is replaced by convergence in L_p or by convergence with probability one?

Exercise 2. If $X_n \xrightarrow{L_p} X$ then $X_n \xrightarrow{L_r} X$, for $1 \leq r \leq p$.

Exercise 3. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then show that $X_n + Y_n \xrightarrow{P} X + Y$ and $X_n Y_n \xrightarrow{P} XY$. What if $\xrightarrow{P}$ is replaced by $\xrightarrow{L_p}$?

Exercise 4. Let $X_n \xrightarrow{P} X$. Show that there is a subsequence (n_k) such that $X_{n_k} \longrightarrow X$ w.p.1. (Choose n_k so that $P(|X_{n_k} - X| > 2^{-k}) < 2^{-k}$.)

Exercise 5. Consider a $\mathcal{U}(0,1)$ variable X. Consider the following sequence of random variables: $Z_1 = I_{(X<1/2)}$; $Z_2 = I_{(X>1/2)}$; $Z_3 = I_{(X<1/4)}$; $Z_4 = I_{(1/4<X<1/2)}$; $Z_5 = I_{(1/2<X<3/4)}$; $Z_6 = I_{(3/4<X<1)}$; etc. It should be clear(?)

how the subsequent Z_n are defined. Show that Z_n does not converge with probability one, but converges in probability to zero.

In conclusion, let us go back to the convergence of observed means to the population mean. Classical Laws of Large Numbers say that convergence here takes place with probability one. In other words, if $Y_1, Y_2, \ldots$ are i.i.d with common finite mean μ, then $X_n = (Y_1 + \cdots + Y_n)/n \longrightarrow \mu$ w.p.1. In fact, this result remains valid even without the assumption of finiteness of the mean, as long as the Y_i are non-negative. The proof of this result is quite involved for presenting here. Instead, we show that convergence in probability holds. For this, let us further assume that $V(Y) < \infty$. In this case, by Markov inequality

$$P(|X_n - \mu| > \epsilon) = P(|X_n - \mu|^2 > \epsilon^2) \leq \frac{1}{\epsilon^2} E|X_n - \mu|^2 = \frac{1}{\epsilon^2} \frac{V(Y)}{n} \to 0\,.$$

In the above, we have used the fact that $V(Y_1 + \cdots + Y_n) = nV(Y)$ because of the i.i.d. hypothesis. It is possible to do away with the finite variance assumption, but the argument becomes a little more complicated. As a special case of the above, if X_n is $B(n,p)$ then $X_n/n \xrightarrow{P} p$.

Exercise 6. If $X_n \sim B(n,p)$ show that $\sum E(X_n - np)^4/n^4 < \infty$ and hence conclude, using Borel-Cantelli lemma that $X_n/n \longrightarrow p$ w.p.1.

Another important mode of convergence is *convergence in distribution.* Since we do not need it for our applications, we do not discuss it. However in Section 0.3, we had an illustration of this kind of convergence. To be specific, what was shown there is that if $X_n \sim B(n,p_n)$ where $np_n \to \lambda$ as $n \to \infty$, then X_n converges 'in distribution' to a $P(\lambda)$ random variable. A classical result in probability, involving the notion of convergence in distribution is what is known as *Central Limit Theorem.* Here is what it says. If $Y_1, Y_2, \ldots$ are i.i.d. random variables with mean μ and finite variance σ^2, then $X_n = (Y_1 + Y_2 + \cdots + Y_n)/\sqrt{n}$ converges in distribution to a $\mathcal{N}(\mu, \sigma^2)$ random variable, that is, for any real number a, $P(X_n \leq a) \longrightarrow \int_{-\infty}^{a} f(u)du$, where f is the density function given in Exercise 3 of Section 0.4.1.

0.6 Characterstic Functions

Definition: For any random variable X, the function $\varphi_X(t) = E(e^{itX})$ defined for $-\infty < t < +\infty$ is called the *characterstic function* of X.

To make sense of this definition, one needs to extend the notion of expectation to a complex-valued random variable. If $Z = U + iV$ where U and V are real random variables with finite expectations, one defines $E(Z)$ to be the complex number $E(U) + iE(V)$. With this definition, it is easy to see that the property

$$E(\alpha_1 Z_1 + \alpha_2 Z_2) = \alpha_1 E(Z_1) + \alpha_2 E(Z_2)$$

holds, where α_1, α_2 are complex numbers and Z_1, Z_2 are complex random variables. The property $|E(Z)| \leq E(|Z|)$ also holds where, as usual, for a complex number z, $|z| = \sqrt{(Re\, z)^2 + (Im\, z)^2}$. Here is a quick proof of the above inequality. It is easy to see that $E(Z) = \alpha |E(Z)|$ for some complex number α with $|\alpha| = 1$. Thus

$$|E(Z)| = \overline{\alpha} E(Z) = E(\overline{\alpha} Z) = E(Re\,(\overline{\alpha} Z)) \leq E(|\overline{\alpha} Z|) = E(|Z|)$$

One can use this to show that the Lebesgue's Dominated Convergence Theorem (see Section 0.2) holds for complex random variables as well.

Returning to characterstic functions it may be noted that φ_X is a complex-valued function of a real variable t, given by the formula

$$\varphi_X(t) = E(\cos(tX)) + i\, E(\sin(tX))\,. \tag{22}$$

Clearly, the real random variables $\cos(tX)$ and $\sin(tX)$ are bounded and hence have finite expectations for all t. From (22) it follows that

(1) $\varphi_X(0) = 1$ and $\varphi_{aX+b}(t) = e^{itb} \varphi_X(at)$.

(2) $\varphi_X(-t) = \overline{\varphi_X(t)} = \varphi_{-X}(t)$. In particular, $\varphi_X(t)$ is a real-valued function if X has a symmetric distribution, that is, X and $-X$ have the same distribution.

Using $|E(e^{itX})| \leq E(|e^{itX}|)$ one also gets

(3) $|\varphi_X(t)| \leq 1$ for all t.

For any real t and h,

$$|\varphi_X(t+h) - \varphi_X(t)| \leq E(|e^{itX}(e^{ihX} - 1)|) = E(|e^{ihX} - 1|)$$

and by the Dominated Convergence Theorem the last expression goes to zero as $h \to 0$. Thus we have proved

(4) $\varphi_X(t)$ is a continuous function — in fact, it is uniformly continuous.

One of the important features of the characterstic function of a random variable X is that the distribution of X is completely determined by its characterstic function φ_X. In other words, two random variables with different distributions cannot have the same characterstic function. We give below the formula, known as the *Inversion formula*, that determines the distribution function F of a random variable X from its characterstic function φ_X.

(5) For any two continuity points $a < b$ of F,

$$F(b) - F(a) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t)\, dt$$

(6) Moreover, if $\int |\varphi_X(t)|dt < \infty$, then the random variable X has a bounded continuous density function given by

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t) dt$$

The proof of (5) is somewhat involved and hence omitted here. Interested reader may consult Chung [2005]. Here is a sketch of a proof of (6). Using an analogue of the Dominated Convergence Theorem valid for general integrals, one can show that if $|\varphi_X|$ has finite integral then the inversion formula can be written as

$$\begin{aligned} F(b) - F(a) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_a^b e^{-itx} dx \right) \varphi_X(t) dt. \end{aligned}$$

Interchanging the order of integration now (which can again be justified in view of $\int |\varphi_X(t)|dt < \infty$), one gets

$$F(b) - F(a) = \int_a^b f(x)dx \quad \text{where} \quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t) dt.$$

Thus X has density $f(x)$ which is bounded because

$$|f(x)| \leq \frac{1}{2\pi} \int |\varphi_X(t)|dt < \infty.$$

Continuity of f follows from the Dominated Convergence Theorem alluded to above.

One consequence of the one-one correspondence between characterstic functions and distributions is that the converse of (2) holds. In other words

(7) $\varphi_X(t)$ is real-valued function if and only if X has a symmetric distribution, that is, X and $-X$ have the same distribution.

(8) Of course, for independent random variables X and Y, we have $\varphi_{X+Y} = \varphi_X \cdot \varphi_Y$.

It is easy to see that if $X \sim B(n,p)$ then $\varphi_X(t) = (q + pe^{it})^n$. Now if X and Y are independent random variables and $X \sim B(n,p)$ and $Y \sim B(m,p)$ then the characterstic function of $X + Y$ turns out to be

$$\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t) = (q + pe^{it})^{m+n}$$

from which we can immediately conclude that $X+Y$ must have $B(m+n,p)$ distribution.

Similarly one can show that if $X \sim \mathcal{N}(0,1)$, then $\varphi_X(t) = E(\cos tX) = e^{-t^2/2}$. From this one can deduce that if $X \sim \mathcal{N}(\mu,\sigma^2)$, then $\varphi_X(t) = e^{it\mu - \frac{1}{2}t^2\sigma^2}$. Therefore, if $X \sim \mathcal{N}(\mu;\sigma^2)$ and $Y \sim \mathcal{N}(\nu,\tau^2)$ are independent then by computing φ_{X+Y}, one can conclude that $X+Y$ is $\mathcal{N}(\mu+\nu,\sigma^2+\tau^2)$.

Characterstic functions can also be used to compute moments of the distribution, when they exist. Here is how the method works. Using the power series expansion $e^{itX} = \sum_{n=0}^{\infty} \frac{(itX)^n}{n!}$ and taking expectations, one gets $\varphi_X(t) = E(\sum \frac{(itX)^n}{n!})$. Assume for the time being that the expectation and the infinite sum could be interchanged. That would give

$$\varphi_X(t) = \sum_{n=0}^{\infty} i^n E(X^n) \frac{t^n}{n!},$$

that is, $\varphi_X(t)$ has a power series expansion in t in which, the coefficient of t^n is $\frac{i^n}{n!}E(X^n)$. From the general theory of power series it would follow that $E(X^n) = \varphi_X^{(n)}(0)/i^n$. It is possible to justify the above formal calculations (using simply Dominated Convergence Theorem and appropriate Mean Value Theorem) and here is the precise result.

(9) If X has finite n-th moment then φ_X has derivatives of orders upto and including n everywhere and, for every $k \leq n$, $E(X^k) = \varphi_X^{(k)}(0)/i^k$.

A very important use of characterstic functions consists of proving convergence in distribution. This is achieved through, what is known as, *Lèvy's Continuity Theorem.* The theorem asserts the equivalence of convergence in distribution and pointwise convergence of characterstic functions. Since we have not formally defined the notion of convergence in distribution, we would not go into the details of this result. The interested reader may consult Chung [2005].

Exercise 1. Calculate the characterstic functions of the following random variables: $P(\lambda)$, $\mathcal{U}[0,1]$, $\mathcal{E}xp(\lambda)$ and $\mathcal{N}(\mu,\sigma^2)$.

Exercise 2. If X has the *double exponential* density

$$f(x) = \frac{1}{2}\, e^{-|x|}, \qquad -\infty < x < \infty,$$

find the characterstic function of X. Use this and property (6) to find the characterstic function of *Cauchy* distribution given by the density

$$f(x) = \frac{1}{\pi}\, \frac{1}{1+x^2}, \qquad -\infty < x < \infty.$$

Exercise 3. Show that $\varphi_X(t_0) = 1$ for some $t_0 > 0$ if and only if X is discrete

with values in the set $\{2\pi n/t_0 : n = 0, \pm 1, \pm 2, \ldots\}$. More generally, show that $|\varphi_X(t_0)| = 1$ for some $t_0 > 0$ if and only if X is discrete with values in the set $\{(2\pi n + \theta)/t_0 : n = 0, \pm 1, \pm 2, \ldots\}$ for some real number θ.

Exercise 4. If φ is a characterstic function, show that both $|\varphi|^2$ and $Re\ \varphi$ are characterstic functions.

Exercise 5. Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables with common characterstic function φ. Let N be a $P(\lambda)$ random variable independent of the sequence (X_n). Find the characterstic function of $Y = X_1 + \cdots + X_N$. If $N = 0$, we define Y to be zero.

0.7 Martingales

In this section we discuss a special class of sequences of random variables known as *martingales*. Martingales constitute a very important and widely useful class of processes. We do not intend to present here an extensive coverage of this topic. Instead we only list a few basic properties of martingales which will be needed for our purposes. A reader interested to learn more can see the book of Leo Breiman.

Consider a sequence of independent tosses of a fair coin. Before each toss you are allowed a bet. If the toss results in heads then you win the amount you wagered; otherwise you loose the same amount. Note that you are allowed to change your wagers at each toss and moreover, your decision is allowed to be based on the outcomes of previous tosses. This can be mathematically formalized by means of a sequence $\epsilon_1, \epsilon_2, \ldots$ of random variables where ϵ_n denotes your wager amount for the n-th toss. If we denote the outcomes of the tosses themselves by a sequence $\eta_1, \eta_2, \ldots$, where each η_i can be $+1$ or -1, then the actual amount you win at the n-th toss is $\eta_n \epsilon_n$. Clearly η_n are i.i.d. random variables. The condition on the ϵ_n is that ϵ_1 is a constant and, for $n \geq 2$, ϵ_n is a random variable that is allowed to depend *only* on $\eta_1, \eta_2, \ldots, \eta_{n-1}$. As an additional technical condition, we shall also assume that each ϵ_n has finite expected value. One of the interesting features of the game is that if the coin is fair, then the game is also fair in the following sense. Denote by X_n, your accumulated fortune upto and including the n-th toss, that is, $X_n = \sum_{i=1}^{n} \epsilon_i \eta_i$. Define $X_0 = 0$. One can easily deduce that, at any stage if you want to find the conditional expectation of your accumulated fortune after the next toss, given all the information upto and including the present time, it equals your present accumulated fortune. That is to say that if you play one more game, it would, on the average, make you neither better off nor worse off. The word 'on an average' is important here, because in the actual play you would really either win or loose. The point is that you cannot be certain of either and the mean change in fortune, based on available information, is zero. This is the mathematical formulation of fairness in the game. This leads to the following formal definition.

All the random variables that we consider below are defined on the same space. Also they are all assumed to be discrete and to have finite expectations. Although the condition that they are discrete is not necessary in general, however it allows us to avoid some technicalities.

Definition: A sequence $(X_n)_{n\geq 0}$ of random variables is said to be a *martingale* if, for every n,

$$E(X_n \mid X_0, X_1, \ldots, X_{n-1}) = X_{n-1}. \tag{23}$$

In particular $E(X_n)$ is same for all n.

We will see plenty of examples of martingales in our applications in the subsequent chapters. However, here are some simple examples.

Example 1: Let $(\eta_i)_{i\geq 1}$ be a sequence of independent random variables with zero means. Set, $X_0 = 0$, and for $n \geq 1$, $X_n = \sum_{i=1}^{n} \eta_i$. Then $(X_n)_{n\geq 0}$ is easily seen to be a martingale. We could easily replace each X_n by $X_n + Z$ where Z is a random variable with finite mean, independent of $(\eta_i)_{i\geq 1}$, and still have a martingale.

