9

An Introduction to Other Types of Stochastic Process

9.1 Markov Processes

9.1.1. The cases treated so far have been considered at some length, this being a convenient way in which to introduce various of the basic notions and most frequently used techniques. They are, however, nothing more than examples of the simplest and most special form of random process; that is, the linear form, or, more explicitly, the homogeneous process with independent increments. We now give some of the basic properties of other cases of interest, although, given the limits of the present work, the treatment will necessarily be brief.

Processes for which 'given the present, the future is independent of the past', or, alternatively, 'the future depends on the past only through the present' are called *Markov processes.* Processes with independent increments are a special case (they are even independent of the present, and the process depends on the latter only through the fact that the future increment, $Y(t) - Y(t_0)$, is added on to the present value $Y(t_0)$); the Markov property is much less restrictive.

The name derives from the fact that Markov considered this property in a particular discrete situation (involving probabilities of 'linked' events, whence *Markov chains*). To give a simple example, let us consider a function Y_n, taking only a finite number of values, 1, 2,..., r, say. For a physical interpretation, which may be more expressive, we could think of it as a 'system' which can be in any one of the r 'states', S_1, S_2,..., S_r, and which passes from one state into another in a sequence of 'steps' (including the possibility that a 'step' could result in the process remaining in the same state: recall, however, what was said in Chapter 8, 8.2.5 concerning the case $\mu_{ii} \neq 0$).

Such a system is said to be a Markov chain if, given that at time n the system is in state i, the probability that it then occupies state j at time $n+1$ is given by some value, $p_{ij}(n)$, which is independent of anything one might know about the past.

The simplest case is that of the *homogeneous* chain, for which the *transition probabilities*, p_{ij}, are also independent of the time n. These probabilities form a matrix, $\boldsymbol{P} = ||p_{ij}||$, and, with the usual definition of matrix product, its square and cube and so on give the analogous matrices of transition probabilities, $p_{ij}^{(2)}, p_{ij}^{(3)}$ and so on for passages from i to j in two steps, three steps and so on. In fact, we have

$$p_{ij}^{(2)} = \sum_h p_{ih} p_{hj} \tag{9.1}$$

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.

(the sum, over h, of the probabilities of going from i to j in two steps when the intermediate state is h) and, in general,

$$p_{ij}^{(m+1)} = \sum_h p_{ih}^{(m)} p_{hj}. \tag{9.2}$$

Of course, the p_{ij} must be non-negative and, for each i, must have sum $\sum_j p_{ij} = 1$.

If the p_{ij} are all non-zero, or if this is the case for the $p_{ij}^{(m)}$ from some given m onwards, we have the so-called *ergodic* case: as m increases, the $p_{ij}^{(m)}$ tend to limit-probabilities p_j, which are independent of i. In other words, for large n, $\mathbf{P}(Y_n = j) = p_j$, independently of one's knowledge concerning the initial state i. Moreover, as n increases, the proportion of the time in which the system occupies state j during the first n steps tends stochastically to p_j (and $1/p_j$ is, in fact, the prevision of the recurrence time; i.e. the time between two successive passages through j).

If, further, we are unaware of the initial situation Y_0, and if our state of uncertainty causes us to attribute precisely these probabilities p_j as the initial, $\mathbf{P}(Y_0 = j)$, then these will remain our probabilities for the occupation of these states throughout the process, and we have what is called a *stationary* process. In fact, the vectors composed of the p_j are characterized by this property of being a fixed point (i.e. an eigenvector with unit eigenvalue) under the transformation P (and, moreover, under the stated conditions it is the unique admissible such eigenvector; i.e. with non-negative components). The ergodic property ensures that, under these conditions, we approach, asymptotically, this stationary situation. The set-up is often applied to statistical problems (involving, for example, a large number of particles or individuals etc.); then the ergodic result has a more concrete interpretation, because it implies the tendency to stationarity of the *statistical distribution.* The reader should compare this situation with those involving illegitimate applications of the ergodic 'principle', outside of the conditions under which the *theorem* holds (see Chapter 8, 8.8.4).

A similar set-up can be obtained in continuous time by assuming that the probabilities of passing from S_i to S_j in the time period from t to $t + \mathrm{d}t$ (given that we are at S_i at time t) are given by $\mu_{ij}\,\mathrm{d}t$, where the μ_{ij} may be constant, or perhaps functions of t. This case has already been considered (Chapter 8, 8.2.5) as background to our discussion of the Poisson process and we shall not add anything here to our previous discussion. As we remarked at that time, we can, without loss of generality, take $\mu_{ij} = 0$ for $i = j$ (and, except for special cases, it is usually convenient to do so).

