

CHAPTER 1

BASIC NOTIONS AND DEFINITIONS

In this chapter the basic definitions and notions of control theory such as a mathematical model, a controlled random process, an observable process, a strategy and a control aim are considered. They serve as the basis for the definition of the adaptive strategies. Some auxiliary notions are considered as well. The general approach to the synthesis of the adaptive strategies and their properties is discussed. Specific features of classic and adaptive control are also discussed.

1.1. Random Processes and Systems of Probability Distributions

We shall begin with the definition of a probability space $(\Omega, \mathfrak{F}, \mathbf{P})$ where Ω is the space of elementary events, \mathfrak{F} is the σ -algebra of measurable subsets from Ω , i.e. a class of the subsets of Ω closed with respect to complements, products and countable sums of these sets and, finally, \mathbf{P} is a probability measure, i.e. it is a non-negative, countably-additive function defined on \mathfrak{F} (i.e. $\mathbf{P}\{A\} \geq 0$ for any $A \in \mathfrak{F}$ and for disjoint sets A_i from \mathfrak{F} we have $\mathbf{P}\{\bigcup_1^\infty A_i\} = \sum_1^\infty \mathbf{P}\{A_i\}$ and $\mathbf{P}\{\Omega\} = 1$.)

Let (X, \mathfrak{X}) be a pair consisting of a measurable space X and a σ -algebra of measurable subsets \mathfrak{X} . In the most interesting cases the space X is Euclidean, i.e. $X = \mathbf{R}^l$, $l = \dim \mathbf{R}^l$. A measurable mapping $\xi : \Omega \rightarrow X$ is called a *random variable* (r.v. for short). It means that the inverse image of any measurable set from \mathfrak{X} belongs to \mathfrak{F} under the mapping ξ or, symbolically, $\xi^{-1}(M) \in \mathfrak{F}$ for all $M \in \mathfrak{X}$. If Ω is a topological space we can consider the smallest σ -algebra \mathfrak{B}_Ω containing all open sets from Ω . It is called the *Borel σ -algebra* and any r.v. ξ which is measurable with respect to \mathfrak{B}_Ω is called a *Borel random variable*. The measure \mathbf{P} on the probability space $(\Omega, \mathfrak{F}, \mathbf{P})$ defines the probabilities of the events pertaining to the r.v. ξ . For example, the event $\xi \in M$ ($M \in \mathfrak{X}$) occurs with probability $\mathbf{P}\{\xi \in M\} = \mathbf{P}\{\omega : \xi \in M\}$. It is convenient to define a measure on \mathbf{R}^1 for the scalar r.v. ξ by using the distribution function $F(x) = \mathbf{P}\{\xi(\omega) \leq x\}$. For the multi-dimensional r.v. this can be done in a similar way.

Using the measure \mathbf{P} the mathematical expectation of ξ is defined as the Lebesgue integral

$$\mathbf{E}\xi = \int_{\Omega} \xi(\omega)\mathbf{P}\{d\omega\}.$$

If $\xi \in \mathbf{R}^1$ we can rewrite this formula as follows

$$\mathbf{E}\xi = \int_{-\infty}^{\infty} x dF(x).$$

Let the variable t be interpreted as time. We shall distinguish discrete and continuous time. In the first case the parameter t is running over the set $\{0, 1, 2, \dots\}$ (sometimes $\{0, \pm 1, \pm 2, \dots\}$) but in the second one t belongs either to a finite interval $[0, T]$ or to the half-axis $[0, \infty)$.

A family of r.v. $\xi_t(\omega)$ in continuous or discrete time is called a *random process* (r.p. for short). The r.p. will be called *non-terminating* if it is defined on an unbounded interval.

For a fixed ω , $\xi_t(\omega)$ is a function of time t which is called a *trajectory* or a *path* of this process. For a fixed t , $\xi_t(\omega)$ is a r.v. which represents the value of this process at time t . The trajectories $\{\xi_t(\omega), \omega \in \Omega\}$ will form a function space and a space of sequences when t is continuous and discrete respectively.

The probabilistic properties of a r.p. are defined by the measure \mathbf{P} given on the appropriate probability space $(\Omega, \mathfrak{F}, \mathbf{P})$. This can be defined in various ways. At first we shall consider the construction of the scalar r.p. by using the family of finite-dimensional distribution functions associated with $\xi_t(\omega)$

$$\mathbf{P}\{\omega : \xi_{t_1}(\omega) \leq x_1, \dots, \xi_{t_m}(\omega) \leq x_m\} \stackrel{\text{def}}{=} F_{t_1 \dots t_m}(x_1, \dots, x_m)$$

for all positive integers m and $t_1 < \dots < t_m$. The distributions from this family have the following consistency properties:

1. For any $x_1, \dots, x_m \in R^1$, $m < l$ and $t_1 < t_2 < \dots < t_m < \dots < t_l$

$$F_{t_1 t_2 \dots t_l}(x_1, \dots, x_m, \infty, \dots, \infty) = F_{t_1 t_2 \dots t_m}(x_1, \dots, x_m);$$

2. If i_1, \dots, i_m is any transposition of the index-set $1, 2, \dots, m$ then

$$F_{t_{i_1} t_{i_2} \dots t_{i_m}}(x_{i_1}, \dots, x_{i_m}) = F_{t_1 t_2 \dots t_m}(x_1, \dots, x_m).$$

Properties 1 and 2 are sufficient for the existence of a unique probability measure coinciding with the distributions $F_{t_1 t_2 \dots t_m}(x_1, \dots, x_m)$ given on the cylindrical sets (Kolmogorov Theorem). Moreover, there is a probability space $(\Omega, \mathfrak{F}, \mathbf{P})$ and r.p. $\xi_t(\omega)$ defined on it (where t is running over the same set from which t_1, \dots, t_m are taken) which corresponds to this family so that

$$\mathbf{P}\{\omega : \xi_{t_1}(\omega) \leq x_1, \dots, \xi_{t_m}(\omega) \leq x_m\} = F_{t_1 t_2 \dots t_m}(x_1, \dots, x_m). \quad (1)$$

The probability measure \mathbf{P} can be constructed in a standard way, i.e. first it is defined on the cylindrical sets of the form

$$M = \{\omega : \xi_{t_1}(\omega) \leq x_1, \dots, \xi_{t_m}(\omega) \leq x_m\}, \quad t_1 < \dots < t_m \quad (2)$$

by (1) and then it is extended to the σ -algebra generated by these sets. If the sets in (2) are open with respect to a topology of the given function space then the σ -algebra will be a Borel σ - algebra.

On the other hand, an r.p. in continuous time can be defined by using the family of conditional probability measures

$$\mu_t(M|\xi_s, 0 \leq s < t) = \mathbf{P}\{\xi_t \in M|\xi_s, 0 \leq s < t\}, \quad \forall M \in \mathfrak{X}$$

where the past history of the process up to time t is considered as the condition. Often, these conditional probabilities are assigned by using an increasing flow of the σ -algebras \mathfrak{F}_t (i.e. for any $t_1 < t_2$ the inclusions $\mathfrak{F}_{t_1} \subseteq \mathfrak{F}_{t_2} \subseteq \mathfrak{F}$ take place). The σ -algebra \mathfrak{F}_t is generated usually by the r.p. past history, i.e. by the system of the following sets

$$M_{t,a} = \{\omega : \xi_s(\omega) < a, s \leq t, a \in \mathbb{R}^1\}.$$

Then one introduces a family of conditional probabilities $\mu_s(\cdot | \mathfrak{F}_{t-})$ where \mathfrak{F}_{t-} means the left-hand limit of the σ -algebras, i.e. $\mathfrak{F}_{t-} = \lim_{s \uparrow t} \mathfrak{F}_s$.

In the discrete time case the r.p. can be defined by using a family of finite-dimensional distribution functions $F_{t_1 \dots t_m}(\cdot)$ or by a family of conditional probabilities $\mu_{t+1}(M | \xi_0, \xi_1, \dots, \xi_t)$. Further the abbreviation $\xi^t \stackrel{\text{def}}{=} (\xi_0, \xi_1, \dots, \xi_t)$ (the past history of the process ξ_t up to time t) will be used to write down this conditional probability as $\mu_{t+1}(M | \xi^t)$. In what follows we shall assume that the above mentioned conditional probabilities satisfy the following conditions:

1. For any past history ξ^t the function $\mu_{t+1}(M | \xi^t)$ is a probability measure with respect to argument M ;
2. For a fixed M this function is measurable with respect to the variables ξ_0, \dots, ξ_t .