Example 2: Let $(\eta_i)_{i\geq 1}$ and Z be as above. Let $(\epsilon_i)_{i\geq 1}$ be a sequence of bounded random variables with ϵ_n depending only on $\{Z, \eta_1, \ldots, \eta_{n-1}\}$ for each n. Set $X_n = Z + \sum_{i=1}^{n} \epsilon_i \eta_i$, $n \geq 0$. Then $(X_n)_{n\geq 0}$ is a martingale. The condition that ϵ_i are bounded is just to ensure that $\epsilon_i \eta_i$ has finite expectation and can be relaxed by the latter. The example given at the beginning of this section with the η_i representing the outcomes of succesive tosses of a coin is just a special case.

Example 3: Let $(\eta_i)_{i\geq 1}$ be as in Example 1, with the additional assumption that $V(\eta_i) = \sigma_i^2 < \infty$. Then $X_0 = 0$ and $X_n = (\sum_1^n \eta_i)^2 - \sum_1^n \sigma_i^2$, $n \geq 1$, defines a martingale. In particular, if each η_i takes the values ± 1 with probability $1/2$ each, then $(\sum_1^n \eta_i)^2 - n$ is a martingale.

Example 4: Here is the famous Polyà Urn Scheme. Start with an urn containing b black balls and r red balls. A ball is drawn at random, its colour noted and then the ball is replaced along with an additional ball of the same colour. This process is repeated. Note that, at each stage the total number of balls in the urn increases by one so that after n turns, the urn will have $b + r + n$ balls. Denoting X_n to be the proportion of red balls in the urn after n turns, with, of course, $X_0 = r/(b + r)$, it is not difficult to check that we get a martingale.

Example 5: Let $(\eta_i)_{i\geq 1}$ be an i.i.d sequence, taking the values ± 1 with probabilities $1/2$ each. Denote $S_n = \sum_{1\leq i\leq n} \eta_i$. Then for any $\theta \in (0, 1)$, the sequence $(X_n)_{n\geq 0}$ defined as $X_0 \equiv 1$ and for $n \geq 1$, $X_n = 2^n \theta^{(n+S_n)/2} (1-\theta)^{(n-S_n)/2}$, defines a martingale. Indeed this is a special case of the next Example.

Example 6: Let $(\eta_i)_{i\geq 1}$ be a sequence of discrete random variables and for each

$n \geq 1$, let $p_n(u_1, \ldots, u_n)$ be the (true) joint p.m.f. of $(\eta_1, \ldots, \eta_n)$. Suppose that $\widetilde{p}_n(u_1, \ldots, u_n)$, $n \geq 1$, be a sequence of joint p.m.f.s satisfying

(i) $\sum_{u_{n+1}} \widetilde{p}_{n+1}(u_1, \ldots, u_n, u_{n+1}) = \widetilde{p}_n(u_1, \ldots, u_n)$ and

(ii) $\widetilde{p}_n(u_1, \ldots, u_n) = 0$ whenever $p_n(u_1, \ldots, u_n) = 0$.

Then $X_0 = 1$ and for $n \geq 1$, $X_n = \widetilde{p}_n(\eta_1, \ldots, \eta_n)/p_n(\eta_1, \ldots, \eta_n)$ can be seen to define a martingale (taking the ratio 0/0 to be 0). Complicated though this example looks, here is the context in which it arises. The $\widetilde{p}_n$ can be thought of as the joint p.m.f.s under some proposed alternative distribution of the sequence $(\eta_i)_{i \geq 1}$. A statistician wants to test the validity of this alternative. Standard tools of statistics often use the X_n (known as *likelihood ratio* in statistical parlance) to test such hypotheses.

Most of the basic theory of martingales is due to J. L. Doob. We proceed to present some of the basic results on martingales, which we need in the sequel. The first result is about convergence with probability one for a martingale. One of the main tools for the proof of this result is an inequality known as *Doob's upcrossing inequality.*

Let $x_0, x_1, \ldots, x_n$ be a finite sequence of real numbers. For $a < b$, let $u_n(a, b)$ denote the number of 'upcrossings' of the interval (a, b) by the sequence. For example, suppose $n = 7$ and $x_0 \leq a, a < x_1 < b, x_2 \geq b, x_3 \geq b, a < x_4 < b, x_5 \leq a, x_6 \leq a$ and $x_7 \geq b$. Then there are exactly two upcrossings. More generally, $u_n(a, b) = k$ if there exist exactly k pairs (and no more) of indices $0 \leq m_1 < n_1 < \cdots < m_k < n_k \leq n$ such that $x_{m_i} \leq a$ and $x_{n_i} \geq b$ for $i = 1, \ldots, k$. Here is a convenient formula for counting the number of upcrossings. Define $v_0 \equiv 1$ and for $0 \leq i \leq n$, $v_{i+1} = 0$ or v_i or 1 according as $x_i \leq a$ or $a < x_i < b$ or $x_i \geq b$. It is then easy to see that $u_n(a, b) = \sum_{i=1}^{n} (v_{i+1} - v_i)^+$. The reader can easily verify the inequality $(b-a)(v_{i+1} - v_i)^+ \leq (x_i - a)(v_{i+1} - v_i)$ for $i = 1, \ldots, n$. This immediately gives $(b-a)u_n(a, b) \leq \sum_{i=1}^{n} (x_i - a)(v_{i+1} - v_i)$. In the above, we have used the notation c^+ to denote $\max\{c, 0\}$ — known as 'the positive part' of a real number c.

Suppose now $X_0, X_1, \ldots, X_n$ are random variables and denote the corresponding number of upcrossings by $U_n(a, b)$, which is also a random variable now. Further the corresponding v_i are now denoted by V_i. For each $i \geq 1$, V_i is also a random variable and depends only on $X_0, \ldots, X_{i-1}$. From the above inequality it follows that, if the X_i have finite means, then $(b-a)E[U_n(a, b)] \leq \sum_{i=1}^{n} E[(X_i - a)(V_{i+1} - V_i)] = \sum_{i=1}^{n} E[(X_i - a)V_{i+1}] - \sum_{i=1}^{n} E[(X_i - a)V_i]$.

Assume now that we have a martingale $(X_i)_{i \geq 0}$ and apply the above to the random variables $X_0, \ldots, X_n$. For each $i = 1, \ldots, n$, we have

$$\begin{aligned} E[(X_i - a)V_i] &= E[E\{(X_i - a)V_i \,|\, X_0, \ldots, X_{i-1}\}] \\ &= E[V_i\, E\{(X_i - a) \,|\, X_0, \ldots, X_{i-1}\}] = E[(X_{i-1} - a)V_i]. \end{aligned}$$

In the above we have used properties of conditional expectations stated in Exercises 22 and 23 of Section 0.2.3 and the martingale property. Using this we get

$$\begin{aligned}(b-a)E[U_n(a,b)] &\leq \sum_{i=1}^{n} E[(X_i - a)V_{i+1}] - \sum_{i=1}^{n} E[(X_{i-1} - a)V_i] \\ &= E[(X_n - a)V_{n+1}] - E[(X_0 - a)V_1].\end{aligned}$$

Since the second term in the final expression is easily seen to be non-negative and the first term is $\leq E(|X_n - a|) \leq E|X_n| + |a|$, we have proved

Theorem (Doob's Upcrossing Inequality): *For any martingale $(X_i)_{i\geq 0}$ and for any $a < b$, $E[U_n(a,b)] \leq (E|X_n| + |a|)/(b-a)$, for all n.*

We are now ready to prove the convergence theorem known as *Doob's Martingale Convergence Theorem.* We need the following simple observation whose proof is left as an exercise. Given any sequence $(x_i)_{i\geq 0}$ of real numbers, the sequence converges if and only if for every pair of rational numbers $a < b$, $u(a,b) \stackrel{\text{def}}{=} \sup_n u_n(a,b) < \infty$. Here by the convergence of a sequence we mean that it either converges to a real number *or* diverges to $+\infty$ *or* diverges to $-\infty$.

Suppose now that $(X_i)_{i\geq 0}$ is a martingale and let $a < b$ be a pair of rational numbers. Consider the sequence of random variables $(U_n(a,b))_{n\geq 1}$ as defined earlier. Clearly this is a non-decreasing sequence of random variables taking non-negative integer values. Thus $U(a,b) = \lim_{n\to\infty} U_n(a,b)$ is well defined (possibly taking the value $+\infty$). Moreover by the Monotone Convergence Theorem, $E[U(a,b)] = \lim_{n\to\infty} E[U_n(a,b)] \leq (\sup_n E|X_n| + |a|)/(b-a)$, where the last inequality uses the upcrossing inequality. Therefore if the martingale $(X_i)_{i\geq 0}$ is assumed to satisfy the condition $\sup_n E|X_n| < \infty$, we will get $E[U(a,b)] < \infty$. This will of course imply $P[U(a,b) = +\infty] = 0$; equivalently, $P[U(a,b) < \infty] = 1$. Since this is true for every pair of rational numbers $a < b$ (and there are only countably many such pairs), we have $P[U(a,b) < \infty$ for every pair of rationals $a < b] = 1$. But this will imply by the earlier observation that $P\{\, X_i \text{ converges } \} = 1$. Further, denoting $Z = \liminf_i |X_i|$, an easy application of the Monotone Convergence Theorem (and the definition of lim inf) gives $E(Z) \leq \liminf_i E|X_i| \leq \sup_n E|X_n| < \infty$. This would imply that if X_i converges to X with probability 1, then X has finite mean. We have thus proved

Theorem (Doob's Martingale Convergence Theorem): *If $(X_n)_{n\geq 0}$ is a martingale with $\sup_n E|X_n| < \infty$, then X_n converges with probability 1 to a random variable X which has finite expectation.*

We are now going to prove that if moreover, $\sup_n E|X_n|^2 < \infty$ then the convergence takes place also in L_2, that is, $E(X^2) < \infty$ and $E(X_n - X)^2 \to 0$

as $n \to \infty$. An immediate consequence of this, which will be used by us, is the following: if $(X_n)_{n\geq 0}$ is a martingale such that $|X_n| \leq c$ for all n, where c is a finite constant, then X_n converges to X with probability 1 as well as in L_2 (hence in L^1) and, in particular, $EX = EX_0$.

For the proof, we first need the following basic result on the expectation of non-negative random variables.

Lemma: *For any non-negative random variable X with finite expectation, one has* $E(X) = \int_0^\infty P(X > \lambda)\, d\lambda$.

Proof: In case X has a density, say $f(x)$, then by an interchange of integrals,

$$\int_0^\infty P(X > \lambda)\, d\lambda = \int_0^\infty \int_\lambda^\infty f(x)dx\, d\lambda = \int_0^\infty f(x) \int_0^x d\lambda\, dx = \int_0^\infty xf(x)dx\,.$$

which equals $E(X)$ as stated. Next let us consider a discrete random variable X taking finitely many values say, $x_1 < \cdots < x_k$ with probabilities $p_1, \ldots, p_k$ respectively. In this case, it is easy to see that

$$\begin{aligned}\int_0^\infty P(X > \lambda)d\lambda = \quad & x_1 + (x_2 - x_1)(1 - p_1) + (x_3 - x_2)(1 - p_1 - p_2) + \\ & \cdots + (x_n - x_{n-1})p_n,\end{aligned}$$

and the right hand side clearly simplifies to $\sum x_i p_i = E(X)$.
If X is a non-negative discrete random variable taking an infinite number of values, say $x_1, x_2, \ldots$, then we can define a sequence (X_n) of non-negative random variables increasing to X with each X_n taking only finitely many values. To be precise, for each n, X_n is defined to be equal to X whenever X takes values from $\{x_1, \ldots, x_n\}$, and is defined to be zero otherwise. An application of the earlier case and Monotone Convergence Theorem completes the proof. ∎

Suppose that $p \geq 1$ and $E(|X|^p) < \infty$. Then by the above Lemma applied to the non-negative random variable $|X|^p$, we get

$$E(|X|^p) = \int_0^\infty P(|X|^p > \lambda)d\lambda = \int_0^\infty P(|X| > \lambda^{1/p})d\lambda.$$

An easy change of variable now leads to

Corollary: For any $p \geq 1$ and any random variable X with $E(|X|^p) < \infty$, $E(|X|^p) = p \int_0^\infty \lambda^{p-1} P(|X| > \lambda)d\lambda$.

Now let $(X_n)_{n\geq 0}$ be a martingale. For each n, let $M_n = \max_{i\leq n} |X_i|$. Fix $\lambda > 0$ and consider the event $A = \{M_n > \lambda\}$. An upper bound for $P(A)$ is

provided by what is known as *Doob's Maximal Inequality* given below.

Lemma (Doob's Maximal Inequality): $P(A) \leq E(|X_n| \cdot I_A)/\lambda$.

Proof: Let $A_0 = \{|X_0| > \lambda\}$ and $A_i = \{|X_0| \leq \lambda, \ldots, |X_{i-1}| \leq \lambda, |X_i| > \lambda\}$, for $1 \leq i \leq n$. Then $A_0, \ldots, A_n$ are disjoint and $A = \bigcup_i A_i$, so that, $P(A) = \sum P(A_i) \leq \frac{1}{\lambda} \sum E(|X_i| I_{A_i})$. An easy consequence of the martingale property of (X_n) is that, for any $i \leq n$, $X_i = E(X_n \,|\, X_0, \ldots, X_i)$. This would therefore give

$$\begin{aligned} P(A) &\leq \frac{1}{\lambda} \sum E(|E(X_n \,|\, X_0, \ldots, X_i)| I_{A_i}) \\ &\leq \frac{1}{\lambda} \sum E(E(|X_n| \,|\, X_0, \ldots, X_i) I_{A_i}) \\ &= \frac{1}{\lambda} \sum E(E(|X_n| I_{A_i} \,|\, X_0, \ldots, X_i)) \\ &= \frac{1}{\lambda} \sum E(|X_n| I_{A_i}) = \frac{1}{\lambda} E(|X_n| I_A). \end{aligned}$$

∎

Applying now the Corollary above with $p = 2$ and $X = M_n$ followed by the Lemma, we get

$$\begin{aligned} E(M_n^2) &\leq 2 \int_0^\infty E(|X_n| I_{\{M_n > \lambda\}})\, d\lambda = 2E\left(|X_n| \int_0^\infty I_{\{M_n > \lambda\}}\, d\lambda\right) \\ &= 2E(|X_n| M_n) \leq 2\sqrt{E(X_n^2)}\sqrt{E(M_n^2)}. \end{aligned}$$

This leads to $E(M_n^2) \leq 4E(X_n^2)$. If now (X_n) is an L_2-bounded martingale, that is, $\sup_n E(X_n^2) < \infty$, then it follows that

$$E\left(\sup_n |X_n|^2\right) \leq 4 \sup_n E(X_n^2) < \infty\,,$$

whence by Dominated Convergence Theorem we get

Theorem: *If $(X_n)_{n \geq 0}$ is an L_2-bounded martingale, then it converges with probability one to a random variable X having finite second moment and moreover, the convergence is also in L_2, that is, $\lim_n E(X_n - X)^2 = 0$.*

0.8 Markov Chains

Consider a system which evolves with time in a random way. For the sake of simplicity, let us consider the set of times to be discrete. Let us also assume that the set S of possible states of the system is countable. S will be called the *state space* of the system and the individual states (i.e. elements of S) will be denoted by i, j, k etc. Let X_n denote the (random) state of the system at time n, $n = 0, 1, 2, \ldots$. We are assuming that each X_n is an S-valued random variable and the entire evolution of the system is described by the sequence $(X_n)_{n \geq 0}$. In particular, X_0 is the *initial state* of the system. Study of such systems in this generality, without any further assumptions, will not lead to any interesting

theory. Usually one imposes additional restrictions on the joint distribution of the sequence to get different kinds of "stochastic processes". One such condition, studied in the last section, is that $E(X_n | X_0, X_1, \ldots, X_{n-1}) = X_{n-1}$ for all n, which gave rise to what are called martingales. In this section, we study one other extremely important and useful condition that leads to a class of processes known as *Markov chains.*

0.8.1 Markov Chains: Transition Probabilities

The property that is imposed can be briefly referred to as 'lack of memory', by which we mean that given the present state of the system the 'future' evolution does not depend on the 'past' history. In mathematical terms, this means that for any two non-negative integers n and m, the conditional distribution of X_{n+m} given $X_0, X_1, \ldots, X_n$ depends only on X_n. It turns out that one needs only to assume this for $m = 1$. Thus we have the following definition.