9.1.2. Within the framework of the very simple cases that formed the basis of our previous discussion, we outlined several of the main problems and features of interest that arise with these processes. The same kinds of problems and features have been studied for general Markov processes and, without going into the details, we shall consider a few of these in order to make some appropriate comments.

The kind of relation which we have encountered in the simple form $p_{ij}^{(2)} = \sum_h p_{ih} p_{hj}$ is typical of the Markov set-up (even in continuous time and with continuous state space). Given that we start from some point P_0 at time t_0, the probability of being in a neighbourhood of a point P_1 at time t_1 satisfies a relation involving the sum (or infinite series, or integral, as the case may be) of the probabilities of getting there by passing

through the various possible points P at some arbitrary, intermediate time t ($t_0 < t < t_1$). These probabilities are evaluated as the product of the probabilities of the two passages: from P_0 to (a neighbourhood of) P in $[t_0, t]$, and then from P to (a neighbourhood of) P_1 in $[t, t_1]$, the latter probability being independent of P_0. This is the probabilistic version of 'Huyghens' principle', by which, in the deterministic case, one regards the evolution of a system in the period from t_0 to t_1 as being the result of what happens between t_0 and t, followed by what happens from t to t_1, starting from the situation reached at t with no need to recall the past. In our case, the same thing applies, not to the evolution of the system, but to the evolution of the probability distribution on the basis of which we foresee the evolution of the system.

In both cases (Huyghens and Markov), these processes are sometimes referred to as *nonhereditary* (in contrast to *hereditary* phenomena, whose evolution is influenced by the past). Examples are provided by the phenomenon of hysteresis, Volterra integral equations and so on. One should note, however, that, in these respects, the distinction between 'present' and 'past' is something of a convention. One often believes that a (deterministic) prediction or a (probabilistic) prevision would be determined by the present if only some (unattainable) data or measurements were known. To compensate for their unavailability, one makes use of available data relating to the past (for example, in the case of hysteresis, the characteristics of the present situation are deduced from the history of the magnetic field which has produced them, since it is impossible to explore the state of magnetization at each point of a body). At an even more basic level, it can happen that for some problems 'the present' can be regarded as the position of a particle (or a body etc.), whereas for others we need, in addition, to know the velocity (or the last movement). This is also true in the probabilistic case and we can consider a *second-order* Markov chain as one in which the probabilities of the possible values for $Y(n+1)$ depend both on the value of $Y(n)$ and of $Y(n-1)$ (but on no others; one could, however, extend the notion and consider chains of arbitrary order). In fact, we could reduce this directly to the first-order case by defining 'the present' at time n to consist of the pair

$$\Big(Y\big(n-1\big), Y\big(n\big)\Big).$$

In other words, we redefine the 'states' to be the r^2 pairs (i, j), with the obvious restriction that from a state (i, j) one can only move, in one step, to one of the r states of the form (j, h).

9.1.3. Although it may seem a 'natural' condition, we are not claiming that the Markov property holds in all 'nonpathological' cases, nor even for the simplest, standard processes. Simple counterexamples that are of practical interest arise in connection with the Poisson process. For example, let $Y(t)$ be the number of telephone conversations in progress on some telephone system at time t and let us assume that $N(t)$, the number of conversations which began between 0 and t, has a Poisson distribution, and that the length of any conversation is a random quantity having the same distribution as all the others, and stochastically independent of them. If this distribution is exponential, the process is Markovian (because every conversation in progress then has the same probability, $\lambda \mathrm{d}t$, of terminating within an infinitesimal time $\mathrm{d}t$, whatever its duration so far) but, in every other case, knowledge of the duration of the conversation so far modifies the prevision. In other examples of this kind, as here, 'age', or something similar, plays a fundamental rôle. A similar kind of example is that where the 'cumulative

effects' have an influence; the prevision at t_0 depends not only on $Y(t_0)$ but also on the sum (or integral, in the continuous case) of the values $Y(t)$ between 0 and t_0. Of course, if the ages or the cumulative values were included as part of the definition of 'the present' (and were observable, or somehow available) then the process, appropriately extended to include these other variables, would turn out to be Markovian.