Using the family of measures $\{\mu_t\}$ one can write down the representations of the finite-dimensional distributions, i.e. an r.p. is given. In what follows we shall prefer to define an r.p. by means of a family $\{\mu_t\}$ which defines the measure uniquely on the state space of the process ξ_t , i.e. on the space of sequences $(\xi_0, \xi_1, \dots, \xi_t, \dots)$. Now the existence problem for an r.p. with given characteristics arises.

Let two sequences of measurable spaces $(\Omega_t, \mathfrak{F}_t)$ and (X_t, \mathfrak{X}_t) , $t = 1, 2, \dots$ be given. A probability measure \mathbf{P}_1 and a conditional measure $\mathbf{P}(\cdot | \omega_1, \dots, \omega_{t-1})$ are defined on $(\Omega_1, \mathfrak{F}_1)$ and $(\Omega_t, \mathfrak{F}_t)$ respectively. For any $A \in \mathfrak{F}_t$ let the function $P(A | \omega_1, \dots, \omega_{t-1})$ be Borelian with respect to the arguments $\omega_1, \dots, \omega_{t-1}$. Let us put for $A_j \in \mathfrak{F}_j$, $j \geq 1$

$$P_t(A_1 \times \dots \times A_t) = \int_{A_1} \mathbf{P}_1(d\omega_1) \int_{A_2} \mathbf{P}_2(d\omega_2 | \omega_1) \dots \int_{A_t} \mathbf{P}_t(d\omega_t | \omega_1, \dots, \omega_{t-1}).$$

Under these conditions the existence problem of a r.p. given on the space $(\Omega, \mathfrak{F}) = \prod_{t \geq 1} (\Omega_t, \mathfrak{F}_t)$ with the paths from $(X, \mathfrak{X}) = \prod_{t \geq 1} (X_t, \mathfrak{X}_t)$ and with the probabilistic characteristics given beforehand is solved by the following theorem:

Theorem 1. (Jonescu Tulcea) *In the space (Ω, \mathfrak{F}) there exist an unique probability measure \mathbf{P} and a process $\xi = (\xi_1(\omega), \xi_2(\omega), \dots)$ such that*

$$\begin{aligned} \mathbf{P}\{\omega : \omega_1 \in A_1, \dots, \omega_t \in A_t\} &= \mathbf{P}_t(A_1 \times \dots \times A_t), \\ \mathbf{P}\{\omega : \xi_1(\omega) \in A_1, \dots, \xi_t(\omega) \in A_t\} &= \mathbf{P}_t(A_1 \times \dots \times A_t) \end{aligned}$$

for all t , $A_j \in \mathfrak{F}_j$.

Sometimes we have to consider the scalar functions (the functionals) $\varphi_t = \varphi(\xi^t)$ defined on the paths of the r.p. which take values from some measurable space (X, \mathfrak{X}) . These functions are measurable mappings $\varphi : X^t \rightarrow \mathbb{R}$. The sequence φ_t is also an r.p. for which it can be found, as a matter of principle, the family of the conditional distributions (or the finite-dimensional distribution functions) provided one is known for the original process. But it is rarely that such calculations can be done in an explicit form. For the mathematical expectation $W(t) = \mathbf{E}\varphi_t$ we have

$$W(t) = \int_{X^t} \varphi(x_1, \dots, x_t) \mu(dx_t | x^{t-1}) \mu(dx_{t-1} | x^t) \cdots \mu(dx_1).$$

It is often necessary that the parameters specifying the distribution of the r.v. (for example, in the case of the binomial distribution it is the probability of the success denoted by p usually; in the case of the normal distribution they are the mathematical expectation and the variance denoted by (m, σ) respectively) be included in the notation of this distribution. This is a standard situation in the theory of the random processes. Such parameters are indicated in the notations of the corresponding conditional probabilities or the distribution functions, i.e. $\mu_t^{(\theta)}$ or $F_{t_1, \dots, t_m}^{(\theta)}$, or, generally speaking, $P^{(\theta)}$. In control theory importance is attached to the situation when there exists a time-varying variable, called a control and usually denoted by u_t , such that the probabilistic characteristics of the process depend on it. We shall now discuss this in more detail.

Let, in addition to (X, \mathfrak{X}) , a measurable space (U, \mathfrak{U}) be given. The points u from U are called the *controls*. Let the conditional probabilities have the form

$$\mu_{t+1}(M | \xi^t, u^t), \quad t \geq 1, \quad M \in \mathfrak{X}.$$

and the following conditions are implemented:

1. The functions μ_{t+1} are the probability measures on X with respect to the first argument for all ξ^t and u^t .
2. For any $M \in \mathfrak{X}$ they are measurable with respect to $(\xi^t, u^t) \in X^t \times U^t$.
3. The functions μ_{t+1} depend on u_t essentially.

We shall call the conditional measures from the family $\{\mu_t, t \geq 1\}$ submitted to these conditions the *controlled conditional measures (distributions)*. Such families given on the space (Ω, \mathfrak{F}) define a class of r.p. taking the values from (X, \mathfrak{X}) with the controls from (U, \mathfrak{U}) . To select some r.p. it is necessary either to fix in advance the sequence of controls or to give the rules of their choice in the course of the r.p. evolution.

Definition 1. The family of conditional distributions $\{\mu_t, t \geq 1\}$ or, which is the same, the class of r.p. associated with it is called the *model of the controlled object*.

Such models will be denoted by ξ, ζ or η .

1.2. Controlled Random Processes

First we shall assume that time is discrete. Let a measurable space (Ω, \mathfrak{F}) be the space of the elementary events, (X, \mathfrak{X}) be the *state space* and (U, \mathfrak{U}) be the *space of controls* (or *actions*). We suppose that a model of the controlled object $\{\mu_t, t \geq 1\}$ is given as well. The r.p. state will be denoted by x_t , i.e. x_t is the value of the r.p. generated by this model at time t . The state x_t may be accessible to observation and measurement but this does not always occur. We can often judge about it only by the circumstantial evidence or by partial observations. Indeed, it is true in the case of many technological, physical and chemical processes where the direct measurement of all components is impossible. By reason of this we are forced to introduce a variable z_t to denote data observed at time t .

We shall consider z_t as a process given onto a measurable space (Z, \mathfrak{Z}) which is called the *space of observations*. The evolution of the process z_t is defined by the family of conditional measures $\{\nu_t, t \geq 1\}$ on Z where the function $\nu_t = \nu(H|x^t, z^{t-1}, u^{t-1})$, $H \in \mathfrak{Z}$ is measurable with respect to the variables included into the condition. The generation of the observations up to a moment t , i.e. $z^t = \{z_s, s \leq t\}$ is called the *history of the observable process* up to the moment t . Another name is an *information image of the model*.

The space Z can be a subspace of X . It means that the object is partially observed, i.e. a piece of information about the states of the process has been lost. If the spaces X and Z have no common points we shall have to judge about the states of the model by using indirect information only. Finally, it often turns out that $X = Z$ (of course, then $\mathfrak{X} = \mathfrak{Z}$). In this case it is said that the model is *completely observable*. Otherwise, it is called *partially observable* (or *with incomplete information*).

Finally, we shall define one more family of conditional measures on the space of the controls U .

Definition 1. We shall call the conditional measure on \mathfrak{U}

$$\sigma^{(t)} \stackrel{\text{def}}{=} \sigma^{(t)}(N|z^t, u^{t-1}), \quad N \in \mathfrak{U}, \quad t \geq 0.$$

a *control choice rule* at time t .

In the singular case the randomized rule $\sigma^{(t)}$ turns into a deterministic one which is represented by the measurable mapping $\sigma^{(t)} : Z^t \times U^{(t-1)} \rightarrow U$. Hence the rule $\sigma^{(t)}$ points to the control u_t which must be chosen by using the informational image of the model up to time t , i.e. the sequence $u_0, z_1, u_1, \dots, z_{t-1}, u_{t-1}, z_t$ in the stochastic or deterministic way. Such laws are called the *non-anticipated* ones.

Definition 2. A family of the choice rules of the controls $\sigma = \{\sigma^{(t)}, t \geq 1\}$ is called a *strategy of the control* or just a *strategy*.

The terms “strategy”, “control algorithm” and “control law” are synonyms. In terms of automatic control theory a “strategy” is the feedback under which the disconnected control system is closed.

Let $\Sigma = \{\sigma\}$ denote a non-empty set of the “admissible” strategies. The choice of Σ depends on the structure of the model. The “admissibility” means that there are some specific circumstances differentiating one model from another which should be taken into account. We shall denote a strategy for the model $\{\mu_t, t \geq 1\}$ with the observations $\{\nu_t, t \geq 0\}$ by $\sigma_{\mu\nu}$.