Definition: A sequence $(X_n)_{n \geq 0}$ of random variables taking values in a countable set S is acIled a *Markov chain* on the state space S if, for any $n \geq 0$ and any $i_0, i_1, \ldots, i_{n-1}, i, j \in S$,

$$P(X_{n+1} = j \,|\, X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j \,|\, X_n = i)\,.$$

Further, the chain is called *time homogeneous*, if the above conditional probabilities do not depend on n and hence are also equal to $P(X_1 = j \,|\, X_0 = i)$.

Markov chains appearing in most applications also happen to be time homogeneous. A rich theory exists for such chains. We shall restrict ourselves to only Markov chains which are time homogeneous. We will denote, for any $i, j \in S$, the probability $P(X_1 = j \,|\, X_0 = i)$ by p_{ij}. Writing the distribution of X_0 as $\{\mu_i, i \in S\}$, it is not difficult to see that all the finite dimensional joint distributions for $(X_n)_{n \geq 0}$ are completely determined by the quantities μ_i, $i \in S$ and p_{ij}, $i, j \in S$. Specifically, for any n and any collection of states $i_0, i_1, \ldots, i_n$,

$$P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \mu_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}\,.$$

$\{\mu_i, i \in S\}$ is called the *initial distribution* and $\{p_{ij}, i, j \in S\}$ are called the *transition probabilities* or *one-step transition probabilities*, to be exact. From the definition it is clear that,

(1) $\mu_i \geq 0$ for all i and $\sum_i \mu_i = 1$,

(2) $p_{ij} \geq 0$ for all $i, j \in S$ and $\sum_j p_{ij} = 1$ for all $i \in S$.

It is often convenient to represent the initial distribution by a row vector $\mu = (\mu_i; i \in S)$ and the transition probabilities by the *transition matrix* $P = ((p_{ij}))_{i,j \in S}$. The property (2) above simply says that P has non-negative entries with each row sum equal to one. Such matrices are called *stochastic matrices.* Much of the theory of Markov chains rests on an analysis of its transition

matrix P and not so much on μ. The matrix P^n, the n-th power (in the sense of matrix multiplication) of the matrix P, is called the *n-step transition matrix* simply because its (i,j)-th element $p_{ij}^{(n)}$ gives the probability of transition from i to j in n steps, that is,

$$p_{ij}^{(n)} = P(X_{n+m} = j \,|\, X_m = i) = P(X_n = j \,|\, X_0 = i)\,.$$

One can easily verify this for $n = 2$ and then use induction to complete the proof.

One useful consequence, known as the *Chapman-Kolmogorov equations*, is

$$p_{ij}^{(m+n)} = \sum_k p_{ik}^{(m)} p_{kj}^{(n)}\,. \tag{24}$$

This can of course be verified directly from Markov property. For the sake of completeness we need to set $P^{(0)} = I$, the identity matrix which is also consistent with the notion of zero-step transition.

A simple yet useful property of a Markov chain $(X_n)_{n\geq 0}$ is that if f is any real function on the state space S satisfying $\sum_j p_{ij} f(j) = f(i)$ for all i, then the sequence $(f(X_n))_{n\geq 0}$ is a martingale, provided, of course, the sum $\sum f(i)\mu_i$ is convergent. Functions f satisfying $\sum_j p_{ij} f(j) = f(i)$ for all i, are known as *harmonic functions* for the transition matrix P.

Example 1: Let $\xi_1, \xi_2, \ldots$ be i.i.d. integer-valued random variables with common distribution $P(\xi_1 = j) = \alpha_j$. Let X_0 be any integer valued random variable independent of the ξ sequence. For $n \geq 1$, let $X_n = X_0 + \sum_{l=1}^{n} \xi_l$. Then (X_n) is a Markov chain with state space $S =$ the set of integers and transition probabilities $p_{ij} = \alpha_{j-i}$. The n-step transition probabilities are also not difficult to get. An elegant formula for these can be obtained using the following notation. For any two probability distributions $\alpha = (\alpha_j)$ and $\beta = (\beta_j)$ on integers, let $\alpha * \beta$ denote the distribution defined by $(\alpha * \beta)_j = \sum_i \alpha_i \beta_{j-i}$. In particular, α^{*n} is defined recursively by $\alpha^{*n} = \alpha^{*(n-1)} * \alpha$. With this notation, the n-step transition probabilities of the chain (X_n) above are given by $p_{ij}^{(n)} = \alpha_{(j-i)}^{*n}$.

It may be noted that here p_{ij} as well as $p_{ij}^{(n)}$ depend on i and j only through $j - i$. In fact, these are the only Markov chains with this property. In other words, if $(X_n)_{n\geq 0}$ is a Markov chain whose transition probabilities p_{ij} depend only on $i - j$, then the random variables $\xi_n = X_n - X_{n-1}$, $n \geq 1$, are i.i.d. The proof is easy. Such Markov chains are called *random walks*. A special case is when the random variables ξ_j take only two values $+1$ and -1. This gives what is known as *simple random walk*. A further special case when $P(\xi_j = +1) = p(\xi_j = -1) = 1/2$, is called *simple symmetric random walk*. Feller's book (vol.1) gives an extensive and interesting account of such random walks. His analysis is based entirely on what is known as 'path counting' and

is therefore easily accessible. Interested reader should consult this book. We will include parts of this material at the end of this section.

Example 2: Consider an urn with a total of D tokens — some numbered $+1$ and some -1. The composition of the urn changes over time as follows. At each turn, a token is picked at random from the urn and its sign changed and put back in the urn. Denote by X_n the number of $+1$ at time n. It is clear that X_n is a Markov chain with state space $S = \{0, 1, \ldots, D\}$ and transition probabilities

$$p_{i,i+1} = 1 - \frac{i}{D} = 1 - p_{i,i-1} .$$

Thus, from a state i, transitions are possible only to states $i-1$ or $i+1$ in one step. Of course if $i = 0$ (respectively D) then the system moves surely to 1 (respectively $D-1$). It is not difficult to see that the two-step transition probabilities are given by:

$$p^{(2)}_{i,i+2} = (1 - \frac{i}{D})(1 - \frac{i+1}{D}), \qquad p^{(2)}_{i,i-2} = \frac{i}{D}\frac{i-1}{D},$$

$$p^{(2)}_{ii} = 1 - \frac{i(i-1)}{D^2} - \frac{(D-i)(D-i-1)}{D^2} .$$

Exercise 1. Suppose X_0 is uniformly distributed on the state space S in the above example. Calculate the distributions of X_1, X_2 and also the joint distribution of (X_1, X_2). Do the same when $X_0 \sim B(D, 1/2)$.

Example 3 ($0-1$ chain): Consider a machine which can be in two states, 'on' and 'off'. Also if the machine is 'on' today, then the probability is α that it will be 'off' tomorrow. Similarly, β is the probability of transition from 'off' state to 'on' state in one day. Denote the 'on' and 'off' states by 0 and 1 respectively. Denoting by X_n the state of the machine on day n, $(X_n)_{n\geq 0}$ is a Markov chain with state space $S = \{0, 1\}$ and transition matrix

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} .$$

We assume that $\alpha + \beta > 0$ (what happens if $\alpha + \beta = 0$?). A trite calculation gives

$$p^{(n)}_{01} = p^{(n-1)}_{00}\alpha + p^{(n-1)}_{01}(1-\beta) = \alpha + p^{(n-1)}_{01}(1-\alpha-\beta) ,$$

from which one deduces

$$p^{(n)}_{01} = \frac{\alpha}{\alpha+\beta}[1 - (1-\alpha-\beta)^n] .$$

Similar calculation shows that

$$p^{(n)}_{10} = \frac{\beta}{\alpha+\beta}[1 - (1-\alpha-\beta)^n] .$$

If one further assumes that $\alpha + \beta < 2$ (what happens if $\alpha + \beta = 2$?), so that $|1 - \alpha - \beta| < 1$, then one gets $p_{01}^{(n)} \to \frac{\alpha}{\alpha+\beta}$ and $p_{10}^{(n)} \to \frac{\beta}{\alpha+\beta}$. Since $p_{00}^{(n)} = 1 - p_{01}^{(n)}$ we deduce that $p_{00}^{(n)} \to \frac{\beta}{\alpha+\beta}$. Similarly $p_{11}^{(n)} \to \frac{\alpha}{\alpha+\beta}$.

From all these, one can deduce that, if $0 < \alpha + \beta < 2$, then the Markov chain has the limiting distribution $\pi = \langle \pi_0, \pi_1 \rangle = \langle \frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta} \rangle$, whatever be the initial distribution of the chain. This distribution represents what is called the 'equilibrium' or 'steady-state' distribution of the chain. It is so called because π also happens to be the only distribution on S with the property that if X_0 has distribution π, then for every n, X_n has the same distribution π. This conclusion also follows from the equations describing $p_{ij}^{(n)}$.

This example leads one to the natural question. Is the above kind of phenomenon true for all Markov chains? That is, does every Markov chain have a limiting distribution? Is the limiting distribution unique (i.e. independent of initial conditions)? Are these limiting distributions, if any, also steady-state distributions in the above sense? The cases $\alpha + \beta = 0$ and $\alpha + \beta = 2$ should convince the reader that the answers to the above questions cannot always be in the affirmative.

To better understand the nature of problems and to identify some of the cases where we do have affirmative answer, we first need to discuss 'classification' of the states. This will be done in the next section. But let us now make a little digression to discuss some interesting properties of simple random walk as described in Example 1.

Simple Random Walk: Recall simple random walk as discussed in Example 1. Here is an illustration of the path-counting argument and the final result will be used in Chapter 4.

In the context of random walks, a path from $(0,0)$ to (n,a) is a polygonal line with vertices (i, s_i) for $0 \leq i \leq n$, with $s_0 = 0$, $s_n = a$ and $s_i - s_{i-1} = \pm 1$ for $1 \leq i \leq n$. It is clear that a has to be an integer between $-n$ and $+n$. Similarly one can talk about paths from (m,a) to $(m+n,b)$, for integers m and n with $n > 0$. Such paths are called paths of length n. It is clear that the total number of paths of length n starting from a given point is 2^n. Also the number of paths from $(0,0)$ to (n,a) is $\binom{n}{\frac{n+a}{2}}$.

A fact, often called the *reflection principle* is that, for integers $a, b > 0$, the number of paths from $(0,a)$ to (n,b) that touch or cross the X-axis is the same as the total number of paths from $(0,-a)$ to (n,b) and hence equals $\binom{n}{\frac{n+a+b}{2}}$. This is done by establishing a one-one correspondence between the two sets of paths. From this one can easily deduce that for any integer $a > 0$, the number of paths from $(0,0)$ to (n,a) that do not hit the X-axis equals $\binom{n-1}{\frac{n+a}{2}-1} - \binom{n-1}{\frac{n+a}{2}}$. Incidentally, this formula also gives a solution to what

is known as the *ballot problem.*

We now turn to an important property of simple, but not necessarily symmetric, random walk. In other words, we consider the Markov chain on the set of integers with transition probabilities $p_{i,i+1} = p = 1 - p_{i,i-1}$. For any integer a, let $T_a = \inf\{n \geq 1 : X_n = a\}$, that is, T_a is the hitting time of a. Fix integers $a < i < b$. Let

$$\varphi(i) = P(T_a < T_b \,|\, X_0 = i) \quad \text{for} \quad a < i < b$$

and $\varphi(a) = 1$ and $\varphi(b) = 0$.

Exercise 2.

(i) Show that for $a < i < b$, $\varphi(i) = p\varphi(i+1) + (1-p)\varphi(i-1)$

(ii) Denoting $d_i = \varphi(i) - \varphi(i-1)$, show that for $a + 1 < i \leq b$,

$$pd_i = (1-p)d_{i-1},$$

and hence that

$$d_i = \left(\frac{1-p}{p}\right)^{i-a-1} d_{a+1}\,.$$

(iii) Assume that $p = 1/2$ and show that $\varphi(i) = 1 + (i-a)d_{a+1}$ for $a \leq i \leq b$, and hence deduce that $\varphi(i) = (b-i)/(b-a)$.

(iv) Assume that $p \neq 1/2$ and denote $(1-p)/p$ by α. Show that

$$\varphi(i) = 1 + \frac{p}{1-2p}\, d_{a+1}\, \left[\alpha^{i-a} - 1\right]$$

for $a \leq i \leq b$ and hence deduce that

$$\varphi(i) = \frac{\alpha^b - \alpha^i}{\alpha^b - \alpha^a}\,.$$

(v) If $p > 1/2$ then show that

$$P(T_a < \infty \,|\, X_0 = i) \quad = \quad \alpha^{i-a} < 1\,.$$

(vi) If $p \leq 1/2$ then show that

$$P(T_a < \infty \,|\, X_0 = i) \quad = \quad 1\,.$$

0.8.2 Classification of States: Recurrence and Transience

For any state i, let us define two random variables, possibly taking value $+\infty$, as follows:

$$T_i = \min(n \geq 1 : X_n = i), \qquad N_i = \#(n \geq 1 : X_n = i)\,.$$

In the definition of T_i, if there is no $n \geq 1$ such that $X_n = i$ we take $T_i = +\infty$. If there are infinitely many such n, then of course, $N_i = +\infty$. T_i represents the time of the first visit to i and N_i represents the total number of visits to i. In both of these, the time point 0 is not counted. It is clear that the events $(T_i < \infty)$ and $(N_i \geq 1)$ are same.

