1 BASIC SCHEMES IN DISCRETE AND CONTINUOUS TIME

1.1 Formulation of the sequential control problem with incomplete information

Study of various controlled dynamical processes by means of dynamic optimization models is widespread in theoretical and applied science. An important place amongst such models is occupied by stochastic models in which the state of the system (object) at some moment of time using a chosen control does not uniquely define the state of system at the next moment. Rather, for this subsequent state it is only possible to specify a probability distribution which depends both on the previous state and chosen control or, more generally, on all the previous history of the system.

The problem of determining model parameters which are known only incompletely *a priori* plays a central rôle in the formulation and analysis of such dynamical models. This problem forms a part of the more general problem of determining optimal controls for a system with incomplete knowledge of model parameters.

The solution of the general problem is considerably complicated by the fact that while the choice of a control at each moment must be based on existing information about model parameters, the information obtained about these parameters depends, generally speaking, on the chosen control. Such problems are well known in the theory of automatic control. Since control plays a double rôle—on the one hand it is necessary to identify the model parameters, while on the other a definite aim must be accomplished (for example, the maximization of a given functional)—they are sometimes called *dual control* problems. Terms such as *sequential control with incomplete information* and *adaptive control* are also often used in the description of such problems. The use of the last term is connected with the fact that in

process control the controlled system is adapting to control objectives which are only known with *partial* information *a priori*.

The areas of general control theory mentioned above are currently enjoying intensive development. However, owing to the complexity of the problems considered, at present there exist comparatively few sufficiently complete results. Most results obtained so far relate to control models in *engineering*. In this section, we present some *economic* process and event models which lead to the formulation of problems of sequential control with incomplete information.

With regard to control in economics, essentially the simplest area of stochastic modelling whose development has reached the stage of practical applications in planning and control concerns *linear econometric models*. In this field it is assumed as a rule that the stochastic character of models consists in the existence of random disturbances added to some deterministic linear relationship between arguments. Usually *regression* and *correlation analyses* are used as the mathematical tools.

However, these assumptions and methods are simply inapplicable to the analysis of situations in which there are a few essentially different hypotheses (for example, various *expert opinions*) concerning the random nature of parameters of the models. At first glance, a possible way of investigating such a situation involves the use of classical methods of *hypothesis testing* to determine the "true" parameters and the subsequent specification of optimal strategies for the identified parameter values.

Two circumstances prevent this approach. First, sequential control *changes* the characteristics of the observed random values; second, as shown by the solution of some of the problems in the sequel, the process of parameter identification and control based on the identified parameter values cannot always be *separated* into two independent stages without essential *loss*.

An adequate description of such situations, and methods for their investigation, is developed in the general theory of statistical decisions from a *Bayesian* viewpoint. According to this approach, some weights (*relative likelihoods*) are attributed to the unknown *a priori* parameters and the quality of the assessment (chosen decisions) is determined by the *expected* (weighted) *loss*, implied by the statistics for values of

the parameters corresponding to these *a priori* weights. Many studies are devoted to a discussion of the Bayesian approach (see DeGroot 1970; Morris 1971) and, in spite of the fact that the universality of its applicability is to some extent questionable, we will accept it as the basis of the following presentation. Briefly, the necessity of its application to the needs of economics is justified, in our opinion, by the following considerations.

At the first stage in the development of mathematical economic modelling, a possibility of constructing completely formalized models was thought to exist. In recent times, however, it has become obvious that many models must contain *expert opinions* amongst their elements. The problems of coordination of differing expert opinions, of construction of a single estimate from these opinions and of methods for its revision all arise in this context. We stress that by the term "expert opinion" we do not necessarily mean an empirical estimate or opinion of a given person. The term may refer, for example, to data obtained through a certain model or method, and so on. One of the possible ways to reconcile experts' assessments for decision making—or to correct a single estimate with updated information—is to apply the ideas and methods of sequential analysis and the planning of experiments as part of the general theory of statistical decisions.

It is intuitively clear that situations with the following three predominant characteristics: *sequential decision making* (choice of controls), a *stochastic* character to the *evolution* of the system and *incomplete information* about parameters (and system states), are so varied and extensive that a general formulation concerning them would be extremely cumbersome and impracticable—if not impossible.

A narrower type of economic decision problem—which may be described by simpler mathematical models and precisely analysed—can be treated if, in addition, the following conditions apply.