Now we can formally define a controlled process in discrete time.

Let a family of conditional probabilities $\{\mu_t, \nu_t, \sigma^{(t)}\}$, $t \geq 0$ be given on the measurable spaces (X, \mathfrak{X}) , (Z, \mathfrak{Z}) , (U, \mathfrak{U}) of the states, the observations and the controls respectively. Here the strategy σ belongs to the set of the strategies Σ . We shall restrict ourselves to the case of topological spaces X, Z, U . Moreover, for all applications of the proposed theory it is sufficient to suppose that these spaces are complete separable metric spaces. It is quite reasonable to assume that the conditional probabilities $\mu_t, \nu_t, \sigma^{(t)}$ are the Borelian functions (with respect to the arguments noted into the conditions). It does not restrict the application field of our theory. According to Jonescu Tulcea Theorem there exists a three-dimensional r.p. (x_t, z_t, u_t) whose paths of which belong to the countable product $(X^\infty \times Z^\infty \times U^\infty, \mathfrak{X}^\infty \times \mathfrak{Z}^\infty \times \mathfrak{U}^\infty)$ of the state space, the observation space and the control space, respectively. The corresponding probability measure on the space (Ω, \mathfrak{F}) is unique.

Definition 3. The sequence $\zeta = \{\mu_t, \nu_t, \sigma^{(t)}, t \geq 1\}$ formed by the family of the controlled conditional distributions μ and ν and the class of the admissible strategies $\Sigma_{\mu\nu}$ is called the *controlled random process* (CRP for short).

The spaces X, Z, U are defined by the measures μ and ν respectively. We shall denote the CRP under a fixed strategy σ by $\zeta(\sigma)$.

Any strategy $\sigma \in \Sigma_{\mu\nu}$ defines a random process $\zeta(\sigma)$ pertaining to the r.p. class given by the family μ .

Let us describe the evolution of a CRP at some fixed strategy from $\Sigma_{\mu\nu}$. At the initial moment $t = 0$ our model is in a state x_0 which generates an observation z_0 . Taking into account this observation a control u_0 is calculated. At the moment $t = 1$ the model passes into a new state x_1 which produces a new observation z_1 and then a new control u_1 is calculated by using the data (z_0, u_0, z_1) . This new control leads to the new observation and state at time $t = 2$ respectively. Using the past history $(z_0, u_0, z_1, u_1, z_2)$ a control u_2 is defined and so on. So step by step the controlled random process is progressed. At each time t it consists of three components $i_t = (x_t, z_t, u_t)$. For the non-terminating processes a result of the control is the infinite sequence of the triplets $(i_0, i_1, \dots, i_t, \dots)$ that represents the trajectory of the process in $X^\infty \times Z^\infty \times U^\infty$. We can consider separately the trajectory of the model x_t in the states space X and the trajectories z_t and u_t in the spaces Z and U respectively.

Definition 4. The finite collection $i^t = (i_0, i_1, \dots, i_t)$ is called the *history of the controlled process* up to time t .

As mentioned above in the important case when $X = Z$ and $x_t = z_t$ one speaks of an observable process. Then the description of the controlled process is simplified.

Let the initial values x_0, u_0 and strategy σ be given. Then the mathematical expectations $E_{x_0, u_0}^{(\sigma)} \varphi_t$ and the higher moments of the function $\varphi = \varphi(x^t, u^{t-1})$ defined on the trajectories of the controlled process can be calculated (if they exist) by the evident formulae.

The deterministic models under the deterministic observations are often met in control problems. It is clear that they fall under the general conception of controlled random processes.

We shall end the list of the main notions of control theory with the notion of a *control aim* which so far had no abstract definition. Instead examples of some concrete goals are usually stated. The control aim is to provide the model with some desirable property that takes place under some strategies but not under others. The control algorithms (the strategies) serve to attain it.

We shall consider some deterministic models as examples of the control goals. As a general rule such a model is represented by a difference equation or a differential one. It is often required that its solution be dissipative or stable (in the Lyapunov sense) or asymptotically stable. Such aims are called the *stabilizational aims*. The *optimization aims* are more complicated. They are associated with some *objective functions* $W(\sigma)$ defined on the set of admissible strategies $\Sigma_{\mu, \nu}$. It is required to find an optimal strategy σ_0 that maximizes this function, i.e. $W(\sigma_0) = \max_{\sigma \in \Sigma} W(\sigma)$ or an ε -optimal strategy σ_ε under which the inequality $W(\sigma_\varepsilon) > \sup_{\sigma \in \Sigma} W(\sigma) - \varepsilon$ holds. Sometimes this inequality is written in the form $W(\sigma_\varepsilon) > (1 - \varepsilon) \sup_{\sigma \in \Sigma} W(\sigma)$. The optimal strategy and the extreme value depend on the given set of the admissible strategies.

The objective function is often additive. So if T is a finite number then

$$W(\sigma) = \sum_{t=1}^T \varphi(x_t, u_{t-1})$$

or

$$W(\sigma) = \lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \varphi(x_t, u_{t-1})$$

in the infinite time case.

Another widespread goal consists of satisfying the inequalities $\varphi(x^t, u^{t-1}) < 0$, $t \geq 1$. Not only the stabilizational aims but also many optimizational aims are its particular cases. The goals mentioned above can refer to “global” aims connected with the non-terminating paths of the model. The “local” aims may be rather important as well. For example, they are the transition from an initial state x_0 into a state \tilde{x} for time T or the optimal high-speed problem, i.e. starting from x_0 it is needed to reach \tilde{x} in minimal time under some restrictions on the controls.

Let us now consider the stochastic controlled models. The main aims of the control are divided into two groups. The first of them is composed of so-called *strong*

aims referred directly to the paths. They have a probabilistic sense. Here some examples of such aims are adduced. It is required to provide the implementation of the following inequalities

$$\varphi(x^t, u^{t-1}) < 0, \quad \forall t \geq 1$$

or of equality

$$\varliminf_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \varphi(x_t, u_{t-1}) = \max_{\sigma} W(\sigma)$$

which take place for almost all paths, i.e. they are true with probability one.

The strong aims are reached under heavy conditions on the considered models. As a general rule the model must be ergodic.

Finally, the second group is composed of the *weak aims*. They refer not to a function defined on the paths of the model but to the average characteristics defined by the chosen strategy. Let us consider the main types of strategies before we write down the explicit formulae of the mathematical expectations for the considered functions under these strategies.

Program strategies are the simplest. They do not depend on the evolution of the model. Such strategies are sequences of unconditional distributions \varkappa_t defined onto the space of controls U fixed in advance. If these distributions are singular then the controls will be some functions of time $u_t = f(t)$ which form the sequence $u_0, u_1, \dots, u_t, \dots$. The program strategies are convenient particularly in the technical applications. Various applied problems of control theory were solved by using them.

Stationary strategies are formed by using the identical rules. If the time t is running over all integers ($t = 0, \pm 1, \pm 2, \dots$) then $\varkappa_t = \varkappa$ for all t . But if the initial moment is fixed, i.e. $t = 0, 1, 2, \dots$ then $\varkappa_t = \varkappa$ beginning at $t = h \geq 0$. In the simplest case when $X = Z$ the considered strategies have the form $\varkappa(\cdot | x_{t-h}, x_{t-h+1}, \dots, x_t; u_{t-h}, u_{t-h+1}, \dots, u_{t-1})$ ($\varkappa(\cdot | x_{t-h}^t, u_{t-h}^{t-1})$ for short). In other words, the rules are different up to time $t = h$ but after this moment they coincide.

Definition 5. The number h is called the *memory depth* of the stationary strategy.

The *stationary program strategy* consists of using the distribution \varkappa only. In the deterministic case the control $u_t = u$ is repeated time after time.