For i and j in S, let

$$f_{ij}^{(n)} = P(T_j = n \,|\, X_0 = i) = P(X_n = j; X_l \neq j, 1 \leq l < n \,|\, X_0 = i), \tag{25}$$

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)} = P(T_j < \infty \,|\, X_0 = i)\,. \tag{26}$$

It is clear from the definitions that $p_{ij}^{(n)} \geq f_{ij}^{(n)}$. In fact one has the following identity often known as the *renewal equation* :

$$p_{ij}^{(n)} \quad = \quad \sum_{m=1}^{n} f_{ij}^{(m)} p_{jj}^{(n-m)}\,. \tag{27}$$

To prove the equation, one has to simply write the event $(X_n = j)$ as the disjoint union of events $\bigcup_{m=1}^{n} (T_j = m, X_n = j)$ and calculate the probabilities of these events by applying the Markov property.

All the states of the Markov chain are classified into two kinds as defined below.

Definition: A state i is called *recurrent* if $f_{ii} = 1$ and is called *transient* otherwise. A state i is called *absorbing* if $p_{ii} = 1$.

Thus a state i is recurrent if the chain starting from i is sure to return to i at some future (possibly random) time. Naturally, for a transient state i there is a positive probability of never returning to i. Clearly every absorbing state is recurrent.

Since the two events $(T_i < \infty)$ and $(N_i \geq 1)$ are identical, it follows that a state i is recurrent iff $P(N_i \geq 1 \,|\, X_0 = i) = 1$, that is, the chain starting from the recurrent state i is sure to make at least one visit to i. We will, in fact, show that starting from a recurrent state i, the chain actually makes infinitely many visits to i with probability one. Intuitively, this should be obvious from the Markov property. To do this rigorously and get some other results we need the following identity:

For states i and j and any $m \geq 1$,

$$P(N_j \geq m \,|\, X_0 = i) = f_{ij} \cdot f_{jj}^{m-1}\,. \tag{28}$$

For $m = 1$, this is just the definition of f_{ij}. Let us prove it for $m = 2$. Clearly,

$$P(N_j \geq 2 \,|\, X_0 = i) = \sum_{n=1}^{\infty} \sum_{n'=1}^{\infty} P(A_{n,n'} \,|\, X_0 = i),$$

where $A_{n,n'} = \left\{ \begin{array}{l} X_n = X_{n+n'} = j; \\ X_p \neq j \text{ for } 1 \leq p < n \text{ and for } n+1 \leq p < n+n' \end{array} \right\}$.

By using the properties of conditional probability and the Markov property, each summand reduces to the product

$$P(X_n = j, X_p \neq j, 1 \leq p < n | X_0 = i) \cdot P(X_{n'} = j, X_p \neq j, 1 \leq p < n' | X_0 = j),$$

so that

$$P(N_j \geq 2 | X_0 = i) = \sum_{n=1}^{\infty} \sum_{n'=1}^{\infty} f_{ij}^{(n)} f_{jj}^{(n')} = f_{ij} f_{jj} .$$

The proof for a general m is similar. Do it for $m = 3$ to make sure that you understand.

Notice that the events $(N_j \geq m)$ are decreasing as m increases with limit being the event $(N_j = \infty)$. If j is a transient state then $f_{jj} < 1$, so that for every i, $P(N_j = \infty | X_0 = i) = 0$. That is, a transient state can be visited at most a finite number of times, no matter where the chain starts. On the other hand, if j is recurrent, then $P(N_j = \infty | X_0 = i) = f_{ij}$. In particular $P(N_j = \infty | X_0 = j) = 1$, as stated earlier.

It also follows that if j is a recurrent state, then $E(N_j | X_0 = i) = 0$ or ∞ according as $f_{ij} = 0$ or $f_{ij} > 0$. In particular, $E(N_j | X_0 = j) = \infty$ if j is recurrent. On the other hand, if j is transient then

$$P(N_j = m | X_0 = i) = f_{ij}(1 - f_{jj}) f_{jj}^{m-1} \qquad \text{for } m = 1, 2, \ldots .$$

Since $f_{jj} < 1$, we have $E(N_j | X_0 = i) = f_{ij}/(1 - f_{jj})$. In particular, it follows that $E(N_j | X_0 = i) < \infty$.

The above analysis leads to another characterization of transience and recurrence, namely, *a state j is recurrent if and only if the series $\sum_n p_{jj}^{(n)}$ diverges.* To see this, one defines random variables $(Y_n, n \geq 1)$ as $Y_n = 1$ if $X_n = j$ and $Y_n = 0$ otherwise. Then $N_j = \sum_{n=1}^{\infty} Y_n$, so that, for any i,

$$E(N_j | X_0 = i) = \sum E(Y_n | X_0 = i) = \sum_{n=1}^{\infty} p_{ij}^{(n)} .$$

The above-stated characterization of recurrence follows now by taking $i = j$.

Further, for a transient state j, the series $\sum p_{ij}^{(n)}$ converges to $\dfrac{f_{ij}}{1 - f_{jj}}$ for every state i. In particular, $p_{ij}^{(n)} \to 0$ as $n \to \infty$.

This last observation can be used to deduce that if the state space is finite then there has to be at least one recurrent state. This is clear since $\sum_{j \in S} p_{ij}^{(n)} = 1$ always for every n. Therefore if S is finite, $\lim_{n \to \infty} \sum_{j \in S} p_{ij}^{(n)} = 1$ also. This makes it impossible that $p_{ij}^{(n)} \to 0$ for all j.

Given states i and j we say that i *leads to* j, in symbols $i \hookrightarrow j$, if $f_{ij} > 0$. This can be seen to be equivalent to requiring that $f_{ij}^{(n)} > 0$ for some $n \geq 1$, which in turn is the same as requiring $p_{ij}^{(n)} > 0$ for some $n \geq 1$. It is a simple consequence of the Chapman-Kolmogorov equations that if $i \hookrightarrow j$ and $j \hookrightarrow k$, then $i \hookrightarrow k$.

An important result is that *a recurrent state does not lead to a transient state.* More specifically, if i is recurrent and $i \hookrightarrow j$ then j is recurrent and $f_{ij} = f_{ji} = 1$. We only need to prove the result when $i \neq j$. We first prove that $f_{ji} = 1$. Since $i \hookrightarrow j$, $p_{ij}^{(n)} > 0$ for some $n \geq 1$. Let m be the smallest such n. Then we can get states $i_1, i_2, \ldots, i_{m-1}$, all different from i, such that $p_{ii_1}p_{i_1i_2}\cdots p_{i_{m-1}j} = \alpha > 0$. Suppose, if possible, $f_{ji} < 1$, that is, $P(X_n \neq i \quad \forall n \,|\, X_0 = j) = \beta > 0$. But $P(X_n \neq i \quad \forall n \,|\, X_0 = i)$ is at least as much as

$$P(X_1 = i_1, \ldots, X_{m-1} = i_{m-1}, X_m = j \text{ and } X_{m+n} \neq i \,\forall n \geq 1 \,|\, X_0 = i)\,.$$

By Markov property, it is easy to see that the right hand side equals $\alpha\beta > 0$, contradicting the recurrence of i. Thus we must have $f_{ji} = 1$. In particular, $j \hookrightarrow i$. Now the recurrence of j is derived as follows. Let $m' \geq 1$ be such that $p_{ji}^{(m')} > 0$. Then

$$\sum_{n=1}^{\infty} p_{jj}^{(n)} \quad \geq \quad \sum_{n=1}^{\infty} p_{ji}^{(m')} p_{ii}^{(n)} p_{ij}^{(m)}$$

and the right hand side diverges because both $p_{ji}^{(m')}$ and $p_{ij}^{(m)}$ are strictly positive and i is recurrent. This implies divergence of the left hand side and hence recurrence of j. That $f_{ij} = 1$ is obtained now by reversing the roles of i and j.

Results of Exercise 2 in the previous section really tell us that for a simple random walk with $p > 1/2$, $f_{ij} < 1$ for all $i > j$. Using the fact that $i \hookrightarrow j$ for any two states i and j, we deduce that all states are transient in case $p > 1/2$. One can similarly show that the same is true if $p < 1/2$. It would be an interesting exercise to go back to the formula for d_i and examine what happens in case $p = 1/2$. The reader should be able to show that now $f_{ij} = 1$ for all $i \neq j$ and deduce from this that all states are recurrent.

0.8.3 Decomposition of the State Space: Irreducible Closed Classes

The limiting and steady state behaviour of a Markov chain is intimately connected with a partition of the state space. One way to get the decomposition is to define an equivalence relation between states. Given states i and j, we will say that they are *communicating* if either $(i = j)$ or $(i \hookrightarrow j$ and $j \hookrightarrow i)$. It is easy to see that this is an equivalence relation so that the whole state space S is partitioned as a disjoint union of equivalence classes. These equivalence classes are called *communicating classes.*

From the earlier result it is clear that in a communicating class either all states are recurrent or all states are transient. A communicating class is called *recurrent* (respectively, *transient*) if all states in the class are recurrent (respectively, transient). It is natural to ask how to interpret these equivalence classes in terms of the behaviour of the chain. Let us make a definition first.

Definition: A set $C \subset S$ is said to be *closed* or *stochastically closed* if $p_{ij} = 0$ for $i \in C$ and $j \notin C$ or equivalently, for $i \in C$, $\sum_{j \in C} p_{ij} = 1$.

The condition in the definition above can easily be seen to be equivalent to $\sum_{j \in C} p_{ij}^{(n)} = 1$ for all $i \in C$ and for all $n \geq 1$. This really means that if the chain starts from $i \in C$, then with probability one it remains in C for ever. More precisely,

$$P(X_n \in C \quad \forall n \geq 1 \,|\, X_0 \in C) = 1\,.$$

The state space S is trivially a closed set. A singleton set $C = \{i\}$ is closed iff i is an absorbing state. It is also easy to see that any recurrent communicating class is closed and moreover it does not have a proper subset which is closed. It is therefore natural to ask "what are the minimal closed subsets of S?"

A closed set C is called *irreducible* if $i \hookrightarrow j$ for every i and j in C. It is easy to see that a closed irreducible set C is minimal in the sense that no proper subset of C is closed. A closed communicating class is irreducible. In particular, any recurrent communicating class is closed and hence irreducible. One can not say the same thing about transient classes simply because they may not be closed. In fact, as already remarked, a finite transient class can never be closed. An infinite transient class may or may not be closed.

Exercise 3. Let the state space S be the set of all integers. The transition matrix is given by $P_{i,i+1} = 3/4$ and $p_{i,i-1} = 1/4$. (This is just the simple random walk with $p = 3/4$.) Show that the chain is transient, S is a transient class and S is closed.

Exercise 4. Let the state space S be the set of all non-negative integers. The transition matrix is given by: $p_{00} = 1$ and, for $i \geq 1$, $p_{i,i+1} = 3/4$, $p_{i,i-1} = 1/4$. Then the set of strictly positive integers is a transient class, but not closed.

In passing let us also note that there may not be any closed irreducible set. For example, let $S = \{0, 1, \ldots\}$ and $p_{i,i+1} = 1$. One can easily see that sets of the form $\{k, k+1, \ldots\} \subset S$ are all closed and these are the only closed sets. Thus no closed set is irreducible. Of course, such a behaviour is impossible for a finite state space Markov chain. In fact, for a finite state space Markov chain, the structure is fairly simple. Since finite state space chains are all that we will be needing for our applications, from now on

let us specialize to the case where the state space is finite.

In this case, we have a unique decomposition of the state space S as follows. $S = S_R \cup S_T$, where S_R is the set of recurrent states (necessarily non-empty)

and S_T those of transient states (possibly empty). Further $S_R = \bigcup_{l=1}^{k} C_l$ where each C_l is an irreducible closed set. If the chain starts in C_l it remains there forever visiting each state an infinite number of times with probability one. Thus if $S_T = \emptyset$, we may, depending on the initial state of the chain, study the chain only on a reduced state space, namely, one of the C_l.

In fact, even if $S_T \neq \emptyset$, a chain starting in S_T will, after a finite (random) number of steps, has to enter one of the C_l — and, of course, will remain there from then on. The long-term behaviour of the chain will therefore be still determined by the analysis of the chain on the restricted state space C_l. The only other things that are pertinent in this case are: How long does it take to enter one of the classes C_l and what are the probabilites of entering the different classes? We address these questions first.

Let $i \in S_T$ and $1 \leq l \leq k$. Let

$$\begin{aligned} \alpha_{il} &= P(X_n \in C_l \text{ for some } n \,|\, X_0 = i) \\ &= P(X_n \in C_l \text{ for all large } n \,|\, X_0 = i)\,. \end{aligned}$$

From what has been said above $\sum_{l=1}^{k} \alpha_{il} = 1$. Let us also denote for $i \in S_T$ and for $1 \leq l \leq k$,

$$\beta_{il} = \sum_{j \in C_l} p_{ij} = P(X_1 \in C_l \,|\, X_0 = i)\,.$$

It is now easy to see from the Markov property that

$$\alpha_{il} = \beta_{il} + \sum_{j \in S_T} p_{ij}\alpha_{jl}\,. \tag{29}$$

In other words, for each l, the numbers $(\alpha_{il})_{i \in S_T}$ satisfy the system of linear equations given by (29). In fact, one can show that it is the unique solution. It is convenient to write this system of equations in matrix form. Let Q denote the submatrix of P of order $S_T \times S_T$ defined as $Q = (p_{ij})_{i,j \in S_T}$. It is convenient to think of the rows and columns of Q indexed by states in S_T. For fixed l, $1 \leq l \leq k$, let $\widetilde{\alpha}_l$ and $\widetilde{\beta}_l$ be the column vectors of size S_T with entries α_{il} and β_{il} respectively for $i \in S_T$. Then $\widetilde{\alpha}_l$ is the unique solution of the equation

$$\widetilde{\alpha}_l = \widetilde{\beta}_l + Q\widetilde{\alpha}_l\,.$$

The uniqueness is a consequence of invertibility of the matrix $(I - Q)$, which in turn follows from the fact that the series $I + Q + Q^2 + \cdots$ is convergent and is indeed the inverse of $I - Q$ (finiteness of S_T plays a role here). Here I is the identity matrix. As a consequence

$$\widetilde{\alpha}_l = (I - Q)^{-1}\widetilde{\beta}_l\,.$$

The duration of time that a chain takes, before it enters S_R, is

$$\tau = \min\{n \geq 1 : X_n \in S_R\}.$$

We want to get a formula for the expected value of τ starting from different transient states. For $i \in S_T$, let

$$m_i = E(\tau \,|\, X_0 = i),$$

and let $\widetilde{m}$ be the column vector of size S_T with entries m_i. Denoting $\widetilde{e}$ to be the column vector of size S_T with all entries 1, it is easy to see that $\widetilde{m}$ satisfies the equation

$$\widetilde{m} \quad = \quad \widetilde{e} + Q\widetilde{m},$$

from which the unique solution for $\widetilde{m}$ emerges as

$$\widetilde{m} = (I - Q)^{-1}\widetilde{e}.$$

From the above analysis, it is clear that the matrix $(I - Q)^{-1}$ plays a fundamental role and is appropriately called the *fundamental matrix* of the chain, denoted by N. The above method is often referred to as the *fundamental matrix method.* A particularly useful special case — which would also be singularly relevant for applications in Markov models in genetics — is what are known as *absorbing chains.*

Definition: A Markov chain on a finite state space, for which all the recurrent states are absorbing, is called an *absorbing chain.*

For an absorbing chain, each C_l obtained in the decomposition of S_R, consists of a single absorbing state. Entering the different C_l really means getting absorbed in one of the absorbing states. For each $i \in S_T$, the numbers α_{il}, for $1 \leq l \leq k$, are called the *absorption probabilities* starting from i.