9.2 Stationary Processes

9.2.1. We have already given the basic idea of a *stationary* process. We discussed it in relation to a Markov process, but this is not a necessary condition for stationarity. A sufficient condition is that the probabilities are invariant with respect to a translation along the time axis. For example, it is sufficient that the probabilities of $Y(t)$, $Y(t+t_1)$,..., $Y(t+t_k)$ satisfying the inequalities $y_i' \leqslant Y(t+t_i) \leqslant y_i''$ are independent of t. The above example of a telephone system (along with similar examples) gives a stationary process if we assume either that the system has been in operation for an infinitely long time or that Y is unknown and that we attribute to the values that it can assume at each instant those probabilities corresponding to the assumption of an infinitely distant beginning. We note that the process is Markovian or non-Markovian according to whether the distributions of the lengths of conversations are or are not independent exponential distributions.

If $Y(t)$ is a stationary process, then the definition implies, in particular, that the distribution of $Y(t)$ does not depend on t, and so neither does the prevision (if it exists) $\mathbf{P}(Y(t)) = m$, nor the variance $\mathbf{P}(|Y(t)|^2) = \sigma^2$ (we assume, for convenience, and without loss of generality, that $m = 0$ and $\sigma = 1$). The same holds for all other moments and parameters of the distribution. If we consider two distinct times t' and t'', the pair of values $Y(t')$ and $Y(t'')$ has a distribution depending only on the difference $t'' - t' = u$,[1] and, in particular, all the quantities defined in terms of this distribution depend only on u; above all, this applies to the *correlation*

$$\phi(u) = r\left(Y(t'), Y(t'')\right) = \mathbf{P}\left(Y(t')Y^*(t'')\right).^2 \tag{9.3}$$

This correlation – usually referred to as the *autocorrelation* function – characterizes the process so far as second-order properties are concerned (in the sense that it enables one to determine $\mathbf{P}(X)$ for all $X = \sum a_{ij} Y_i Y_j^*$; i.e. for functions of the second degree in the values $Y_i = Y(t_i)$ of $Y(t)$ at any number of arbitrary time points t_i). If we have $m \neq 0$ and $\sigma \neq 1$, we can get back to the original process from the standardized case by noting that the former is equal to $m + \sigma Y(t)$. Similar conclusions hold in the nonstationary case, also, provided that $\mathbf{P}(Y(t))$ is constant and

$$\mathbf{P}\left(Y(t')Y^*(t'')\right) = \Gamma\left(t', t''\right) \tag{9.4}$$

---

1 The choice of u is a deliberate attempt to exploit an analogy that makes it convenient to use the same notation, $\phi(u)$, for both the autocorrelation and the characteristic function (see the next section).

2 The asterisk denotes 'complex conjugate'. For the present, it is superfluous, as we are only considering real functions; shortly, however, we shall need the extension to the complex field.

depends only on the difference between t' and t''. Putting $t' = t''$ gives the second moment and if this is bounded so is Γ, and the process is called 'second-order stationary'.

9.2.2. In dealing with this topic it is convenient to allow $Y(t)$ to be complex (for the same reasons for which it is convenient to represent harmonic oscillations with e^{it} rather than sines and cosines). The product $Y(t')Y(t'')$ therefore has to be replaced by the Hermitian product; that is by $Y(t')Y^*(t'')$ (as we already indicated when we defined ϕ and Γ). This implies that

$$\Gamma\left(t'', t'\right) = \Gamma^*\left(t', t''\right),$$

and, in particular, that $\phi(-u) = \phi^*(u)$. The latter is the more important because it relates directly to the stationary case that we are discussing. Moreover, the real part of $\phi(u)$ is continuous if (and only if) the process is 'mean-square continuous' (a stationary process enjoying this property is known as a *Khintchin process*). This property requires, in the notation of Chapter 6, 6.8.3, that $Y(t) \xrightarrow{\bullet} Y(t_0)$ as $t \to t_0$, but – and one should be clear about this – it says nothing about the continuity of the function $Y(t)$. We require that the prevision of $[Y(t) - Y(t_0)]^2$ tends to 0. This happens, for example, for a Poisson process, or variants thereof, even for generalized Poisson processes (these only change through discontinuities, which, in the latter case, are everywhere dense), provided the standard deviation is finite (in this case, in fact, $\mathbf{P}(Y(t) - Y(t_0))^2 = K|t - t_0| \to 0$).