Let us write down the representation of the mathematical expectation of the function $\varphi_t = \varphi(x^t, u^{t-1})$ under a program strategy having the form (u_0, u_1, u_2, \dots) . We have

$$\begin{aligned} W(u^{t-1}) &\stackrel{\text{def}}{=} \mathbf{E}\varphi_t \\ &= \int_{X^{t+1}} \varphi(x_0, \dots, x_t; u_0, \dots, u_{t-1}) \mu(dx_t | x^{t-1}, u^{t-1}) \dots \mu(dx_0 | u_0). \end{aligned} \quad (1)$$

Following this $\mathbf{E}\varphi_t$ is a measurable function of u . If the space U is topological then the conditional probabilities $\mu(\cdot|x^t, u^t)$ will be continuous with respect to u for all t and the integral from (1) converges uniformly. Then the functions $W(u^t)$ are continuous with respect to all of its arguments jointly. If the program strategy is randomized then $\mathbf{E}\varphi_t$ will not be a function but only a number. Indeed (again for the sake of simplicity we assume that $X = Z$)

$$\begin{aligned} W(t) &\stackrel{\text{def}}{=} \mathbf{E}\varphi_t \\ &= \int_{X^{t+1} \times U^t} \varphi(x_0, \dots, x_t; u_0, \dots, u_{t-1}) \mu(dx_t|x^{t-1}, u^{t-1}) \cdots \mu(dx_0|u_0) \\ &\quad \times \sigma^{(t-1)}(du_{t-1}) \cdots \sigma^{(0)}(du_0). \end{aligned}$$

Definition 6. A strategy generated by a sequence of rules which have memory depth equal to one, i.e. by the distributions $\sigma^{(t)}(\cdot|x_t)$ or $\sigma^{(t)}(\cdot|z_t)$ given on U is called a *Markov strategy*.

For the Markov strategy the current value of the model or the observation is important only. The deterministic Markov strategy is formed by using the functions of either $f_t(x_t)$ type or $f_t(z_t)$ type.

Definition 7. A strategy generated by a distribution $\sigma(\cdot|x)$ (i.e. at every instant of time the controls are chosen according to the same rule) is called a *stationary Markov strategy*.

Finally, a combination of the notions mentioned above leads to the *simple strategies*, i.e. the *stationary deterministic Markov strategies* consisting in using the same function f of the current state, i.e. $f(x_t) = u_t$.

This notion is used especially often since the solution of many problems can be obtained by means of the simple strategies. Then the mathematical expectation for any function can be calculated by analogy with the previous formula

$$\begin{aligned} \mathbf{E}\varphi_t &= W(t) \\ &= \int_{X^{t+1}} \varphi(x_0, \dots, x_t; f(x_0), \dots, f(x_t)) \mu(dx_t|f(x_t)) \cdots \mu(dx_0|f(x_0)). \end{aligned}$$

For any strategy (non-stationary and randomized) depending on the whole history of the model evolution the measure generated by it on the set of paths leads to the expression

$$\begin{aligned} \mathbf{E}_\sigma \varphi_t &= W(\sigma, t) \\ &= \int_{X^{t+1} \times U^t} \varphi(x^t, u^t) \prod_{i=0}^t \mu_i(dx_i|x^{i-1}, u^{i-1}) \sigma^{(i)}(du_i|x^i, u^{i-1}), \end{aligned}$$

where x^{-1}, u^{-1} should be treated as a formal notation of the absent variables. The partially observable models are given in a more complicated way. The analogues of the integral considered above have the rather cumbersome form. For this reason the appropriate formulae are omitted.

Now we can finally point out the typical aims of the control. It is assumed that the class Σ of the admissible strategies is chosen so that $\mathbf{E}_\sigma \varphi_t$ should be finite for all t . The following functions $\varphi_t = \varphi(x_t, u_{t-1})$ are the most common ones. Their role is connected with the fact that sometimes we can draw a conclusion about a current state x_t of the model knowing the value of some numerical characteristic of the state (for example, the function φ_t). The next aim is quite reasonable with respect to functions

$$\lim_{t \rightarrow \infty} \mathbf{E}_\sigma \varphi_t = \sup_{\sigma \in \Sigma} W(\sigma) = \bar{W}.$$

Here $W(\sigma)$ is the limiting objective function obtained by passing to the limit, i.e. $W(\sigma) = \lim_{t \rightarrow \infty} W(\sigma, t)$. This aim is called *asymptotic optimality*. In connection to the above we shall cite the appropriate terminology. So φ_t is the *reward* at time t , $\mathbf{E}_\sigma \varphi_t$ is the *average reward* at time t under the strategy σ , $W(\sigma)$ is the *limiting reward* and, at last, \bar{W} is the *maximum limiting average reward*. We are often forced to restrict to approximate optimality. For a fixed $\varepsilon > 0$ it is required to find a strategy σ_ε such that the inequality

$$\lim_{t \rightarrow \infty} \mathbf{E}_{\sigma_\varepsilon} \varphi_t > \bar{W} - \varepsilon$$

holds. Such an aim is called ε -*optimality* (in the weak sense).

The weakest optimal aims are concerned with the Cesaro averages such as

asymptotic optimality

$$\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \mathbf{E}_\sigma \varphi_n = \bar{W}$$

and ε -*optimality*

$$\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \mathbf{E}_\sigma \varphi_t > \bar{W} - \varepsilon.$$

Stabilization problems can sometimes refer to the weak aims. For instance, the asymptotic stability in the mean square sense can be reasonably interpreted as the minimization problem (with respect to the limit) of the function $\mathbf{E}_\sigma \|x_t\|^2$, i.e. $\lim_{t \rightarrow \infty} \mathbf{E}_\sigma \|x_t\|^2 = 0$.

The other aims of control are the fulfilment of the “goal inequalities” such as $a(t) \leq W(\sigma, t) \leq b(t)$ under some given numerical sequences $a(t), b(t)$. The more general form of such inequalities is connected with a family of functions $(\varphi^{(1)}, \dots, \varphi^{(l)})$. It is required that the inclusions

$$(W^{(1)}(\sigma, t), \dots, W^{(l)}(\sigma, t)) \in G$$

where $W^{(i)}(\sigma, t) = \mathbf{E}_\sigma \varphi_t^{(i)}$, $i = 1, \dots, l$, $G \in \mathbb{R}^l$ take place for all $t \geq t_0$.

The majority of well-known aims of control may be reduced to aims of this kind. For example, the optimizational aims (in the weak sense) or the stability

problems (in the last case the goal is the fulfilment of the inequality $E_\sigma \|x_t\|^2 \leq \gamma$) are considered as such aims.

We shall now discuss methods of solving control problems for some types of models.

A *process with independent values* (PIV for short) defined by a family of controlled conditional distributions of the form

$$\mu_{t+1}(\cdot|x^t, u^t) = \mu_{t+1}(\cdot|u_t),$$

is the simplest kind of CRP. This means that only the last of the controls used appears in the condition. In this case the class of the admissible strategies Σ consists only of the sequences of conditional distributions on U . In the physical point of view these processes are inertia-free and independent on the previous evolution. Sometimes it is more convenient to write down PIV in the form $x_t(u_{t-1})$.

Homogeneous processes with independent values (HPIV for short) have fundamental importance. Their conditional probabilities do not depend on time, i.e. $\mu_t(\cdot|u) \equiv \mu(\cdot|u)$. The measurable function $\varphi(x_t, u_{t-1})$ defined on the HPIV path is also a HPIV. The name of these processes is derived from the fact that the program strategy inverts HPIV into a sequence of the independent random variables identically distributed under $u_t \equiv u$. As we shall soon see the control strategy for HPIV has such a form. In the deterministic case, i.e. when the measures μ are degenerate the HPIV will be represented by a single-valued function of the form

$$x_t = g(u_{t-1}),$$

i.e. it is the optimization standard object. The more complex strategy transforms the HPIV. For example, the simple strategy $\sigma = \{f(x)\}$ turns it into a homogeneous Markov process with the transition function $\mu(M|f(x))$. The strategies of the general form having large or increasing memory depth transform the HPIV into some Markov process or, generally speaking, into a random process of general form.

The control aims for PHIV are usually double, i.e. the optimization problem and the achievement of a given level. They are formulated in the terms of “average rewards” for one step.

For the sake of simplicity we shall not touch the function $\varphi(x_t, u_{t-1})$ but we shall consider the scalar model only, i.e. $X = R^1$. Then the average reward at time t under a control u is equal to

$$W(u) = \int_{-\infty}^{\infty} y\mu(dy|u).$$

The average reward at the same time under a strategy $\sigma = \{\sigma^{(j)}\}$ is equal to

$$\mathbf{E}_\sigma x_t = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} W(u_{t-1}) \prod_{j=1}^t \mu(dx_j|u_{j-1})\sigma^{(j)}(du_{j-1}|x^{j-1}, u^{j-2}).$$

The following estimate is evident

$$\mathbf{E}_\sigma x_t \leq \max_u W(u).$$

We shall use the same symbols $W(u)$ and $W(\sigma)$ for the mathematical expectation at a fixed moment and for the limiting reward considered as a function given on Σ respectively. They differ by the arguments u or σ only. Having this in mind we shall obtain the obvious inequality

$$W(\sigma) \leq \max_u W(u).$$

Here the equality will take place if the control $u_0 = \arg \max W(u)$ is applied for all t , i.e. the stationary strategy $\sigma_0 = \{u_0, \dots, u_0, \dots\}$ is used. Then

$$\sup_{\sigma} W(\sigma) = \max_u \bar{W}(u).$$

Applying the strong law of the large numbers to the sequence $x_0, x_1, \dots, x_k, \dots$ obtained of the independent identically distributed random variables, we shall obtain

$$\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T x_t = \bar{W} \quad \text{a.s.}$$

For a task to be solved on the achievement of the given average level it is necessary to find a strategy σ^* such that the equality $W(u) = \alpha$ (where α is a number from the domain of the function W) takes place for all time t . The solution is obvious, namely, it is needed to calculate the root of the equation $W(u) = \alpha$.