First, we assume that the *decision* at each moment consists of a choice of one of the *controls* $a^1, \ldots, a^m$, or in the choice of an m-dimensional vector α with nonnegative coordinates $\alpha^1, \ldots, \alpha^m$, $\sum_{j=1}^m \alpha^j = 1$ which prescribe the *mixed* (i.e. *randomized*) *use* of controls $a^1, \ldots, a^m$ with corresponding *weights* $\alpha^1, \ldots, \alpha^m$. In other words, the decision concerns the *distribution* of a single resource at each moment in time between m possible uses. In each practical situation it is neces-

sary to define precisely how one should understand the use of a control with a weight different from 0 or 1.

Secondly, as a result of the choice of the control a^j a random variable with distribution function given by $F^j(x)$ is observed whose values do not depend on the values of the random observations at past steps. Such a *Markov property* (independence of the future from the past for a fixed present) is, of course, an additional constraint. However, it is well known that models with a finite dependence on the past can be considered to fall under the situation discussed above, although the appropriate embedding leads to a considerable increase in the dimension of the problem.

Finally, the third condition concerns a finite number of *hypotheses* $H_1, \ldots, H_N$ with regard to the distribution functions of the random variable observed upon the choice of one of the controls $a^1, \ldots, a^m$. Under the hypothesis H_i the random variable observed with choice a^j has a distribution function F_i^j, where the F_i^j, $i = 1, \ldots, N$, $j = 1, \ldots, m$, are supposed known.

Now we turn to economic applications and describe informally three examples of simplified mathematical economic models which have the above-mentioned properties.

The first example is related to the choice of selling policy for some good. Such a selling policy is defined by the choice at every moment of time of some *condition of sale* depending on the past history of sales. If we speak about the sale of a specific good, then the conditions of sale may be a fixed price, or a fixed production volume for the market, or a combination of a price and packaging of a fixed type, and so on. Suppose that m different possible conditions of sale exist and we have N experts. The i^{th} expert declares what, in his opinion, will be the distribution function F_i^j of sales in a fixed accounting period under condition of sale j, $j = 1, \ldots, m$, i.e. $F_i^j(x)$ gives the probability that the sales volume does not exceed x under the j^{th} condition of sale.

The possible combinations of conditions of sale and *expert opinions* define an $m \times N$ *hypothesis matrix* $\{F_i^j\}$ whose elements are probability distribution functions. At time 0, an *a priori weighting* $\xi_1, \ldots, \xi_N$ of the *experts* is given by considerations outside the model. A *sales strategy* is naturally defined by the rule which for each time t, $t = 1, \ldots, T$, specifies the condition of sale on the time interval

$[t, t+1)$ on the basis of the *a priori* experts' weighting, the hypothesis matrix and the actual sales and chosen controls at all previous time periods. It is assumed that the sales in each such interval $[t, t+1)$ do *not* depend on the corresponding sales in previous intervals, but depend only on the control chosen (it is justified if the good is not storable, e.g. goods or services). The aim of the firm is the maximization of profit on sales over some time interval.

From general theory it follows that to make a decision at each period it is not necessary to remember all previous statistics. At each period it is sufficient to update the *a posteriori* experts weighting in terms of sales in the previous interval and then to make a decision based on these updated values. We emphasize that the experts' opinions (the rows in the hypothesis matrix) remain *fixed.* The results of observation will eventually show the most correct expert and the decision maker will then tend to follow his advice. Models of this type were initially studied by Rothschild (1974); see also §7.5. Models in which the hypothesis matrix depends upon the observation process may also be formulated in the present structure, but in this case the search for optimal strategies becomes considerably more complicated.

Next we will describe a second economic situation whose mathematical formulation leads to the same structure. This model is presented in detail in Sonin (1976).

Suppose that the realization of some scientific or technological programme is possible through several different *development methods* (for example, several research projects or institutions) which may be operated in parallel, and that the total programme may be divided into a specified number of *stages.* The *completion time* of each stage, for example by the i^{th} method, is considered to be a random variable, whose distribution depends, on the one hand, on some internal characteristic of the i^{th} method and, on the other, on the amount of resources involved and their distribution over time. It is naturally assumed that *a priori* only partial information about the various development method parameters is known and that this information is obtained gradually and in relation to the intensity of their implementation, i.e. with the amount of resources applied to each development method. Using the Bayesian approach, we assume that there exist N *hypotheses* (experts' opinions) regarding the characteristics of the development methods

and *a priori* probabilities (*weights*) of these hypotheses. At each period a single divisible resource is distributed between the development methods. It is required to find the allocation of the resource over a fixed number of periods which maximizes one of the following functionals: (a) the *probability* of completion of a given number of stages, or (b) the *average* number of stages completed.