Under these conditions, it can be shown that the class of possible correlation functions coincides with the class of characteristic functions (and, of course, in the case of $Y(t)$ with even-valued, and hence real-valued, characteristic functions, we have

$$\phi\left(-u\right) = \phi^*\left(u\right) = \phi\left(u\right),$$

which correspond, as characteristic functions, to symmetric distributions $F(-x) = 1 - F(x)$). In any case, the distribution F has an important significance so far as the process is concerned, not only from a mathematical point of view but also practically, in all applications, especially to problems in physics, where it has a connection with energy. It gives, in fact, the *spectral function* of the process: that is $F(\omega_2) - F(\omega_1)$ is the prevision of the energy corresponding to the frequencies in the interval $\omega_1 \leqslant \omega \leqslant \omega_2$. Expressed in an informal manner, the actual meaning of this in relation to the random function $Y(t)$ defined by the process is the following: let $U(\omega)$ (in general, complex) be the function expressing $Y(t)$ as a mixture of harmonic components (i.e. as a Fourier–Stieltjes transform)

$$Y\left(t\right) = \int_{-\infty}^{\infty} e^{i\omega t} dU\left(\omega\right). \tag{9.5}$$

The prevision of the energy corresponding to an individual $d\omega$ is

$$dF\left(\omega\right) = \mathbf{P}\left(\left|dU\left(\omega\right)\right|^2\right), \tag{9.6}$$

and in terms of F we obtain the correlation function

$$\phi\left(u\right) = \int_{-\infty}^{\infty} e^{i\omega u} dF\left(\omega\right), \tag{9.7}$$

which is therefore the characteristic function of the energy distribution.

The spectrum F could contain both concentrated masses (jumps of F)

$$U_k = F(\omega_k + 0) - F(\omega_k - 0),$$

corresponding to 'lines' ω_k and diffused masses (segments where F is increasing and continuous). To make things as clear as possible, we repeat and extend the previous discussion in the simpler case where we just have concentrated masses U_k, corresponding to a set of particular frequencies ω_k. In this case, we have

$$Y(t) = \sum_k U_k e^{-i\omega_k t}, \tag{9.5'}$$

and we can deduce that

$$\begin{aligned} U_k &= \lim_{a\to\infty} \frac{1}{2a} \int_{-a}^{a} e^{-i\omega_k t} \sum_h U_h e^{i\omega_h t} dt \\ &= \lim_{a\to\infty} \frac{1}{2a} \int_{-a}^{a} e^{-i\omega_k t} Y(t) dt. \end{aligned} \tag{9.8}$$

The U_k are, therefore, random quantities that depend on the global behaviour of $Y(t)$; conversely, knowledge of these random quantities determines $Y(t)$ in the way we have indicated. To give the probability distribution for all the U_k is an indirect way of giving all the probabilities of the process leading to $Y(t)$. From the energy viewpoint, we could say that the energy for the frequency ω_k is $|U_k|^2$ with prevision $\mathbf{P}(|U_k|^2)$; the U_k are uncorrelated (i.e. 'orthogonal') in the sense that $\mathbf{P}\left(U_h U_k^*\right) = 0$ $(h \neq k)$, and the total energy for frequencies up to and including ω is given by

$$F(\omega) = \sum_k \mathbf{P}\left(|U_k|^2\right) \tag{9.6'}$$

(the sum being taken over all the k for which $\omega_k \leqslant \omega$).

Because we have standardized the process ($\sigma = 1$), the total energy equals 1. The correlation function is given by

$$\phi(u) = \sum_k \mathbf{P}\left(|U_k|^2\right) e^{i\omega_k u}. \tag{9.9}$$

Cramèr and Loève have proved that in the case we have considered (with discrete spectrum) the U_k are mutually orthogonal and that this also holds in the general case (either second-order stationary or Khintchin processes) for the $dU(\omega)$ for disjoint intervals $d_1\omega$ and $d_2\omega$;

$$\mathbf{P}\left(d_1 U . d_2 U^*\right) = 0.$$

Conversely, all such processes can be obtained in this way (a result also proved by Cramèr and Loève).

9.2.3. The concepts and techniques that we have discussed are not only applicable to problems in physics – from which we have borrowed the particularly expressive form

of terminology – but also to problems in other fields, such as statistics ('time-series' analysis and so on). However, this was a good place to give an outline of these ideas; it is useful to be able to 'see' the various problems concerning Fourier transforms and their mathematical properties in terms of some concrete framework. The two applications we have encountered are, in some sense, mutual inverses one of the other. In the case of the characteristic function, the concrete datum, or, at any rate, the most immediate, was the distribution, and the transform mainly provided a 'useful image' of it; in the case we have just dealt with, the function $Y(t)$ and the autocorrelation function are more concrete, and the corresponding distributions U and F are the 'images' in a certain wave interpretation.