As seen earlier, we can only solve for α_{il} simultaneously for all $i \in S_T$ with l held fixed. In other words each vector $\widetilde{\alpha}_l$ is supposed to be solved separately for each l. However, in case there are only two absorbing states — that is $k = 2$ — then solving for one l is enough (why?).

In case of absorbing chains, it is notationally convenient to list the states so that the absorbing states come before the transient states. To avoid triviality, we assume that there is at least one transient state. With this ordering of states, the transition matrix P takes the form

$$P \quad = \quad \begin{pmatrix} I & O \\ R & Q \end{pmatrix},$$

where Q is as before. The entries of NR are precisely the absorption probabilities α_{il} for $i \in S_T$ and $1 \leq l \leq k$. Entries of $N\widetilde{e}$ are precisely the mean times till absorption.

Exercise 5. Consider a chain with four states and transition matrix

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

Calculate the fundamental matrix, the absorption probabilities and mean times till absorption.

Exercise 6. For an absorbing chain, show that the vector of variances of the time till absorption starting from different transient states is given by $\widetilde{V} = (2N - I)\widetilde{m} - \widetilde{m}^2$ where $\widetilde{m}^2$ denotes the vector whose entries are the squares of the entries of $\widetilde{m}$.

0.8.4 Ergodic Chains: Limiting Behaviour and Stationary Distributions

We go back to a general finite state space Markov chain and recall the decomposition of the state space

$$\begin{aligned} S &= S_R \cup S_T \\ &= \bigcup_{l=1}^{k} C_l \cup S_T \end{aligned}$$

where S_T and S_R are the sets of transient and recurrent states respectively and C_l, $1 \leq l \leq k$, are the closed irreducible subsets of S_R.

As noted already, the long-term behaviour of the chain is determined by the analysis of the chain on the restricted state spaces C_l. This is what we want to pursue now. Accordingly, let us assume that the whole state space S of the chain is one single closed irreducible class of recurrent states. Such a chain is called an *irreducible recurrent chain*, also sometimes referred to as an *ergodic chain*. For such a chain, we are going to show that there exists a unique $\pi = \{\pi_i, i \in S\}$ with $\pi_i > 0$ and $\sum \pi_i = 1$ such that $\pi P = \pi$ or equivalently, there is a unique probability π on S such that if $X_0 \sim \pi$ then $X_1 \sim \pi$ (in fact, $X_n \sim \pi$ for all n). It is this property that is described by saying that π is an *invariant distribution* for the chain — also called a *steady-state* or an *equilibrium* or a *stationary* distribution. Formally,

Definition: By an *invariant distribution* of a Markov chain is meant a probability distribution π on the state space S for which the equality $\pi P = \pi$ holds.

Exercise 7. If $X_0 \sim \pi$ where π is an invariant distribution, show that the joint distribution of (X_n, X_{n+1}) is same for all n. Generalize from pairs to triplets etc.

A Markov chain may not have any invariant distribution. In case it has, it may have more than one invariant distributions. It is not difficult to show that symmetric simple random walk has no invariant distribution. On the

other extreme, the Markov chain on $S = \{0, 1\}$, with $p_{00} = p_{11} = 1$, is an easy example of a chain with plenty of invariant distributions! Our following analysis will show, among other things, that a finite state space Markov chain will always have at least one invariant distribution.

Indeed, we will show that for an ergodic chain with finitely many states there is one and only one invariant distribution. This invariant distribution π will also turn out to be the limiting distribution of the chain in the following sense:

$$\text{for all} \quad i, j \in S, \quad \lim_n \frac{1}{n} \sum_{l=1}^{n} p_{ij}^{(l)} \quad = \quad \pi_j \,. \tag{30}$$

It may be noted that the above limit does not depend on i. In other words, the effect of the initial distribution wears off in the long run.

We first prove that the limits in the left-hand side of (30) exist and are free of i. We start with some notations. For any state i, $P_i(A)$ will denote the conditional probability of the event A, given $X_0 = i$. Thus $P_i(A)$ can be thought of as the probability of the event A when the initial distribution is concentrated at i. Expectation wih respect to this probability will be denoted by E_i. Let us, from now on, fix a state j. Let $T_1, T_2, \ldots$ be the times of successive visits to the state j. Only times greater than or equal to one are considered for these visits. That is,

$$T_1 \quad = \quad \min\{n \geq 1 : X_n = j\},$$

and for $r \geq 2$,

$$T_r \quad = \quad \min\{n > T_{r-1} : X_n = j\} \,.$$

Since S is a closed irreducible recurrent class, $P_i(T_r < \infty \text{ for all } r) = 1$. Set $Z_0 = T_1$ and for $r \geq 1$, $Z_r = T_{r+1} - T_r$. We claim that the Markov property of (X_n) implies that for any i, the random variables $Z_1, Z_2, \ldots$ are i.i.d. under P_i. Moreover, the common distribution is that of T_1 under P_j (and hence does not depend on the initial state i). To see this, note first that $P_i(Z_1 = l_1 \,|\, Z_0 = l_0)$ equals the conditional probability of $(X_{l_0+m} \neq j, 0 < m < l_1; X_{l_0+l_1} = j)$, given $(X_0 = i; X_m \neq j, 0 < m < l_0; X_{l_0} = j)$ which by Markov property at time l_0 equals

$$P(X_m \neq j, 0 < m < l_1; X_{l_1} = j \,|\, X_0 = j) = P_j(T_1 = l_1) \,.$$

Next, $P_i(Z_2 = l_2 \,|\, Z_0 = l_0, Z_1 = l_1)$ equals the conditional probability of the event A given the event B, where

$$A = \{X_{l_0+l_1+m} \neq j, 0 < m < l_2, X_{l_0+l_1+l_2} = j\},$$

and

$$B = \left\{ \begin{array}{l} X_0 = i, X_{l_0} = X_{l_0+l_1} = j, X_m \neq j, 0 < m < l_0, \\ X_{l_0+m} \neq j, 0 < m < l_1 \end{array} \right\} .$$

Once again, by the Markov property at time $l_0 + l_1$, the above conditional probability equals

$$P(X_m \neq j, 0 < m < l_2; X_{l_2} = j \,|\, X_0 = j) = P_j(T_1 = l_2)\,.$$

One can use similar technique to show that for any $r \geq 1$,

$$P_i(Z_{r+1} = l_{r+1} \,|\, Z_0 = l_0, \ldots, Z_r = l_r) = P_j(T_1 = l_{r+1})\,.$$

From this, our claim about the P_i distribution of the sequence $(Z_r)_{r\geq 1}$ can easily be proved.

Now an application of the strong law of large numbers yields

$$\frac{Z_1 + \cdots + Z_r}{r} \longrightarrow E_j(T_1) \quad \text{as } r \to \infty$$

with P_i-probability one, for any i. It should be remarked that the strong law of large numbers used here does not need any apriori assumption on finiteness of the expectation $E_j(T_1)$. This is because the random variables Z_i are all non-negative. (See the paragraph following Exercise 5 in Section 0.5.) By the definition of the sequence $Z_1, Z_2, \ldots$, we have

$$Z_1 + \cdots + Z_r = T_{r+1} - T_1\,.$$

Thus

$$\frac{T_{r+1} - T_1}{r} \longrightarrow E_j(T_1) \quad \text{as } r \to \infty,$$

with P_i-probability one. But then

$$\frac{T_r - T_1}{r} = \frac{T_r - T_1}{r-1}\,\frac{r-1}{r} \longrightarrow E_j(T_1) \quad \text{as } r \to \infty,$$

with P_i-probability one. Since $P_i(T_1 < \infty) = 1$ and hence $\frac{T_1}{r} \longrightarrow 0$ as $r \to \infty$, we have with P_i-probability one

$$\frac{T_{r+1}}{r} \longrightarrow E_j(T_1) \quad \text{as well as} \quad \frac{T_r}{r} \longrightarrow E_j(T_1) \quad \text{as } r \to \infty\,.$$

For each $n \geq 1$, let us consider the random variable

$$N_n \quad = \quad \#\{1 \leq l \leq n : X_l = j\}\,.$$

Since we have an ergodic chain, $N_n \to \infty$ as $n \to \infty$ with P_i-probability one, for any i, so that

$$\frac{T_{N_n+1}}{N_n} \longrightarrow E_j(T_1) \quad \text{and} \quad \frac{T_{N_n}}{N_n} \longrightarrow E_j(T_1) \quad \text{as} \quad n \to \infty\,.$$

But by definition of N_n, one clearly has $T_{N_n} \leq n < T_{N_n+1}$ so that

$$\frac{T_{N_n}}{N_n} \quad \leq \quad \frac{n}{N_n} \quad \leq \quad \frac{T_{N_n+1}}{N_n}\,.$$

Thus we have, for every i, $\dfrac{n}{N_n} \to E_j(T_1)$ as $n \to \infty$ or, equivalently,

$$\frac{N_n}{n} \quad \longrightarrow \quad \frac{1}{E_j(T_1)} \quad \text{as} \quad n \to \infty,$$

with P_i-probability one. Since $0 \leq \frac{N_n}{n} \leq 1$ for all n, we have by the dominated convergence theorem that

$$\frac{1}{n} E_i(N_n) \quad \longrightarrow \quad \frac{1}{E_j(T_1)} \quad \text{as} \quad n \to \infty\,.$$

It is easy to identify $E_i(N_n)$ as $\sum_{l=1}^{n} p_{ij}^{(l)}$. This proves that $\lim_{n\to\infty} \frac{1}{n} \sum_{l=1}^{n} p_{ij}^{(l)}$ exists for all i and j, and equals $1/E_j(T_1)$. Denote this quantity by π_j. It is clear that $\pi_j \geq 0$ for each j. Of course, $\pi_j = 0$ is not yet ruled out but will be ruled out soon. Also, since each P^n is a stochastic matrix of finite order one gets $\sum \pi_j = 1$. Thus $\pi = (\pi_j : j \in S)$ is a probability on the state space. We next show that $\pi P = \pi$, or in other words, $(\pi P)_j = \pi_j$ for each j. Fix any arbitrary state k.

$$\begin{aligned}(\pi P)_j &= \sum_i \pi_i p_{ij} = \sum_i \lim_n \frac{1}{n} \sum_{l=1}^{n} p_{ki}^{(l)}\, p_{ij} = \lim_n \frac{1}{n} \sum_{l=1}^{n} \sum_i p_{ki}^{(l)} p_{ij} \\ &= \lim_n \frac{1}{n} \sum_{l=1}^{n} p_{kj}^{(l+1)} = \lim_n \frac{1}{n} \left[\sum_{l=1}^{n} p_{kj}^{(l)} - p_{kj}^{(1)} + p_{kj}^{(n+1)} \right] = \pi_j\,.\end{aligned}$$

This shows that our π is indeed an invariant distribution. To show uniqueness, let $\widetilde{\pi} = (\widetilde{\pi}_j)$ be any invariant distribution. From $\widetilde{\pi} P^n = \widetilde{\pi}$, one easily gets that for each j,

$$\frac{1}{n} \sum_{l=1}^{n} \sum_i \widetilde{\pi}_i p_{ij}^{(l)} \quad = \quad \widetilde{\pi}_j\,.$$

Letting $n \to \infty$, the left-hand side equals $\sum_i \widetilde{\pi}_i \pi_j = \pi_j$ showing that $\widetilde{\pi} = \pi$. It is only appropriate to draw the attention of the reader to an important fact lest it be overlooked. In our analysis above, we have repeatedly taken the liberty of interchanging sum and limits at our will. This was sponsored by the assumption of finiteness of the state space. The case of infinite state space could be a very different ball game.

Thus we have proved that for an ergodic finite state Markov chain, there is a unique invariant distribution π which is also the limiting distribution in the sense of (30). Indeed our proof also shows that for each j, $1/\pi_j$ is nothing but the expected time to return to j, given that the chain started at j. As noted already, this expected value could potentially be infinite for some j, leading to $\pi_j = 0$.

We now go a step further and show that for ergodic finite state Markov chains, $\pi_j > 0$ for all j which, in turn, would also imply that, starting from any

state j, the expected time to return is finite. Indeed, suppose that for some j, $\pi_j = 0$. Fix any $i \neq j$ and an $l \geq 1$ such that $p_{ij}^{(l)} > 0$. Since $\pi P^l = \pi$, we have $\pi_j = \sum_k \pi_k p_{kj}^{(l)} \geq \pi_i p_{ij}^{(l)}$, so that $\pi_i = 0$. Thus $\pi_j = 0$ for some j implies that $\pi_i = 0$ for all i which contradicts $\sum \pi_i = 1$.

For an irreducible recurrent chain we already knew that starting from a state j, we are sure to return to j sometime or the other. What we have just shown is that if the state space is moreover finite then the expected time to return is also finite. This property is often referred to in the literature as *positive recurrence.* This is not true in general, that is, a recurrent state may fail to be positive recurrent, if the state space is infinite. Such recurrent states are called *null recurrent.*

Another natural question that arises out of (30) is : why do we not consider simply the $\lim_n p_{ij}^{(n)}$ instead of the averages $\frac{1}{n}\sum p_{ij}^{(l)}$ as was done above? It is not difficult to see that $\lim_n p_{ij}^{(n)}$ may fail to exist, in general. In fact, a two state chain with transition matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

will illustrate this. What is happening in this example is that, for any i and j, exactly one of $p_{ij}^{(n)}$ and $p_{ij}^{(n+1)}$ is positive for each n. In fact, for $i = j$, $p_{ij}^{(n)}$ is positive (indeed, equals 1) if and only if n is even, while for $i \neq j$, this happens if and only if n is odd.

Usually, it is only such periodic behaviour of the chain, as illustrated in the example above, that prevents the existence of $\lim_n p_{ij}^{(n)}$. We are not going to pursue the periodicity properties and their consequences here. However, for subsequent applications, we are going to describe now (without proofs) what happens if such periodic behaviour is ruled out.

For a Markov chain, a state j is said to be *aperiodic* if $\{n \geq 1 : p_{jj}^{(n)} > 0\}$ has greatest common divisor 1. It is immediate that none of the two states in the above example are aperiodic — the g.c.d is 2 for both. It can be shown that in an irreducible Markov chain, either all states are aperiodic or none are and, in the first case, the chain is said to be *aperiodic.* Now we can state the main result without proof.

Theorem: *If an ergodic finite state chain is aperiodic, then for all states i and j, the limit $\lim_{n\to\infty} p_{ij}^{(n)}$ exists and equals π_j where $\pi = (\pi_j, j \in S)$ is the unique invariant distribution.*

In effect what it says is that for an aperiodic ergodic chain, π is the limiting distribution of the chain, irrespective of how it starts.