The model just described allows a wider interpretation if "development methods" are understood in a broader sense as a complex multistage programme. The mathematical structure described above may serve as the simplest model of changing priorities and redistribution of resources in a multistage programme.

The last model of the type under consideration is formulated in terms of a classical *search problem.* We have m cells, and there are one (or more) objects known to be in them. Usually, the *a priori* probabilities of the object being in each cell are considered to be given, together with the probability of *discovering* the object by a *search* process (only one cell can be searched in each period). Suppose that the *a priori* information has a more complicated character, precisely that we have N *hypotheses* and the i^{th} hypothesis states that the probability of discovery of the object in the j^{th} cell in a single search equals λ_i^j. (The classical problem is obtained if $\lambda_j^j := q^j$, $\lambda_i^j := 0$ for $i \neq j$.) It is required to maximize the probability of discovering the object within a fixed time. We will not elaborate the interpretations of such problems, but mention here only that such search problems model some problems of technical diagnostics, industrial maintenance, and so on (see, for example, Rastrigin 1968; Stone 1975).

Some other examples of situations whose modelling leads to the same mathematical structure are considered in the studies mentioned in Chapter 7 (see §§7.5 and 7.6).

1.2 Basic schemes in discrete time

The mathematical model can be described as follows. We suppose given a *matrix of probability distribution functions* $\{F_i^j\}$ with rows $i = 1, \dots, N$ and columns $j = 1, \dots, m$, a *control set* $\{a^1, \dots, a^m\}$ and a set of *hypotheses* $\{H_1, \dots, H_N\}$. The distribution function F_i^j corresponds to the control a^j and to the hypothesis H_i. At each

time $n = 1, \dots, \nu$, the decision maker (statistician) chooses either one of the controls $a^1, \dots, a^m$ or a probability distribution over the control set according to which a *realization* of the control occurs. If the hypothesis H_i is true and the control a^j is realized, a random variable $X(n)$ is observed with distribution F_i^j which is *independent* of previous observations and controls. In choosing a control, the statistician knows the vector of *a priori* probabilities ξ in $(N-1)$-dimensional simplex $S^N := \{\xi = (\xi_1, \dots, \xi_N) : \xi_i \geq 0,\ i = 1, \dots, N,\ \sum_{i=1}^N \xi_i = 1\}$, the matrix of distribution functions $\{F_i^j\}$ (the *hypothesis matrix*) and the values of previous controls and observations $a(1)$, $X(1), \dots, a(n-1)$, $X(n-1)$.

A function defined for all $n = 1, 2, \dots, \nu$ on the system *history* up to the present time n, i.e. on the vectors $h(n-1) = (a(1), X(1), \dots, a(n-1), X(n-1))$, with values in the set of probability distributions over the control set $a^1, \dots, a^m$, or equivalently in the $(m-1)$-dimensional simplex S^m, is called a (*randomized*) *action rule.* If each distribution is concentrated on a single control, i.e. on a vertex of the simplex S^m, then the action rule is termed *nonrandomized.* For future reference, the set of vertices of the simplex S^m will be denoted by $\widehat{S}^m$.

Sometimes it will be useful to employ the following informal interpretation and related terminology for simplicity. We will say that we have m *devices* generating random values. A choice of control a^j corresponds to the *use* of device j and, under the i^{th} hypothesis, to the observation of random values with distribution F_i^j on the j^{th} device.

Under hypothesis H_i the fixed *action rule* β generates the probability distribution designated by P_i^β on the space of controls and observations. For a given ξ the distribution $P_\xi^\beta = \sum_{i=1}^N \xi_i P_i^\beta$ corresponds to the action rule β. The corresponding expectations are designated by E_i^β and E_ξ^β.

Besides the vector ξ and the hypothesis matrix $\{F_i^j\}$ the criterion function $f_n(i, h(n))$, interpreted as the *profit* (*cost*) at time n if hypothesis H_i is true and $h(n)$ is the *history* of the system up to time n, is considered to be known. For fixed ξ the purpose of control is to choose the action rule β which maximizes (minimizes) the expected value of the resulting profit (cost) for ν steps,

$$\sup_\beta E_\xi^\beta \sum_{n=1}^{\nu} f_n(i, h(n)),$$

where $\nu \leq \infty$ is often termed the *horizon* of the problem.