It is also simple to give methods of solving other control problems by HPIV. Those methods are based on the assumption that all characteristics of the process are known exactly. Before we had used the average reward only.

There is another interpretation of the HPIV notion (in so-called “wide sense”) which, in turn, means that

- (a) for the fixed u_1, \dots, u_{t-1} the r.v. $x_{t+1}(u)$ does not depend on the r.v. $x_i(u_{i-1})$, $i \leq t$;
- (b) $\mathbf{E}x_t(u) = W(u)$;
- (c) $\sup_t \mathbf{E}x_t^2(u) < \infty$.

Let us consider the next class of processes called *controlled Markov chains*. These are specified by the 5-symbol collection $C = \{X, U, P^{(u)}, \bar{p}, \zeta\}$ where $X = \{x_1, \dots, x_m\}$, $U = \{u_1, \dots, u_k\}$ are the sets of states and controls respectively, the stochastic matrices $P^{(u)} = (p_{ij}^{(u)})$ are formed by the controlled one-step transition probabilities from x_i to x_j under the control u , i.e. $p_{ij}^{(u)} = \mathbf{P}\{x_i \xrightarrow{u} x_j\}$, $\bar{p} = (p_1, \dots, p_m)$ is an initial distribution and, finally, $\zeta = \zeta(x, u, \omega)$ is a numerical r.v. denoting the reward in the state x under the applied control u . The average rewards $r_x^u = \mathbf{E}\zeta(x, u)$ are required to be finite. The choice of the admissible strategy leads to the “evolution” of the Markov chain, i.e. the state x_t and the reward $\zeta_t = \zeta(x_t, u_t, \omega)$ become some functions of time. Unless otherwise stated these functions (the paths) will be assumed non-terminating.

The HPIV considered above are controlled Markov chains with a single state.

In the theory of controlled Markov chains one usually considers the discounted reward

$$W_\beta(\sigma, \bar{p}) = \sum_{t=0}^{\infty} \beta^t \mathbf{E}_{\sigma, \bar{p}} \zeta_t, \quad 0 \leq \beta < 1,$$

or the limiting average (for one step) reward

$$W(\sigma, \bar{p}) = \lim_{T \rightarrow \infty} T^{-1} \sum_{t=0}^T \mathbf{E}_{\sigma, \bar{p}} \zeta_t, \quad 0 \leq \beta < 1$$

as the objective functions. It is required to find an optimal strategy maximizing one of these functions. The structure of the optimal strategy is well known and simple. There are only k^m different functions defined onto X which take the values from U and generate the simple strategies. These functions form the set Σ . Linear programming provides the necessary calculation tools to find the optimal strategies. Let us state this problem for the ergodic chains having the communicating states only, i.e., there is a positive probability that for a finite number of transitions each state x_i can be reached from any state x_j . We shall restrict ourselves here to the case of the second objective function; the indexes i, j and l correspond to the states and to the controls respectively.

$$\begin{aligned} \sum_{i,l} r_i^l x_i^l &\rightarrow \max; \\ \sum_l x_j^l - \sum_{i,l} p_{ij}^l x_i^l &= 0, \quad j = 1, \dots, m; \\ \sum_{i,l} x_i^l &= 1; \\ x_i^l &\geq 0, \quad i = 1, \dots, m; \quad l = 1, \dots, k. \end{aligned}$$

Here the quantities x_i^l mean the choice probabilities of the control u_l in the state x_i^l or, in other words, the optimal strategy is being searched in the class of randomized strategies. The analysis shows that for any i there exists the unique $l(i)$ such that $x_i^{l(i)} > 0$ and $x_i^l = 0$ for $l \neq l(i)$. The solution of this problem points to

- (a) the optimal control law $u(x_i) = u_{l(i)}$;
- (b) the limiting probabilities of the states under the optimal strategy $\pi_i(\sigma_{\text{opt}}) = x_i^{l(i)}$;
- (c) maximum of the objective function

$$W(\sigma_{\text{opt}}) = \sum_{i=1}^m r_i^{l(i)} x_i^{l(i)}.$$

According to the strong law of large numbers for the ergodic Markov chains not only the objective condition concerned with the average rewards but the stronger

equality

$$\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \zeta_t(u_{t-1}) = \max_{\sigma} W(\sigma), \quad \text{a.s.}$$

holds. The limit value on the left-hand side of the last equality does not depend, naturally, on the initial state of the chain.

For the non-ergodic controlled Markov chains the problem of linear programming is more complicated, i.e.

$$\begin{aligned} \sum_{i,l} r_i^l x_i^l &\rightarrow \max; \\ \sum_{i,l} (\delta_{ij} - p_{ij}^l) x_i^l &= 0, \quad j = 1, \dots, m; \\ \sum_l x_i^l - \sum_{i,l} (\delta_{ij} - p_{ij}^l) y_i^l &= p_j; \quad j = 1, \dots, m; \\ x_i^l, y_i^l &\geq 0, \quad i = 1, \dots, m; \quad l = 1, \dots, k. \end{aligned}$$

Here δ_{ij} is the Kronecker symbol, p_i is an initial distribution, the variables x_i^l and y_i^l correspond to states from the ergodic classes and to all states respectively. Here we do not go into details. This problem with 2^{mk} unknown variables can be solved by one of the numerical methods of linear programming. Frequently some variant of the simplex method is used.

As seen from what has been said, to solve the optimization control problems of Markov chains it is required to know information about all chain parameters, i.e. the collection of $m^2 k$ controlled transition probabilities p_{ij}^l must be known exactly. Similarly, for HPIV information about $\mu(\cdot|u)$ is required. This *a priori* information may be replaced by the explicit form of the average reward $W(u)$.

The greatest difficulties arise for the *partially observable Markov chains* differing from the above-mentioned ones by the structure of the admissible strategies. They are formed by the laws $\sigma^{(t)} = \sigma(\cdot|z^t, u^{t-1})$ where z_t signifies the observable variable. It can be:

- (1) a reward ζ_t ;
- (2) a "pseudostate" $z \in Z = \{z_1, \dots, z_n\}$, z_j being observed with probability q_{ij} in the state x_j . The numbers q_{ij} form the stochastic matrix Q ;
- (3) the pair (ζ_t, z_t) .

For the partially observable chains the main control problem is still unsolved.

Now we shall give another description of Markov chains that has an abstract form. To that end, we shall use the branch of mathematical logic called the theory of automata. We shall start with *deterministic finite automata*. Let three finite sets $X = \{x_1, \dots, x_m\}$, $Y = \{y_1, \dots, y_k\}$, and $S = \{s_1, \dots, s_n\}$ be given. They are called the alphabets of the input signals, the output signals and the states respectively.

We shall denote a transition function by \varkappa , $\varkappa : X \times S \rightarrow S$ and an output function by λ , $\lambda : X \times S \rightarrow Y$ but s_0 is an initial state.

Definition 8. The collection

$$\mathfrak{A} = \{X, S, Y; s_0, \varkappa, \lambda\}$$

is called a *Mili automaton*.^a

Clearly it is treated as a one-one mapping of the semi-group of the words $G_X = \{x_{i_1}x_{i_2}\dots x_{i_p}, \forall p, i_1, \dots, i_p\}$ from the alphabet X into that $G_Y = \{y_{j_1}, \dots, y_{j_m}, \forall q, j_1, \dots, j_q\}$ from the alphabet Y . It is natural to consider this mapping as a time-varying one. So starting from some initial state s_0 the automaton receives the input symbols x_{i_1}, \dots, x_{i_t} which are transformed by the transition function \varkappa into the corresponding states $s_{\nu_1}, s_{\nu_2}, \dots, s_{\nu_t}$ and by the output function λ into the sequence of the output signals $y_{j_1}y_{j_2}, \dots, y_{j_t}$ having the same length. The initial segments of the input word correspond one-one to that of the output word. This mapping is called an *automaton mapping*. In mathematical logic automata characterize a class of algorithms.