Exercise 8. Consider a Markov chain with r states. Suppose that the transition matrix has the property that each column sum is one (remember that for a transition matrix each row sum is one). If the chain is irreducible then

show that the uniform distribution on the state space is the unique stationary distribution. What do you infer about the expected times to return in this case? What if it is not irreducible?

Exercise 9. Let a be a probability vector with strictly positive entries and length 10. Consider a Markov chain with 10 states. Let the transition matrix have identical rows, each row being a. What is the stationary distribution? What chain are we talking about? What if the vector is not strictly positive?

Exercise 10. Consider a chain with 4 states and the following transition matrix.

$$\begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

Denoting by τ the time (≥ 1) of the first visit to the state 2, calculate $E_i(\tau)$ for each state i. Suppose that μ is an initial distribution on the state space. Calculate $E_\mu(\tau)$.

0.8.5 Absorbing Chains: Limiting Behaviour, Rate of Convergence

Recall that an absorbing chain is a finite state Markov chain consisting only of absorbing and transient states. Since the state space is finite, there is at least one absorbing state. To avoid trivialities, we assume that there is at least one transient state also. Specifically, let us assume that there are m states, $\{1, 2, \ldots, m\}$, of which the first k are absorbing and the remaining transient. As already seen, the transition matrix has the structure

$$P = \begin{pmatrix} I & O \\ R & Q \end{pmatrix},$$

where I is the identity matrix of order k and R and Q are of orders $(m-k) \times k$ and $(m-k) \times (m-k)$ respectively. As already observed, the fundamental matrix $N = (I - Q)^{-1}$ plays an important role. For example, the matrix NR equals $((\alpha_{ij}))$ where α_{ij} for $k+1 \leq i \leq m$ and $1 \leq j \leq k$ are the absorption probabilities. For an absorbing state j, it follows from the continuity property of probability that

$$\alpha_{ij} = \lim_{n \to \infty} P_i(X_n = j) = \lim_n p_{ij}^{(n)}. \tag{31}$$

Recall that, for $k+1 \leq j \leq m$, $\lim_n p_{ij}^{(n)} = 0$ for all i. All of these can be stated in matrix form as

$$P^n \longrightarrow \begin{pmatrix} I & O \\ A & O \end{pmatrix}$$

where $A = NR = ((\alpha_{ij}))$.

We now want to show that the convergence in (31) happens geometrically fast and also calculate the exact rate of convergence. Interest in the rate of convergence lies in the fact that just like the expected time till absorption, this rate also gives another indication as to how fast the chain gets trapped in one of the absorbing states.

To begin with, let us recall that a number λ (possibly complex) is called an *eigenvalue* of P if there is a non-null vector $\widetilde{u}$ (with possibly complex entries) such that $\widetilde{u}P = \lambda\widetilde{u}$. Such a non-null vector $\widetilde{u}$ is called a *left eigenvector* corresponding to the eigenvalue λ. Recall also that the set of all vectors $\widetilde{u}$ with $\widetilde{u}P = \lambda\widetilde{u}$ is a vector space, called the *left eigenspace* associated to λ.

Let us now observe that a transition matrix P cannot have an eigenvalue λ with $|\lambda| > 1$. If possible, suppose $|\lambda| > 1$ and $\widetilde{u} = (u_1, u_2, \ldots, u_m)$ is a non-null vector with $\widetilde{u}P = \lambda\widetilde{u}$. Then clearly, $\widetilde{u}P^n = \lambda^n\widetilde{u}$ for all $n \geq 1$. Let j be such that $u_j \neq 0$. We then get a contradiction from the fact that

$$\sum_i u_i p_{ij}^{(n)} = \lambda^n u_j \,,$$

where the left-hand-side remains bounded by $\sum |u_i|$ for all n, while the right-hand-side is unbounded.

We next show that if P is the transition matrix of an absorbing chain with k absorbing states, then the dimension of the eigenspace associated to $\lambda = 1$ is exactly k. Indeed, let $\widetilde{u}$ be any vector with $\widetilde{u}P = \widetilde{u}$ which, of course, implies $\widetilde{u}P^n = \widetilde{u}$ for all n. Then, for any $j \geq k+1$, $u_j = \sum u_i p_{ij}^{(n)} \to 0$ as $n \to \infty$ showing that $u_{k+1} = \cdots = u_m = 0$. Thus, the dimension of the eigenspace is at most k. On the other hand, for each i, $1 \leq i \leq k$, the vector $\widetilde{u}^i$ with i-th coordinate equal to 1 and other coordinates 0, can easily seen to be a left eigenvector corresponding to $\lambda = 1$. So there are exactly k linearly independent left eigenvectors for $\lambda = 1$.

Finally, for P as above, we show that $\lambda = -1$ cannot be an eigenvalue. Suppose that $\widetilde{u}$ satisfies $\widetilde{u}P = -\widetilde{u}$ and hence $\widetilde{u}P^n = (-1)^n\widetilde{u}$ for all n. For $j \geq k+1$, $(-1)^n u_j = \sum u_i p_{ij}^{(n)}$ again yields that $u_j = 0$. For $j \leq k$,

$$-u_j = \sum_{i=1}^{m} u_i p_{ij} = \sum_{i=1}^{k} u_i p_{ij} = u_j$$

implying again that $u_j = 0$. Thus any $\widetilde{u}$ satisfying $\widetilde{u}P = -\widetilde{u}$ must be null.

To continue with our discussion of the rate of convergence, let us assume that there are m real eigenvalues $\lambda_1, \ldots, \lambda_m$ (not necessarily distinct) of P, with associated left eigenvectors $\widetilde{u}^1, \ldots, \widetilde{u}^m$, which are linearly independent. As shown earlier, we can and do take $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 1$ and $\widetilde{u}^i$, $1 \leq i \leq k$, as defined above. If the remaining eigenvalues are listed in decreasing order of magnitude, we will clearly get

$$1 = \lambda_1 = \cdots = \lambda_k > |\lambda_{k+1}| \geq |\lambda_{k+2}| \geq \cdots \geq |\lambda_m| \,.$$

Denote by L the $m \times m$ matrix whose i-th row is $\widetilde{u}^i$. The above equations can be reformulated as

$$LP = DL$$

where $D = \text{Diag}(\lambda_1, \ldots, \lambda_m)$. Since the vectors $\widetilde{u}^i$ are linearly independent, the matrix L is invertible and therefore

$$P = L^{-1} D L. \tag{32}$$

Readers initiated to linear algebra would quickly recognize the above as the *spectral representation* of P. Our assumption therefore really amounts to P having a spectral representation. It follows from (32) that

$$P^n = L^{-1} D^n L,$$

where clearly $D^n = \text{Diag}(\lambda_1^n, \ldots, \lambda_m^n)$. In particular, for any $1 \leq i, j \leq m$,

$$p_{ij}^{(n)} = \sum_{r=1}^{m} \lambda_r^n (L^{-1})_{ir} L_{rj}.$$

Since $\lambda_1 = \cdots = \lambda_k = 1$, we have

$$\begin{aligned} |p_{ij}^{(n)} - \sum_{r=1}^{k} (L^{-1})_{ir} L_{rj}| &\leq \sum_{r=k+1}^{m} |\lambda_r|^n |(L^{-1})_{ir} L_{rj}| \\ &\leq |\lambda_{k+1}|^n \sum_{r=k+1}^{m} |(L^{-1})_{ir} L_{rj}|. \end{aligned}$$

Of course, $|\lambda_{k+1}| < 1$ implies that the left side goes to zero as $n \to \infty$ and the convergence is geometrically fast with rate not larger than $|\lambda_{k+1}|$. Incidentally this argument also shows that

$$\sum_{r=1}^{k} (L^{-1})_{ir} L_{rj} = \lim_{n \to \infty} p_{ij}^{(n)}$$

for all i and j. Of course, if i is a transient state (that is, $i \geq k+1$) and j is an absorbing state (that is, $j \leq k$), then this quantity is precisely α_{ij}, the probability of absorption in j starting from i. We leave it as an exercise to verify that (i) if i is absorbing, then $\sum_{r=1}^{k} (L^{-1})_{ir} L_{rj}$ equals δ_{ij} and (ii) if j is transient then this is zero.

Exercise 11. In the above discussion of convergence rate for absorbing chain, we assumed spectral representation for the transition matrix. However, a transition matrix need not always admit a spectral representation (32). Show that the following transition matrices do not admit spectral representation. For chains with these transition matrices, find the rates of convergence.

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2/3 & 0 & 1/3 & 0 \\ 2/3 & 0 & 0 & 1/3 \\ 2/3 & 1/3 & 0 & 0 \end{pmatrix}, \qquad P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 2/3 & 1/3 & 0 & 0 \end{pmatrix}.$$

0.9 Continuous Time Processes

Discrete time stochastic processes are used to describe evolution of systems that change only at discrete instants of time. The relevant time set is the set of time points at which changes may take place, and, is usually taken as $\{0, 1, 2, \ldots\}$. In the earlier two sections, we discussed some special types of such processes, namely, martingales and Markov chains.

In this section, we discuss stochastic processes evolving over a continuum of time or in other words, *continuous time stochastic processes.* Even though the process evolves over a continuous time, distinction would be made as to the nature of the evolution. Let us consider two simple examples to make the distinction clear.

Imagine a telephone exchange through which calls pass at random instants of time. If we consider the number of calls passing through upto time t, then we have a stochastic process (X_t) over time set $t \in [0, \infty)$. However, the state space of the process is $\{0, 1, 2, \ldots\}$, which is a discrete set and the process evolves only through jumps.

A different example would be the kinetic movement of a gas molecule where the position of the particle changes continuously with time and not through jumps. In other words, here the state space is also a continuum.

In the first subsection, we will discuss a special class of processes of the first type, namely *Markov chains in continuous time.* Such processes will be used in connection with temporal spread of epidemics in Chapter 4.

The second subsection would be devoted to a special class of processes of the second type — known as *diffusion processes.* It is worth noting here that a diffusion process may sometimes serve also as an approximation to a discrete time Markov chain and often allows us to get good approximations to quantities of interest, related to the original discrete chain. Indeed, it is mainly this application of diffusion processes which will be used in connection with Markov models in genetics in Chapter 3.

The interval $[0, \infty)$ is usually taken as the time set for a continuous time process. Thus, we consider a family of random variables X_t, indexed by all real numbers $t \geq 0$, each taking values in a set S. This constitutes a continuous time process and, in analogy with discrete time processes, is denoted $(X_t)_{t \geq 0}$. The set S is called the *state space* of the process. The notion of Markov property for stochastic processes has been already encountered in the discrete set-up. It simply means that at any point of time, given the present state of the process, the future evolution is (conditionally) independent of the history of the past. A simple formulation of this idea in the continuous time case is as follows.

Definition: A process $(X_t)_{t \geq 0}$ with state space S is said to be a *Markov process* if, for any choice $0 \leq s_1 < s_2 < \cdots < s_n < s < t + s$ of time points, the conditional distribution of the random variable X_{t+s}, given $(X_{s_1}, \ldots, X_{s_n}, X_s)$,

depends only on X_s, that is, for $B \subset S$,

$$P(X_{t+s} \in B \,|\, X_{s_1} = x_1, \ldots, X_{s_n} = x_n, X_s = x) = P(X_{t+s} \in B \,|\, X_s = x)\,.$$

If, furthermore, these conditional distributions are the same for all s, that is, the right side of the above equation depends only on t and not on s, then the process $(X_t)_{t\geq 0}$ is said to be *time-homogeneous*.

In the above definition, the set B can be any subset of S in case S is a countable set. However, as discussed in Section 0.4, one has to be more selective, in case S is not countable. Of course, for almost any conceivable B, the above property has to hold.

We will consider here only time-homogeneous Markov processes. Thus, for all $t > 0$ and $s \geq 0$, we have

$$P(X_{t+s} \in B \,|\, X_s = x) = P(X_t \in B \,|\, X_0 = x)\,.$$

Let us denote this by $P_t(x, B)$. Clearly, for each $t \geq 0$ and each point x in the state space S, $P_t(x, \cdot)$ is a probability distribution on S. This family of distributions, as t and x vary — called the family of *transition probabilities* — play the same role as that of the transition matrix and its powers for a Markov chain in capturing the evolutionary mechanism of the whole process $(X_t)_{t\geq 0}$.

Just like in the case of discrete Markov chains, the family of transition probabilities $P_t(x, \cdot)$ here also satisfy the Chapman-Kolmogorov equations, which now reads as

$$P_{t+s}(x, B) = \int P_s(y, B) P_t(x, dy) \quad \text{for all } t \geq 0, s \geq 0\,.$$

The interpretation of the integral is not difficult. It is simply a notation for $E(P_s(X_t, B) \,|\, X_0 = x)$. We will return to this in the next two subsections and see that the equations take on simpler forms under special assumptions.

0.9.1 Continuous Time Markov Chains

In this section, we assume that the state space is countable, that is, each X_t is a discrete random variable taking values in a countable set S. The time-homogeneous Markov property reduces to

$$P(X_{t+s} = j \,|\, X_{s_1} = i_1, \ldots, X_{s_n} = i_n, X_s = i) = P_{ij}(t) = P(X_t = j \,|\, X_0 = i)$$

for all $0 \leq s_1 < s_2 < \cdots < s_n < s < t + s$ and $i, j \in S$. It can be shown that the above equation actually implies that for any $s > 0$, the conditional distribution of $(X_{t+s})_{t\geq 0}$, given $(X_u)_{u<s}$ and $X_s = i$, is the same as that of $(X_t)_{t\geq 0}$, given $X_0 = i$. In particular,

$$P(X_{t+s} = j \,|\, X_u, u < s; X_s = i) = P_{ij}(t)\,.$$

If $P(t)$ denotes the $S \times S$ matrix whose (i, j)-th entry is $P_{ij}(t)$, then each $P(t)$ is clearly a stochastic matrix. Thus we have a family $\{P(t), t \geq 0\}$ of stochastic

matrices. Here $P(0)$ is the identity matrix of size S. The Chapman-Kolmogorov equations are easily seen to correspond to the *semigroup* property

$$P(t+s) = P(t) \cdot P(s) .$$

As mentioned earlier, the family $(P(t))_{t\geq 0}$ plays the same role as the sequence $(P^n)_{n\geq 0}$ of the n-step transition matrices in case of discrete Markov chains. The notable difference is that while the P^n are all determined by the one-step transition matrix P, it is not clear how to get one single matrix that will determine all the $P(t)$ for $t \geq 0$. We are now going to show how to do this.

One may recall that if $P : [0,\infty) \longrightarrow R$ is a continuous function with $P(t+s) = P(t) \cdot P(s)$ and $P(0) = 1$, then $P'(0)$ exists and determines $P(t)$ for all values of t. Indeed if $P'(0) = Q$, then $P(t) = e^{Qt}$ for all t. In particular $P'(t) = P(t) \cdot Q = Q \cdot P(t)$ for all t. Indeed this differential equation along with the initial condition $P(0) = 1$ also characterizes the function $P(t) = e^{Qt}$.