A corresponding optimal expected value of total profit (cost) in a maximization (minimization) problem is called the *value function* $F_\nu(\xi)$ and the action rule at which the extremum is reached, if it exists, is called the *optimal rule* for the point ξ.

Further, as a rule we consider the case when the distribution functions F_i^j are *Bernoulli*, i.e. are of random variables with values *success* and *failure* (1 or 0). In this case it suffices to replace the matrix $\{F_i^j\}$ of functions by the matrix $\lambda = \{\lambda_i^j\}$ of numbers, where λ_i^j is the probability of success under the i^{th} hypothesis using the j^{th} control. We term this model the *basic (discrete time) scheme*. The *value function* corresponding to maximizing the *number of successes* in ν steps for the *a priori* distribution ξ is designated by $V_\nu\{\xi\}$.

We conclude this section by formulating one of the simplest special cases of the basic scheme, the *two-armed bandit* problem mentioned in the introduction, which will be used to illustrate many facts presented in this chapter. Only two controls and two hypotheses exist. The matrix is symmetric with regard to both diagonals, i.e. $\lambda_1^1 = \lambda_2^2 = \lambda^1$, $\lambda_1^2 = \lambda_2^1 = \lambda^2$, $0 \leq \lambda^1 < \lambda^2 \leq 1$. It is required to maximize the number of appearances of 1 in a sequence of ν trials, so that the profit function $f_n(i, X(n)) := 1$ if $X(n) = 1$ and $:= 0$ if $X(n) = 0$, $n \leq \nu$.

1.3 Relation of the basic scheme to the general theory of sequential control with incomplete information

The basic scheme described in §1.2 relates to the general theory of control with *incomplete information* (see Shiryaev 1967; Dynkin & Yushkevich 1976; Schäl 1979). For the case of Markov stationary processes in discrete time, the Bayesian approach to the formulation of the problem has the following form.

At each moment of (discrete) time the state of the system is described by pairs (y, x); the second element x corresponds to the *observed* component of the system, and the first element y to the *unobserved*. Suppose given a family of *transition probabilities* $p^a(y', x'|y, x)$ defining the probability distribution of the position of the system at the next moment and depending on the previous state (y, x) and a control a chosen from some set of *admissible controls*. At each moment the choice of control is made on the basis of knowledge of the *a priori* distribution of the state of unobserved components, on the observed states and on the previously used controls. The purpose of control (in the *additive* case) consists in a choice of *strategy* which gives a minimum (maximum) in expectation of the resulting criterion

$$\sum_{n=1}^{\nu} f_n(y(n), x(n), a(n)) \, .$$

In the case of the basic scheme of §1.2, the rôle of the unobserved component is played by the index i of the hypotheses $i = 1, \ldots, N$, and that of the observed component by the values of random variables $X(n)$. So, we have a *specific* case of control with incomplete information in which the state of the unobserved component does *not* change with time. This variant of the general structure of control with incomplete information is presented in §2.2. Further, in problems with this structure the distribution $p(\cdot|\cdot)$ does not depend on the previous state but, by virtue of the independence of observations for a fixed control, only on the control.

The results of general control theory with incomplete information (the Markov additive case) are thus applicable to the basic scheme.

The most important consequence of this observation is that this control problem with *incomplete* information is equivalent to the familiar problem of control of a *Markov chain*, i.e. to a stochastic control problem with *complete* information. (The latter formulation may be derived from that described above with incomplete information if it is supposed that the unobserved component is absent.) The pairs (*a posteriori distribution* of the unobserved component, observed component) serve as the states of an equivalent Markov chain. The *a posteriori distribution* is calculated from all the previous history of the system. In our case when, as mentioned above, the transition probabilities and cost functions do not depend on the previous state of the observed component, the states of the equivalent Markov chain are simply the *a posteriori* probabilities of the hypotheses. In other words, the *a posteriori* hypothesis probabilities together with the last previous control are Markov *sufficient statistics*.