Moor automata differs from Mili ones by the output function λ which depends on the state only, i.e. in this case $y = \lambda(s)$. It means that for Moor automaton the state space can be decomposed into a sum of disjoint subsets, i.e. $S = S^{(1)} + \dots + S^{(k)}$ where the same output signal y_j corresponds to all states from $S^{(j)}$. With this difference, however, both types of automata are equivalent, i.e. they realize the same mapping.

The inner logic of automata theory and its applications have led to the creation of more complicated types of automata realizing more general automaton mappings. The probabilistic (stochastic) automata were the first of them.

Definition 9. The collection

$$\mathfrak{A} = \{X, S, Y, \bar{p}, \Pi(x), \mu\}$$

is called a *stochastic Moor automaton*.

Here $\bar{p} = (p_1, \dots, p_m)$ is the initial distribution of the automaton states, $\Pi(x) = (\pi_{ij}(x))$ is the stochastic matrix of transition probabilities from the state s_i to the state s_j under the input signal x , $\mu = \mu(y|s)$ is the conditional distribution onto Y under the automaton state s . The stochastic Mili automaton can be defined in a similar way.

It is easy to see that the stochastic Moor automaton (and Mili one as well) are controlled Markov chains. Indeed, the input signals x are the controls, the states of automaton serve as the states of the chain, the matrices $\Pi(x)$ are the analogue of the controlled conditional transitional probabilities $\mathbf{P}\{s(t+1)|s(t), x(t)\}$, the initial distribution \bar{p} has the same meaning in both cases and, finally, the output

^aThe *non-initial automata* are used often. Their initial state is not specified.

signals y are the abstract analogue of the rewards to appear with the probabilities $\mu(y|s)$. In the case of numerical rewards one may not introduce the set Y and the conditional distribution μ . Instead, it is enough to give the family of r.v. $\zeta(s, \omega)$ or $\zeta(s, x, \omega)$ which means the presence of the corresponding conditional distribution. Under such an agreement the automaton can be written in the form that does not differ from the notation of a Markov chain (if, additionally, we write $U = \{u\}$ instead of $X = \{x\}$)

$$\mathfrak{A} = \{S, U, \bar{p}, \Pi(x), \zeta(s, x)\}.$$

Automata with variable structure whose matrices $\Pi_t(x)$ and the r.v. ζ_t depend on time t are regarded as another type of automata. They will also be useful for us in the future. These automata can be reduced to the common ones but with some infinite S .

The terminology of automata theory often has a number of advantages. One of them is that an automaton can be naturally interpreted as a control algorithm. It is true in technique. For example, the controlling calculators and discrete devices of automatic control represent by itself the finite deterministic automata.

The strategy realized by finite automata can be directly applied not only to HPIV but to Markov chains as well. As we shall see later the infinite automata can provide attainment of the aims of control for the most complicated models. The process of the control of the model represented in the form of an automaton A_μ by means of an automaton A_σ can be visually illustrated as an interaction of two automata. The first one A_μ sends the sequence of states of the model x_t or the observable values z_t to the automaton A_σ and the latter replies to it by the sequence of the controlling signals u_t . Such a system denoted by $A_\mu \otimes A_\sigma$ realizes the one-one mapping.

As the last example of control theory problem we shall choose the *linear-quadratic problem* (LQP). In the simplest form it is connected with the linear difference equation with constant coefficients

$$x_t + a_1x_{t-1} + \cdots + a_nx_{t-n} = b_1u_{t-1} + \cdots + b_mu_{t-m} + \xi_t, \quad t \geq 0,$$

or, in another form, with the system of the first order equations

$$y_{t+1} = Ay_t + Bu_t + \eta_t, \quad t \geq 0. \quad (2)$$

The numbers (a_i, b_i) and matrices A, B are supposed to be known, ξ_t and η_t are the additive noises, i.e. the r.p. of some nature. Let us consider Eq. (2) supposing that $y_t, \eta_t \in \mathbb{R}^l, u_t \in \mathbb{R}^m$. We assume that η_t is a sequence of independent, identically distributed r.v. where $E\eta_t = 0$ and the matrix R of their second moments is finite. The initial value y_0 is supposed to be a r.v. with the distribution P and with the matrix of the second moments R_0 but $Ey_0 = m$. We shall choose the function

$$W(\sigma, \mathbf{P}) = \sum_{t=0}^{T-1} [y_t^T Q_1 y_t + u_t^T Q_2 u_t] + y_T^T Q_0 y_T,$$

defined on the finite time interval $[0, T]$ as an objective function. Here Q_0 and Q_1 are some non-negative matrices, Q_2 is positive definite. The aim is to minimize the non-negative function $W(\sigma, \mathbf{P})$. The solution of this problem has been carefully studied. The optimal control law is linear (by the state), deterministic and non-stationary. It has the form

$$u_t = K(t)x_t$$

where the amplification matrix $K(t)$ can be expressed in terms of the numerical parameters of the equation and of the objective function, namely,

$$K(t) = [Q_2 + B^T S(t+1)B]^{-1} B S(t+1)A.$$

Here the non-negative definite matrices $S(t)$ are the solution of the Riccati matrix recurrent system

$$S(t+1) = A^T S(t)A + Q_1 - A^T S(t) [Q_2 + B^T S(t+1)B]^{-1} B S(t)A$$

with the following boundary condition (on the right end) $S(T) = Q_0$. This system will be solved **before** the solution of the optimization problem is found. The minimal value of the objective function is equal to

$$\bar{W}(T) \stackrel{\text{def}}{=} \min_{\sigma} W(\sigma, T) = m^T S(0)m + \mathbf{sp} S(0)R_0 + \sum_{n=0}^{T-1} \mathbf{sp} S(n+1)R_1.$$

If the observations are incomplete, i.e. $z_t = Ly_t + \zeta_t$ we shall have to use additional considerations to obtain the useful signal y_t from the accessible information z_t . With this aim in view the ‘‘Kalman filter’’ is used.

We remark that solving the control problem for the new model again requires full information about characteristics of the linear difference equations, i.e. it is necessary to know $\mathbf{E}y_0$, the matrix R for the noise η_t , and five matrices A , B , Q_0 , Q_1 , Q_2 . For the values of the amplification matrix $K(t)$ to be calculated this *a priori* information is needed. Otherwise, if there is a lack of *a priori* information or inaccuracies into the description of the model then the linearly-quadratic problem cannot be solved and it remains only to sympathize with the designer of such a control system.

Let us return to the controlled model with the strategy. Let the model and the strategy be represented by automata A_μ and A_σ correspondingly and it is assumed that the spaces of the states X , the observations Z and the controls U are some measurable (and may be topological) spaces. Supposing that $X = Z$ we shall accept the controlled Markov process A_μ with the state space X and with the control space U as a model. The properties of the process are specified by the controlled transition function $\mu(\cdot|x, u)$. We shall choose the stochastic Moor automata $A_\sigma = \{X, S, Y; \Pi(x), q\}$ as a class of admissible strategies Σ_μ . As arranged above an interaction between A_μ and A_σ is denoted by $A_\mu \otimes A_\sigma$.

We shall connect with the object $A_\mu \otimes A_\sigma$ an *associated Markov process* (C, P) where $C = X \times S$ is the state space of this process but its transition function P is given by

$$P(U|c) = \int_{X \times U} \pi_{ij}(w) \mu(dw|x, u) q(du|s_i)$$

where $U = M \times S$ is a subset of C , $M \in \mathfrak{X}$, the pair $c = (x, s_i)$ being the previous state of the associated Markov chain. It is easy to write down the transition probability from c into a set $M \times \hat{S}$ where $\hat{S} \subset S$. This process is homogeneous in time. Its paths are the sequences $(x_0, s_0; x_1, s_1; \dots; x_t, s_t, \dots)$. For the discrete spaces X and U the integrals in the above mentioned formula are replaced by the sums and we can speak about the transition from the state $s' = (x', s')$ into $s'' = (x'', s'')$. If the spaces X and S are finite then (C, P) will become the *associated Markov chain*. It is interesting to know the conditions under which it is regular, i.e. it is ergodic without cyclic subclasses. Here we shall point to one sufficient condition, namely,

the chain is regular but the automaton is strongly tied up and it has no cyclic states.^b

The regular chains have positive limiting probabilities. It enables us to calculate the limiting mathematical expectation of the reward (if it is defined for this chain). The form of the limiting average reward depends on the strategy A_σ used.