Our present situation is quite similar to this except that, instead of real-valued functions, we are dealing with an $S \times S$ matrix-valued function $P(t)$. We want to show that under certain conditions, $P(t)$ also satisfies the matrix differential equations

$$P'(t) = Q \cdot P(t) = P(t) \cdot Q$$

for some matrix $Q = (q_{ij})$. In other words, we have the following two systems of equations

$$P'_{ij}(t) = \sum_k q_{ik} P_{kj}(t) \qquad i,j \in S \tag{33}$$

$$P'_{ij}(t) = \sum_k P_{ik}(t) q_{kj} \qquad i,j \in S \tag{34}$$

Unlike in the real-valued case, the two systems are not identical. System (33) is always true and is known as Kolmogorov's *Backward Equations*. System (34) which is true under some additional regularity conditions, is known as Kolmogorov's *Forward Equations*, also known as *Fokker-Planck* equations. We proceed to give a derivation of the above equations, assuming that the state space is finite. Indeed, it is only in the proof of the forward equations that the finiteness of the state space will be used. Our derivation will also identify the matrix Q, frequently known as the *Q-matrix* of the chain.

We first prove two basic lemmas which will give us a description of how the chain evolves with time. Let

$$T = \inf\{t > 0 : X_t \neq X_0\} .$$

In other words, T is the first time the system leaves the initial state.

Lemma 1: *For any* $s, t \geq 0$,

$$P(T > t+s \,|\, X_0 = i) = P(T > t \,|\, X_0 = i) P(T > s \,|\, X_0 = i) .$$

Proof: First assume that $s > 0$ and $t > 0$.

$$\begin{aligned} P(T > t+s \,|\, X_0 = i) &= P(T > t+s, T > s \,|\, X_0 = i) \\ &= P(T > t+s, X_s = i, T > s \,|\, X_0 = i) \\ &= P(T > s, X_s = i \,|\, X_0 = i) \\ &\quad \times P(T > t+s \,|\, X_0 = i, T > s, X_s = i) \\ &= P(T > s \,|\, X_0 = i) \cdot P(T > t \,|\, X_0 = i)\,, \end{aligned}$$

where the equality $P(T > t+s \,|\, X_0 = i, T > s, X_s = i) = P(T > t \,|\, X_0 = i)$ follows from the assumed Markov property. The case $s = 0$ and/or $t = 0$ follows by taking limits. ∎

A consequence of the above is that, for any $i \in S$, there is a $\lambda_i \in [0, \infty]$ such that, $P(T > t \,|\, X_0 = i) = e^{-\lambda_i t}$ for all $t \geq 0$. In particular, $P(T > 0 \,|\, X_0 = i)$ is either one or zero (according as λ_i is finite or not). Also, the case $\lambda_i = 0$ corresponds to $P(T = \infty \,|\, X_0 = i) = 1$. Clearly, $0 < \lambda_i < \infty$ refers to an exponential distribution as encountered in Section 0.4. However, we agree here to use the term exponential distribution in a broad sense even when λ_i equals 0 or ∞.

Lemma 2: *For any i, j with $i \neq j$ and any $s > 0$,*

$$P(T > s, X_T = j \,|\, X_0 = i) = P(T > s \,|\, X_0 = i) P(X_T = j \,|\, X_0 = i)\,.$$

Proof: This is clearly true if $P(T > 0 \,|\, X_0 = i) = 0$. We assume therefore that $P(T > 0 \,|\, X_0 = i) = 1$.

$$\begin{aligned} P(T > s, X_T = j \,|\, X_0 = i) &= P(T > s, X_T = j, X_s = i \,|\, X_0 = i) \\ &= P(X_s = i, T > s \,|\, X_0 = i) \times P(X_T = j \,|\, X_0 = i, X_s = i, T > s) \\ &= P(T > s \,|\, X_0 = i) P(X_T = j \,|\, X_0 = i, T > 0) \\ &= P(T > s \,|\, X_0 = i) P(X_T = j \,|\, X_0 = i)\,. \end{aligned}$$

∎

The content of the two lemmas is the following. For every state i, there is a number $\lambda_i \in [0, \infty]$ and transition probabilities p_{ij} for $j \neq i$. If the chain starts in the state i, it remains there for an exponentially distributed random time T_1 with mean $1/\lambda_i$ and then moves to state j with probability p_{ij}, independently of T_1. Subsequently, the chain behaves as if started from state j. It may be pointed out that $\lambda_i = \infty$ corresponds to $P(T_1 = 0 \,|\, X_0 = i) = 1$, meaning that the chain instantaneously jumps from the state i. Such states are called *instantaneous states*. It can be shown that this contingency is not possible in a finite state chain. In general, we assume that there are no such states. It may also be pointed out that $\lambda_i = 0$ corresponds to $P(T_1 = \infty \,|\, X_0 = i) = 1$, meaning that the chain starting at i remains there forever. That is, i is *absorbing*. Thus only $\lambda_i > 0$ corresponds to the case when the waiting time in state i is a proper exponential random variable. In any case, from the above description it is clear that the evolution of the chain is completely captured

by the parameters $(\lambda_i, i \in S)$ and the stochastic matrix $((p_{ij}))_{i,j\in S}$ with zero diagonal entries.

From the above description, it should also be clear that if $T_1, T_2, \ldots$ represent the successive (random) times of jumps of the chain then the sequence of random variables defined as

$$Y_0 = X_0, \quad Y_n = X_{T_n} \quad \text{for } n \geq 1$$

would form a discrete time Markov chain with state space S. The one-step transition probabilities of the Markov chain are given by p_{ij}, if $\lambda_i > 0$. For i such that $\lambda_i = 0$, we have $p_{ii} = 1$.

The chain $(Y_n)_{n\geq 0}$ is usually called the *embedded chain.* For many of the properties of the continuous time chain, like classification of states, asymptotic behaviour and existence of invariant distributions, it suffices to examine only the embedded chain. Of course, some important features that explicitly make use of the waiting times at various states would not be captured by the embedded chain. For more on embedded chains, the reader may look at the book of Bhattacharya and Waymire.

We now proceed towards proving Kolmogorov's backward equations (33). Let $i \in S$ be such that $\lambda_i > 0$. Then for any $j \in S$ and any $t > 0$, we have by conditioning on the time of the first jump from i,

$$\begin{aligned} P_{ij}(t) &= \sum_{k\neq i} \int_0^t \lambda_i e^{-\lambda_i s} p_{ik} P_{kj}(t-s)ds + e^{-\lambda_i t}\delta_{ij} \\ &= \sum_{k\neq i} e^{-\lambda_i t} p_{ik} \int_0^t \lambda_i e^{\lambda_i u} P_{kj}(u)du + e^{-\lambda_i t}\delta_{ij}\,. \end{aligned}$$

Note that, in case $j = i$, the process starting from i may be in state $j\,(= i)$ at time t by simply waiting at the initial state at least till time t. The term $e^{-\lambda_i t}\delta_{ij}$ occurs to take care of this contingency. Of course, for $j \neq i$, this contingecy does not arise and therefore the term has no contribution. Here δ_{ij} is the usual Kronecker delta, that is, δ_{ij} equals 1 or 0 according as $i = j$ or $i \neq j$. The above equation shows that $P_{ij}(t)$ is continuous in t. In case the state space is finite, this is immediate because each summand is continuous in t. In general, one needs to use the Dominated Convergence Theorem. The continuity of $P_{ij}(t)$, in turn, gives differentiability also and indeed, the sum on the right side can be differentiated term by term. This requires the fundamental theorem of calculus as well as the Dominated Convergence Theorem. The upshot is

$$\begin{aligned} P'_{ij}(t) &= -\lambda_i \left(\sum_{k\neq i} e^{-\lambda_i t} p_{ik} \int_0^t \lambda_i e^{\lambda_i u} P_{kj}(u)du + e^{-\lambda_i t}\delta_{ij} \right) \\ &\qquad + \sum_{k\neq i} e^{-\lambda_i t} p_{ik} \lambda_i e^{\lambda_i t} P_{kj}(t) \\ &= -\lambda_i P_{ij}(t) + \sum_{k\neq i} \lambda_i p_{ik} P_{kj}(t)\,. \end{aligned}$$

In other words, denoting

$$q_{ik} = \lambda_i p_{ik} \quad \text{for} \quad k \neq i; \quad \text{and} \quad q_{ii} = -\lambda_i, \tag{35}$$

we get

$$P'_{ij}(t) = \sum_k q_{ik} P_{kj}(t)$$

Clearly the same equation also holds in case $\lambda_i = 0$ because in that case $P_{ij}(t) = \delta_{ij}$ so that $P'_{ij}(t) = 0$. Thus we have proved the backward equations (33) with q_{ij} for $i, j \in S$ defined by (35).

We now proceed to derive the forward equations (34). Let us first observe that a consequence of the Equations (33) is that

$$P'_{ij}(0) = q_{ij} \quad \text{for all } i, j \in S\,.$$

Of course the derivative at zero is only the derivative from the right, that is,

$$\lim_{h \downarrow 0} \frac{P_{ij}(h) - \delta_{ij}}{h} \quad = \quad q_{ij}\,. \tag{36}$$

By the Chapman-Kolmogorov equations $P_{ij}(t+h) = \sum_k P_{ik}(t) P_{kj}(h)$ so that,

$$\frac{P_{ij}(t+h) - P_{ij}(t)}{h} = \sum_{k \neq j} P_{ik}(t) \frac{P_{kj}(h)}{h} + P_{ij}(t) \frac{P_{jj}(h) - 1}{h}$$

Now letting $h \downarrow 0$ and using (36), one obtains the forward equations. It is in the last step — interchanging the limit and sum — that finiteness of the state space is used. It should be noted that because of the differentiability of $P_{ij}(t)$, the limit $\lim_{h \downarrow 0} \frac{P_{ij}(t+h) - P_{ij}(t)}{h}$ equals $P'_{ij}(t)$ for all $t > 0$.

The matrix $Q = ((q_{ij}))$ is often called the *infinitesimal matrix* or *rate matrix* or *Q-matrix* of the chain. This Q-matrix has the property that all the off-diagonal entries are non-negative and each row sum equals zero. Accordingly the diagonal entries must be non-positive and are determined by the off-diagonal entries. The equation (35) shows that the Q-matrix is determined by the parameters $(\lambda_i, i \in S)$ and $(p_{ij}, i, j \in S, i \neq j)$. What is more important is that the Q-matrix, in turn, determines these parameters. Indeed $\lambda_i = -q_{ii} = \sum_{j \neq i} q_{ij}$, and, for any i, j with $j \neq i$, $p_{ij} = -q_{ij}/q_{ii}$. Of course, if $q_{ii} = 0$, then clearly for each j, q_{ij} is also zero and the above ratio should be interpreted as zero. In a nutshell, the Q-matrix of a chain completely captures the evolution of the chain. The elements of the Q-matrix are often called the *transition rates*, not to be confused with transition probabilities.

A simple but important class of continuous time Markov chains are what are known as *Birth and Death* chains. The state space is $\{0, 1, 2, \ldots\}$. The transition rates are given as follows:

$$q_{i,j} = 0 \quad \text{for all } i, j \text{ with } |i - j| > 1;$$

$$q_{0,1} = b_0; \qquad q_{i,i+1} = b_i, \quad q_{i,i-1} = d_i \quad \text{for } i \geq 1 .$$

It is clear that $\lambda_i = b_i + d_i$, so that the chain starting at i, waits there for an exponential time with mean $1/(b_i + d_i)$, at the end of which it jumps to $i-1$ or $i+1$ with probabilities $d_i/(b_i + d_i)$ and $b_i/(b_i + d_i)$ respectively. If we think of i as population size, then a jump to $(i-1)$ can be treated as death whereas a jump to $(i+1)$ can be regarded as birth. So the population evolves only through a death or a birth. Obviously from size 0, there can only be a birth. The parameters b_i (respectively, d_i) are called the *birth rates* (respectively, *death rates*). The Kolmogorov equations take on a simple form and are often not too difficult to solve. For example, the forward equations will now read

$$P'_{ij}(t) = b_{j-1}P_{i,j-1}(t) + d_{j+1}P_{i,j+1}(t) - (b_j + d_j)P_{ij}(t) .$$

If furthermore $b_i = 0$ for all i, that is, there are no births, the underlying chain is called a *pure death chain.* Clearly, 0 would always be an absorbing state for such a chain. For some special forms of the birth and death rates, the reader may consult the book of Karlin.

0.9.2 Diffusion Processes

To simplify matters, we will assume that the state space of the process is a bounded interval I and, more importantly, that for each t and x, the distribution $P_t(x, \cdot)$ is absolutely continuous with density $p(t, x, \cdot)$. The probability density functions $p(t, x, \cdot)$ — known as the *transition densities* —are then easily seen to satisfy an equation similar to (24) of Section 0.8, namely, that for all $t, s > 0$,

$$p(t+s, x, y) \quad = \quad \int_I p(t, x, z)p(s, z, y)dz . \tag{37}$$

These are the *Chapman-Kolmogorov equations* for transition densities in the continuous time case.

Suppose now that we have a process $(X_t)_{t\geq 0}$ that satisfies, in addition to the above, the following properties:

$$E(X_{t+h} - X_t \,|\, X_t = x) \quad = \quad a(x)h + o(h), \tag{38}$$

$$E(|X_{t+h} - X_t|^2 \,|\, X_t = x) \quad = \quad b(x)h + o(h), \tag{39}$$

$$E(|X_{t+h} - X_t|^k \,|\, X_t = x) \quad = \quad o(h), \qquad \text{for } k \geq 3 . \tag{40}$$

Recall that a function $g(h)$ is said to be of *smaller order* than h, written $o(h)$, if $g(h)/h \to 0$ as $h \to 0$. For subsequent use, let us also recall that $g(h)$ is said to be *at most of the order of* h, written $O(h)$, if $g(h)/h$ remains bounded as $h \to 0$. In both places we are only considering the behaviour near zero.

That the left sides of the equations (38) through (40) are independent of t is, of course, a consequence of the time-homogeneity property. Here $a(\cdot)$ and $b(\cdot)$ are two functions on the state space I and are known as the *drift coefficient*

and *diffusion coefficient* respectively. The equations (38)–(40) can equivalently be expressed in terms of the transition densities as:

$$\int (y-x)p(h,x,y)dy = a(x)h + o(h), \tag{41}$$

$$\int |y-x|^2 p(h,x,y)dy = b(x)h + o(h), \tag{42}$$

$$\int |y-x|^k p(h,x,y)dy = o(h), \quad \text{for } k \geq 3\,. \tag{43}$$

Definition: By a *diffusion process*, we will simply mean a time homogeneous Markov process $(X_t)_{t\geq 0}$ with transtion density $p(t,x,y)$ that satisfies the properties (37) and (41)–(43).