1.4 Evolution of *a posteriori* probabilities in discrete time

Let $\xi(n)$ be an N-dimensional vector of *a posteriori* probabilities at time n. Using the Bayesian formula it is easy to show how to compute $\xi(n+1)$ if $\xi(n) := \xi$, $a(n) := a^j$ and $X(n+1)$ is observed. We give these formulae for the basic scheme, *viz.*

$$\xi_i(n+1) = \xi_i\lambda_i^j / \sum_{k=1}^{N} \xi_k\lambda_k^j \qquad \text{if } X(n+1) = 1,$$

$$\xi_i(n+1) = \xi_i(1-\lambda_i^j) / \sum_{k=1}^{N} \xi_k(1-\lambda_k^j) \quad \text{if } X(n+1) = 0. \quad (1.1)$$

The denominator of the first fraction is equal to the probability of a realization of a 1 upon applying the control a^j with current value ξ of the vector of hypothesis probabilities. We will designate this probability as $p^j(\xi)$. According to the discussion in §1.3, the fundamental problem is equivalent to the control of a Markov chain whose states are the vectors ξ belonging to the $(N-1)$-dimensional simplex S^N, with transition probabilities $p^a(\xi'|\xi)$ such that with $a := a^j$ the system makes a transition to the state $\Gamma_1^j\xi$ defined by the first formula of (1.1) with probability $p^j(\xi)$ and to the state $\Gamma_0^j\xi$ defined by the second formula of (1.1) with complementary probability $1 - p^j(\xi)$.

The evolution of *a posteriori* probabilities in this system may be described by a random process whose state $\xi(n+1)$ at time $(n+1)$ is determined by $\xi(n)$, the control realized at time n and the value of the observation at time $n+1$. It is well known from control theory with incomplete information (and it is also easy to check directly) that for any action rule the process $\xi(n)$ is a *martingale* (with respect to the measure P_ξ^β), i.e. the conditional mean of the increment of the process equals 0. Another important fact—also well known in mathematical statistics—is that under a fixed hypothesis H_i, $i = 1, \ldots, N$ (or, equivalently, for the measure P_i^β) the process $\xi(n)$ will "usually" exhibit *consistency* as $n \to \infty$, as a result of which $\xi_i(n) \to 1$ and $\xi_r(n) \to 0$, $r \neq i$. The expression "usually" relates to the fact that if, for example, an action rule is used which at each step chooses the control a^j and $\lambda_i^j = \lambda_s^j \neq \lambda_r^j$ for $r \neq i, s$, then for the true hypothesis H_i (or measure P_i^β), $\xi_r(n) \to 0$ for $r \neq i, s$ and $(\xi_i(n)/\xi_s(n)) = \text{const}$ holds. If all elements of the matrix $\{\lambda_i^j\}$ are different, then for any i with respect to the measure P_i^β the vector $\xi(n)$ always converges to the vector e_i^N, where e_i^N is the N-dimensional vector $(0, 0, \ldots, 1, 0, \ldots, 0)$, with a 1 in the i^{th} position. Such a situation is naturally called *discrimination* of hypotheses. As mentioned above, this phenomenon will *not* occur for *all* matrices (these questions will be discussed in §2.5).

Further, in many cases we will make a transformation of the coordinates of ξ to the coordinates $\eta := (\eta_1, \ldots, \eta_N)$ related to ξ by *one* of the N one-to-one relations (i.e. for *some* $s = 1, \ldots, N$)

$$\eta_i := \ln(\xi_i/\xi_s), \quad \xi_i := \exp\eta_i / \sum_{k=1}^{N} \exp\eta_k, \quad i = 1, \ldots, N. \quad (1.2)$$

The vector η represents the *logarithmic maximum likelihood function* often used in mathematical statistics. In our case the transformation to the coordinates η is convenient since the increment of the process $\eta(n)$ depends only on the value of the observation $X(n)$ and does *not* depend on the value $\eta(n-1)$, unlike the increments of $\xi(n)$ which *do* have such a dependence.

A notion of strategy is useful when the transition is made from the basic scheme in original form to the controlled Markov chain. A *strategy* is a function of previous controls and previous states of the vector $\xi(n)$ (but it is *not* a function of previous observations). Each strategy corresponds to a whole *set* of action rules β_ξ defined for *all* possible values of the initial point ξ.

An important rôle among strategies is played by the *Markov* strategies or, in terminology of optimal control, strategies given in *feedback* form, i.e. by a function defined on pairs (ξ, n) and taking values in S^m. In this case, the control equals $a(\xi(n), n)$ at time n. The important rôle of Markov strategies in the theory of controlled Markov chains derives from the fact that for a wide set of assumptions, one of which surely holds in our case, it is sufficient to search for the optimal strategy in this class.

In some studies devoted to the problems deriving from the basic scheme, a formulation is immediately given in terms of the control problem of a Markov chain whose states are *a posteriori* probabilities and only Markov strategies are considered as admissible. However,