Let us consider the special case of the controlled HPIV. In this case the structure of the associated process (C, P) is simple enough. The set $C = S$, where S is the state set of this automaton, serves for the state set of the process whose transitions are regulated by the matrix $P = (p_{ij})$

$$p_{ij} = P(s_i \rightarrow s_j) = \int_{X \times U} \pi_{ij}(w) \mu(dw|u) q(du|s_i).$$

This process is again homogeneous. If the PIV is considered as a model (a lack of homogeneity takes place) or automaton A_σ has changeable structure then the process (C, P) will be non-homogeneous. If the set S is finite, we shall again call (S, P) a chain. It is regular under the condition stated above and there exists a positive stationary distribution $\pi_1, \dots, \pi_{|s|}$ which does not depend on the initial state. The limiting average reward (HPIV is scalar) is equal to

$$W(\xi, A) = \sum_{l=1}^k W(u_l) \tilde{\pi}_l$$

where $\tilde{\pi}_l = \sum_j \pi_j q(u_l|s_j)$ are the stationary probabilities of the choice of the control u_l and $W(u) = \int x \mu(dx|u)$ is the average reward under the control u .

^bThe strong tie-up means that from each state of the chain it can pass into any other. The state of automaton will be called a cyclic if the greatest common divisor of the lengths of the input sequences which transform this state into itself with a positive probability is greater than one.

If the initial state is s_0 and the automaton A realizes the strategy u_1, \dots, u_k then the mathematical expectation of the reward at time t will be given by

$$W(\xi, A, s_0, t) = \sum_{l=1}^k W(u_l) \sum_j p_{s_0 j}^{(t-1)} q(u_l | s_j).$$

As known from the theory of Markov chains

$$\lim_{t \rightarrow \infty} W(\xi, A, s_0, t) = W(\xi, A)$$

with exponential convergence rate. (It is defined as the second largest absolute value of the eigenvalues of the matrix P .)

Our consideration has been concerned with both controlled models and control problems in discrete time. Fundamental difficulties prevent any formal and correct definition of controlled processes in continuous time. Therefore such definitions are not stated. Nevertheless, this will not pose us problems since from the whole set of such processes we shall confine ourselves to these topics:

- (1) the semi-Markov countable processes;
- (2) the ordinary differential equations of the form

$$\dot{x} = Ax + Bu + h(t)$$

where $h(t)$ is the deterministic or stochastic additive noise;

- (3) the stochastic differential Ito equations

$$dx_t = (Ax_t + Bu_t)dt + Cdw_t$$

where w_t signifies a Wiener process.

For all above-mentioned cases the structure of optimal and admissible strategies is well known.

1.3. Definition of Adaptive Control

The control problems described above are characterized by the presence of complete information about the model required for their solution. We have to know not only the model structure but all functions and constants entering into the description of this model. For short we shall name the control theory worked out for such cases as the *classical control theory*. For a long time this theory has been working and has reached an advanced stage of development and has proved its significance in the many practical applications. Let us turn to another situation initiated mainly by practical interest but (partially) also by theoretical one. We shall suppose now that *a priori* information about the model is incomplete. More precisely, we shall assume about the model and the appropriate controlled random process (CRP) that some common properties of their structure are known only. Let a class \mathcal{K} of the CRP be given only.

Definition 1. By *adaptive control theory* we mean the part of control theory devoted to the study of the whole class of control processes \mathcal{K} , instead of a particular process, due to the lack of complete information.

We shall begin our consideration with the class $\mathcal{K} = \{\xi\}$ of the CRP each element of which is characterized by the triplet $(\mu_t, \nu_t, \Sigma_{\mu, \nu}, t \geq 0)$, i.e. there is a collection of conditional distributions (μ_t) and (ν_t) and a set of admissible strategies $\Sigma_{\mu, \nu}$. The control aim, related to any process from \mathcal{K} is given. It is supposed that this class contains an infinite number of elements and the intersection $\cap_{\mu, \nu} \Sigma_{\mu, \nu}$ of all admissible strategies for CRP from \mathcal{K} is non-empty. It is rather convenient to specify the class \mathcal{K} by using an auxiliary parameter θ belonging to some parameter set Θ . Having this in mind both numerical and other (unknown) characteristics of the conditional distributions (μ_t) and (ν_t) are combined in a collection denoted by the single symbol θ . To emphasize the dependence of the considered distributions upon parameter θ their notations are supplied with a corresponding symbol, i.e. we shall write $(\mu_t(\theta), \nu_t(\theta))$, $t \geq 0$. Then instead of CRP ξ and \mathcal{K} we can write $\xi(\theta)$ and $\mathcal{K}(\Theta)$. Here the parameter set Θ is determined exactly. The advantages of such a notation consist in the obvious description of an “*a priori* uncertainty” set that differentiates the given class of the CRP from the others. Now we shall give examples of parameterization for some classes of CRP.

1. HPIV ξ with the two-element state space $X = (1; -1)$ and the finite set of the controls $U = (u_1, \dots, u_k)$. This process is completely determined by the collection of probabilities $q_j = \mathbf{P}\{x = 1 | u_j\}$, $j = 1, \dots, k$ which are considered as k -dimensional parameter.
2. The class of linear difference equations $x_{t+1} = Ax_t + Bu_t + h_t$. Here the parameters are the elements of the matrix pair $\theta = (A, B)$.
3. The class of Markov chains $(S, P^{(u)}, U)$ with matrices of transition probabilities $P^{(u)}$, $u \in U$ which form the parameter θ . If the rewards are given on the chain then their average values must be included in θ .

If an aim contains the mathematical expectations of the functions φ_t then, without reservation, they are assumed to exist and to be finite. Such a statement as “the strategy σ has led CRP ξ to the aim” means that the result of the interaction between ξ and σ generates the process that has the properties declared into the definition of the considered aim. We formulate the main definition below:

Definition 2. An admissible strategy (for all processes from \mathcal{K}) which leads any process from \mathcal{K} to the given aim C is called an *adaptive control* (or *adaptive strategy*) with respect to the class \mathcal{K} under the aim C .

The “adaptability” of a strategy signifies that it is intended not only for the individual CRP from \mathcal{K} but for all processes entering into \mathcal{K} . In the course of the control process we do not know, generally speaking, what process from \mathcal{K} concretely is under the control. In the adaptive control theory the aims are the same as in the

classical one. Indeed, practice puts forward the aims which are indifferent to our knowledge about the controlled object. A customer of a control system does not usually worry about the design knowledge of the particular features of the model. He is, generally, interested in the final result only.

In this connection two questions arise. Firstly, whether the control system can be designed under a lack of *a priori* information? But also what the suppositions must be done for it to be realized? Secondly, how (by what means) can the adaptive strategies be constructed? The most of the remaining part is devoted to the answer the first question. Now we would like to give a short answer the second one.

The classification of the adaptive strategies contains three main points:

- (1) *the identification strategies;*
- (2) *the direct strategies;*
- (3) *the searching strategies.*

An identification strategy consists of a combination of two operations simultaneously. The first of them is an evaluation of the unknown parameters of the model (we have denoted them by symbol θ united as the scalars and the vectors as the matrices and the points of some function spaces and so on). The second operation is the calculation of the controls by using the received estimates and according to the choice rules forming the required strategy. Apparently, there are two rather serious restrictions on the identification approach. Namely,

- (1) the “good” (converging) estimates of the parameter θ have to exist;
- (2) for known θ we must know how (by what method?) an optimal strategy is found.

These suppositions are not trivial. To come nearer to the given aim we have to choose the controls by using the rule which corresponds to the current value of the estimate. Unfortunately, this causes a deterioration of the estimates quality. In this case the convergence of these estimates to the true values of the parameters may be lost. Therefore, it is necessary to “mar” the controls, for example, to randomize them so that the optimal controls provided the achievement of the goal for CRP should appear more often. Doing so we have to choose the “incorrect” controls with a positive probability though it moves the controlled process from the desirable course. Due to the appearing deflexions we can obtain necessary information which gives the appropriate estimates of the parameters. Along with the explicit identification which consists of using the estimates converging to the true values of the parameters the partial (indirect) identification is used as well. In the last case the parameter θ is estimated approximately, i.e. only to such an extent which is necessary for the given problem to be solved (to attain one’s aim). The indirect identification methods produce the estimate of the parameter θ belonging to such a region of the parameter space Θ where the control aim can be attained at least approximately.