A substantial and mathematically deep theory of diffusion processes exists. See, for example, the book of Bhattacharya and Waymire. One of the major concerns of the theory is to show that, under suitable conditions on the functions $a(\cdot)$ and $b(\cdot)$, a unique diffusion process with required properties exists which, moreover, has nice additional features like, for example, having 'continuous paths'. Further, by imposing suitable conditions on $a(\cdot)$ and $b(\cdot)$, one can also ensure that the transition density of the resulting diffusion is sufficiently smooth in the state variables x and y. However, the mathematical depth of formal diffusion theory is inappropriate at this level, and also, high technical rigour is somewhat unnecessary for our present purposes. Accordingly, we choose not to get into the theory here. We will assume much of what we need and, instead, try to focus on how to apply it. In particular, we assume without question that a unique diffusion process with given drift and diffusion coefficients does exist and that its transtion density $p(t,x,y)$ is twice continuously differentiable in both the state variables x and y.

Before proceeding any further, let us also assume that the state space I of the diffusion process is the unit interval $[0,1]$. Now let g be any twice continuously differentiable function on $[0,1]$ with $g(0) = g(1) = g'(0) = g'(1) = 0$. Using (37) we have

$$\int g(z)p(t+h,x,z)\,dz = \int\int g(z)p(t,x,y)p(h,y,z)\,dy\,dz\,. \tag{44}$$

Using the Taylor expansion of g around y, namely,

$$g(z) = g(y) + (z-y)g'(y) + \frac{1}{2}(z-y)^2 g''(y) + O(|z-y|^3)$$

on the right side of (44), we get

$$\begin{aligned} \int g(y)p(t,x,y)\,dy &+ \int g'(y)[\int (z-y)p(h,y,z)\,dz]\,dy \\ &+ \tfrac{1}{2}\int g''(y)p(t,x,y)[\int (z-y)^2 p(h,y,z)\,dz]\,dy \\ &+ \int p(t,x,y)[\int O(|z-y|^3)p(h,y,z)\,dz]\,dy\,. \end{aligned}$$

Making use of (41)–(43), equation (44) can now be rewritten as

$$\int g(y)[p(t+h,x,y)-p(t,x,y)]\,dy$$
$$=\left[\int g'(y)p(t,x,y)a(y)\,dy+\frac{1}{2}\int g''(y)p(t,x,y)b(y)\,dy\right]h+o(h)\,.$$

Dividing both sides by h and taking limits as $h\downarrow 0$, we obtain

$$\int g(y)\frac{\partial}{\partial t}[p(t,x,y)]\,dy$$
$$=\int g'(y)a(y)p(t,x,y)\,dy+\frac{1}{2}\int g''(y)b(y)p(t,x,y)\,dy\,.$$

Applying integration by parts once on the first term of the right side and twice on the second term, and, using the assumed boundary conditions satisfied by g, we get

$$\int g(y)\frac{\partial}{\partial t}p(t,x,y)\,dy$$
$$=\int g(y)\{-\frac{\partial}{\partial y}(a(y)p(t,x,y))+\frac{1}{2}\frac{\partial^2}{\partial y^2}(b(y)p(t,x,y))\}\,dy\,.$$

Since this equation is valid for all functions g satisfying the assumed conditions, we must have

$$\frac{\partial}{\partial t}p(t,x,y)=-\frac{\partial}{\partial y}(a(y)p(t,x,y))+\frac{1}{2}\frac{\partial^2}{\partial y^2}(b(y)p(t,x,y))\,. \tag{45}$$

This partial differential equation (45) for the transition density function is known as the Kolmogorov's *Forward Equation* or the *Fokker-Planck Equation* and is of fundamental importance in diffusion theory and its applications. A similar equation, called the Kolmogorov's *Backward Equation* for the transition density, can be derived much more easily as follows.

From (37), we have

$$p(t+h,x,y)=\int p(h,x,z)p(t,z,y)\,dy \tag{46}$$

Using the Taylor expansion of $p(t,z,y)$ as a function of z around the point $z=x$, that is, the expansion

$$p(t,z,y)=p(t,x,y)+(z-x)\frac{\partial p(t,x,y)}{\partial x}+\frac{1}{2}(z-x)^2\frac{\partial^2 p(t,x,y)}{\partial x^2}+O(|z-x|^3)$$

on the right side of (46), we get

$$p(t+h,x,y)=p(t,x,y)+\frac{\partial p(t,x,y)}{\partial x}\int(z-x)p(h,x,z)\,dz$$
$$+\frac{1}{2}\frac{\partial^2 p(t,x,y)}{\partial x^2}\int(z-x)^2p(h,x,z)\,dz+\int O(|z-x|^3)p(h,x,z)\,dz\,.$$

Using properties (41)–(43) again, we obtain

$$p(t+h,x,y) - p(t,x,y) = \left\{ a(x)\frac{\partial p(t,x,y)}{\partial x} + \frac{1}{2}\frac{\partial^2 p(t,x,y)}{\partial x^2} \right\} h + o(h)\,.$$

Dividing both sides by h and taking limits as $h \downarrow 0$ leads finally to

$$\frac{\partial p(t,x,y)}{\partial t} = a(x)\frac{\partial p(t,x,y)}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2 p(t,x,y)}{\partial x^2} \tag{47}$$

which is the so called backward equation and will be more useful in the sequel.

We now proceed to show some examples as to how the equation (47) can be used to evaluate certain quantities of interest related to the underlying diffusion. Let us consider, for example, the function

$$F(t,x,y) = \int_0^y p(t,x,z)\,dz, \qquad 0 < x < 1\,.$$

Clearly $F(t,x,y) = P(X_t \leq y \,|\, X_0 = x)$. It follows easily from (47) that the function F satisfies the differential equation

$$\frac{\partial F(t,x,y)}{\partial t} = a(x)\frac{\partial F(t,x,y)}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2 F(t,x,y)}{\partial x^2}\,. \tag{48}$$

This, of course, involves several interchanges of differentiation and integration. But, as mentioned earlier, we will not worry about such technical issues. We will simply put it on record that they can all be justified with some work.

Suppose now that for the diffusion process under study, both the states 0 and 1 are absorbing states. For $i = 0, 1$, let $A_i(t,x)$ denote the probability that the diffusion process starting at state x gets absorbed in state i at or before time t. It is clear then that

$$A_0(t,x) = \lim_{y \downarrow 0} F(t,x,y) \quad \text{and} \quad A_1(t,x) = 1 - \lim_{y \uparrow 1} F(t,x,y)\,.$$

By passing to the limits in (65) as $y \downarrow 0$ or as $y \uparrow 1$ we obtain,

$$\frac{\partial A_i(t,x)}{\partial t} = a(x)\frac{\partial A_i(t,x)}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2 A_i(t,x)}{\partial x^2}\,. \tag{49}$$

It should be noted that though both $A_0(t,x)$ and $A_1(t,x)$ satisfy the same partial differential equation, the solutions would be different (as they should be) because they satisfy different boundary conditions, namely, $A_0(t,0) = 1$ and $A_0(t,1) = 0$ whereas $A_1(t,0) = 0$ and $A_1(t,1) = 1$.

Let us denote by $A_i(x)$, for $i = 0, 1$, the probability that the process starting at the state x ever gets absorbed in the state i. Clearly

$$A_i(x) \quad = \quad \lim_{t \uparrow \infty} A_i(t,x)\,.$$

By a standard result of calculus, since $\lim_{t\uparrow\infty} A_i(t,x)$ exists, $\dfrac{\partial A_i(t,x)}{\partial t} \to 0$ as $t \to \infty$. It thus follows, by letting $t \to \infty$ in (49), that $A_i(x)$ satisfies the differential equation

$$a(x)\frac{dA_i(x)}{dx} + \frac{1}{2}b(x)\frac{d^2A_i(x)}{dx^2} = 0\,. \tag{50}$$

It should again be noted that, although $A_0(x)$ and $A_1(x)$ satisfy the same differential equation, the boundary conditions are different for the two. For $A_0(x)$, for example, the boundary conditions are $A_0(0) = 1$ and $A_0(1) = 0$. Using these, one can easily solve (50) explicitly to get

$$A_0(x) = \frac{\int_x^1 \psi(y)\,dy}{\int_0^1 \psi(y)\,dy}\,, \tag{51}$$

where

$$\psi(y) = \exp\left\{-2\int_0^y \frac{a(z)}{b(z)}\,dz\right\}. \tag{52}$$

Similarly, for $A_1(x)$, using the boundary conditions $A_1(0) = 0$ and $A_1(1) = 1$, one gets

$$A_1(x) = \frac{\int_0^x \psi(y)\,dy}{\int_0^1 \psi(y)\,dy}\,. \tag{53}$$

Of course, $A_1(x) = 1 - A_0(x)$, as it should be.

Having thus obtained simple formulae for the absorption probabilities, let us next turn to the time until absorption. Let τ denote the random variable representing the time until absorption. Let us write

$$A(t,x) = A_0(t,x) + A_1(t,x)$$

where $A_i(t,x)$ are as defined earlier. Then $A(t,x)$ also satisfies the same partial differential equation (49). Notice, however, that $A(t,x)$ is just the probability that $\tau \le t$ given $X_0 = x$; in other words, $A(t,x)$ is the probability distribution function (in t) of τ, conditional on the initial state being x. Suppose now that for each $x \in (0,1)$, this conditional distribution is absolutely continuous with density function $\varphi(t,x)$, $t \ge 0$. Since $A(t,x)$ satisfies the equation (49) we will then have

$$\varphi(t,x) = \frac{\partial A(t,x)}{\partial t} = a(x)\frac{\partial A(t,x)}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2 A(t,x)}{\partial x^2},$$

so that

$$\varphi(t,x) = a(x)\frac{\partial}{\partial x}\left\{\int_0^t \varphi(s,x)\,ds\right\} + \frac{1}{2}b(x)\frac{\partial^2}{\partial x^2}\left\{\int_0^t \varphi(s,x)\,ds\right\}.$$

On differentiating with respect to t (and, of course, assuming again that integration with respect to s and differentiation with respect to x in the above equation can be interchanged) one obtains that

$$\frac{\partial\varphi(t,x)}{\partial t} = a(x)\frac{\partial\varphi(t,x)}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2\varphi(t,x)}{\partial x^2}. \tag{54}$$

Suppose now that we are interested in the mean time till absorption, that is, in

$$T(x) = E(\tau \mid X_0 = x) = \int_0^\infty t\varphi(t,x)\,dt. \tag{55}$$

Let us assume that $t\varphi(t,x) \to 0$ as $t \to \infty$. One then has

$$1 = \int_0^\infty \varphi(t,x)\,dt = [t\varphi(t,x)]\Big|_{t=0}^{t=\infty} - \int_0^\infty t\frac{\partial\varphi(t,x)}{\partial t}\,dt = -\int_0^\infty t\frac{\partial\varphi(t,x)}{\partial t}\,dt.$$

Now using (54) we have

$$1 = -\int_0^\infty t\left\{a(x)\frac{\partial\varphi(t,x)}{\partial x} + \frac{1}{2}b(x)\frac{\partial^2\varphi(t,x)}{\partial x^2}\right\}dt.$$

Assuming once again that the t-integration and x-differentiation can be interchanged, one obtains $T(x)$ to satisfy the ordinary differential equation

$$a(x)\frac{dT(x)}{dx} + \frac{1}{2}b(x)\frac{d^2T(x)}{dx^2} = -1.$$

The obvious boundary conditions now are $T(0) = T(1) = 0$. Using the standard method of integrating factors, one obtains the solution to be

$$T(x) = -2\int_0^x \psi(z)\left(\int_0^z \frac{1}{b(y)\psi(y)}\,dy\right)dz + 2\frac{\int_0^x \psi(z)\,dz}{\int_0^1 \psi(z)\,dz}\int_0^1 \psi(z)\left(\int_0^z \frac{1}{b(y)\psi(y)}\,dy\right)dz, \tag{56}$$

or equivalently

$$T(x) = -2\int_0^x \frac{1}{b(y)\psi(y)}\left(\int_y^x \psi(z)\,dz\right)dy + 2\frac{\int_0^x \psi(z)\,dz}{\int_0^1 \psi(z)\,dz}\int_0^1 \frac{1}{b(y)\psi(y)}\left(\int_y^1 \psi(z)\,dz\right)dy, \tag{57}$$

where ψ is as defined in (52). After some algebra, this solution can equivalently be expressed in the form

$$T(x) \quad = \quad \int_0^1 t(x,y)\,dy, \tag{58}$$

where

$$t(x,y) \quad = \quad \begin{cases} 2A_0(x) \left[b(y)\psi(y) \int_0^y \psi(z)\,dz \right]^{-1} & \text{if} \quad 0 \le y \le x \\ 2A_1(x) \left[b(y)\psi(y) \int_y^1 \psi(z)\,dz \right]^{-1} & \text{if} \quad x \le y \le 1 \end{cases} \tag{59}$$

where $A_i(x)$ are as defined earlier. The above representation is not fortuitous. It can be shown, although we skip it here, that the function $t(x,y)$ has the following interpretation. For $0 \le y_1 < y_2 \le 1$, the integral $\int_{y_1}^{y_2} t(x,y)\,dy$ is the mean time that the diffusion process starting at x spends in the interval (y_1, y_2). In particular, if g is a well-behaved function on the state space, then

$$E\left(\int_0^\tau g(X_s)\,ds \mid X_0 = x \right) \quad = \quad \int_0^1 g(y)t(x,y)\,dy\,.$$

For each fixed non-absorbing state x, the function $t(x,\cdot)$ is what is called the *sojourn time density* of the diffusion starting at the state x.

We end this section here by simply mentioning that it is possible to derive the higher moments of the absorption time — more generally, of $\int_0^\tau g(X_s)\,ds$ — by proceeding in exactly the same way, except that the formulae become complicated.

0.10 References/Supplementary Readings

[1] Bhattacharya, R. N. and Waymire, E. [1990]: *Stochastic Processes with Applications*, John Wiley.

[2] Breiman, L. [1968]: *Probability*, Addison-Wesley.

[3] Chung, K. L. [1974]: *Elementary probability theory with stochastic processes* Academic Press.

[4] Chung, K. L. [2005]: *A course in Probability Theory*, Third edition, Elsevier India Ltd, New Delhi.

[5] Feller, W. [1968]: *An Introduction to Probability Theory and its Applications*, vol. I, Third edition, John Wiley & Sons.

[6] Karlin, S. [1966]: *First Course in Stochastic Processes*, Academic Press.

[7] Karlin, S. and Taylor, H. M. [1975]: *First Course in Stochastic Processes*, Second edition, Academic Press.

[8] Ross, S. M. [1982]: *Stochastic Processes*, John Wiley & Sons.

[9] Ross, S. M. [1998]: *First Course in Probability*, Fifth edition, Prentice Hall International.

For supplementary reading on the material discussed in this chapter the reader may refer to the above books in the manner listed below.
Sections 0.1 through 0.6: [2], [3], [4], [5], [7], [8], [9].
Section 0.7 : [2], [6].
Section 0.8 : [1], [5], [6], [7], [8].
Section 0.9 : [1].