When constructing an adaptive strategy the use of the identification approach is grounded on two stages. At the first stage the convergence of the estimates $\hat{\theta}_t$ of the parameter θ to the true value of this parameter as $t \rightarrow \infty$ should be proved and then, at the second one, it is necessary to prove that the control aim is reached since it does not result, generally speaking, from the convergence of the estimates.

The direct strategies can be applied when the structure of the choice rules for the controlled models from a given class of the CRP is known. For the sake of definiteness we shall suppose that such a strategy is stationary and it is generated by a deterministic rule $h(x_{t-h}^t, u_{t-h}^t; \varkappa(\theta))$ where $\varkappa(\theta)$ will denote the parameter determining the rule h if the model is specified by the parameter θ .^c

The direct approach ignores the dependence of the rule h on the model parameter θ . The parameter \varkappa of the rule h serves for the unknown parameter instead of the parameter θ and the problem is to find the “proper” estimates $\hat{\varkappa}_t$ of this parameter by using the observations of the process x_t (or by using the process z_t in the case of partial observations). If the function $h(\cdot, \cdot; \varkappa)$ is continuous with respect to \varkappa then the convergence $\hat{\varkappa}_t \rightarrow \varkappa$ as $t \rightarrow \infty$ will imply the convergence of $h(\cdot, \cdot; \hat{\varkappa}_t)$ to the true law $h(\cdot, \cdot; \varkappa(\theta))$. It remains to make sure that the control aim would be achieved. An application of the direct strategies usually requires stronger restrictions to the CRP class in contrast with the identification strategies and, besides, the proofs of the corresponding assertions are more complicated.

We shall adduce now an example of a direct strategy based on a wayfaring (a fluctuation) on the given set of rules. Such a wayfaring arises when the structure of the optimal control is well-known in advance and, moreover, is simple enough. For example, in the case of control problems by the HPIV class, the optimal strategy is to repeat unlimitedly the same optimal action u_{opt} . Let the control set $U = \{u_1, \dots, u_k\}$ be finite and neither the measure $\mu(\cdot|u)$ of the process nor the average reward $W(u)$ associated with the control aim be known. We shall use randomized rules of control choice $\bar{p}_t = (p_t^{(1)}, \dots, p_t^{(k)})$ where $p_t^{(i)} = P\{u(t) = u_i\}$. The set of all such rules forms the $k-1$ -dimensional simplex $S = \{\bar{p} : \sum_1^k p_i = 1, p_i \geq 0, \forall i\}$. Under lack of information about the controlled process from the CPIV class, the vector rules \bar{p} are transformed in the course of control so that they converge to the top of the simplex $e_j = (0, \dots, 0, 1, 0, \dots, 0)$ corresponding to the optimal action u_j . To that end, a family of operators $T^{(x)}$ depending on the observed value x of the process received in response to the previous action is constructed on the given simplex. The operators $T^{(x)}$ are represented as some automata or some recurrent procedures. So, $\bar{p}_{t+1} = T^{(x)}\bar{p}_t$. Then the sequence of rules $\bar{p}_1, \dots, \bar{p}_t, \dots$ generates a random wayfaring on the simplex S . The “adaptability” of this direct strategy

^cIn case, for example, of a linear quadratic problem the control law has the following form $u_t = -Kx_t$. The elements of the amplifier matrix $K(t)$ can be expressed by using the well-known procedure in terms of the initiative parameters entering into the description of the task or, more exactly, in terms of five matrices entering both into the control law and into the minimized function [see Sec. 2, Chap. 1].

takes place on the condition that the top set of this simplex corresponding to the optimal action $u_{\text{opt}} = u_{j_0}$ is the absorbing set for the random wayfaring.

The searching strategies differ from the above-mentioned strategies because the search for the required rule is carried out in the set of **all** deterministic rules, i.e. among the rules having the following form $[\{h(x)\}, \{h(x^{(1)}, x^{(2)})\}, \dots, \{h(x^{(1)}, \dots, x^{(n)})\}, \dots]$ where $\{h(x, \dots)\}$ stands for the set of all admissible functions having the necessary number of arguments belonging either to the state space X or to the observation space Z . The control of the CRP consists of constructing the optimal strategy using increasing memory depth (or with infinite depth when time is running over the set $\{\dots, -1, 0, 1, \dots\}$). The construction difficulties of the searching strategies, even in the case of discrete spaces X, U , are concerned with the choice of search direction in the set of all competitive rules having an increasing number of arguments. An enumeration of all finite-valued functions with the increasing number of finite-valued arguments is considered as another problem. In practice such an enumeration must be effective, i.e. the function's number defines its explicit form by using either the tables or the formulae which can be used immediately. For this reason the searching method interpreted as a wayfaring on the set of complex functions is used for the most difficult problems but it has no applied importance up to now. This method demonstrates the principal mathematical resolvability of the problem considered. In the present monograph the use of searching strategies is limited to Chaps. 6–8. The identification and direct strategies can be realized in practice without special difficulties. These two types of strategies afford a basis for the application of adaptive control theory in industry, communication and elsewhere.

From what has been outlined above some conclusions about the common features of adaptive strategies can be drawn. First of all, the most typical property is non-stationarity. Next, these strategies are non-Markovian, i.e. they depend not only on the last state of the model but on the more distant past history. Finally, to control the non-deterministic objects it is necessary to use randomization, i.e. the control choice rules are some probability distributions defined on U . Note, however, that there are control problems with the strategies of elementary form, for example, $u(t) = Cx(t)$. Unfortunately, such problems are exceptions.

The adaptive control has a fundamental difference from the classical one. It consists of an uncertainty, as a general rule, of the moment when the control aim gets “near-by”. We shall explain this by two examples given below.

First, we shall consider the weak aims for the models $\mu(\theta)$ where $\theta \in \Theta$ is a parameter which differentiates one model from another. Let $W_\theta(\sigma)$ be an objective function, $\bar{W}_\theta = \sup_\sigma W_\theta(\sigma)$. It is required that the following inequality $\lim_{t \rightarrow \infty} E_\sigma \varphi_t > \bar{W}_\theta - \varepsilon$ should hold (φ_t is some function defined on the paths of the model). For a given model θ it is possible, in principle, to calculate either the first moment $t^*(\theta)$ when the inequality $E_\sigma \varphi_{t^*(\theta)} > \bar{W}_\theta - \varepsilon$ holds or the upper estimate of this moment (and also to estimate the convergence rate or some other characteristic of the transition process). As indicated in the corresponding notation

this moment depends on the model considered. In adaptive theory we have another situation. In the case of lack of *a priori* information about the model the magnitude $t^*(\theta)$ can take any possible value and, as a matter of fact, is unbounded on the class $\mathcal{K} = \{\mu_\theta, \theta \in \Theta\}$. Therefore, the estimates of the convergence rate must be uniform on the class \mathcal{K} .

In the case of strong aims the situation is more complicated. If, for example, it is required to ensure the implementation of the inequality

$$\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \varphi_t > \bar{W}_\theta - \varepsilon, \quad \text{a.s.}$$

then we shall take an interest in the random moment $\tau(\theta)$ starting from which the following inequality

$$\frac{1}{T} \sum_{t=1}^T \varphi_t > \bar{W}_\theta - \varepsilon$$

always holds. The moment $\tau(\theta)$ is non-Markov, i.e. it is not adapted to the information provided by the past history of the process and, therefore, is unobserved. In other words, at every moment of control we have no confidence that the ε -optimal adaptive strategy “has understood” the situation and one “has adjusted” to the proposed model and it has already used almost the optimal rules.

Thus appearance of the non-Markov moments $\tau(\theta)$ and other unknown characteristics such as $t^*(\theta)$ forms an uncertainty entering into the description of our model. The “uncertainty” is the subject-matter of such branches of mathematics as game theory and statistics. The distinction between a decision process in statistics and one in adaptive control theory is based on the fact that in the first case the observations are performed over a finite interval of time (it is fixed beforehand or it ends at some Markov moment) and, hence, it may lead to errors with a positive probability but in the second case the decisions of the adaptive strategies after the moments $\tau(\theta)$ or $t^*(\theta)$ are faultless. In the control problems by the stochastic models it means that we have to consider such problems on unlimited time intervals.

The subject matter of adaptive control theory consists in finding sufficient conditions for existence of adaptive strategies for various aims and classes of CRP. These conditions, with a few exceptions, have constructive character and they can be realized in practice without difficulties. For theoretical and practical purposes it is important to know the necessary conditions for adaptive strategies to exist. However, up to now they are known only for a few classes of the CRP. Heuristic considerations together with experience enable us to state some hypothetical assumptions about the necessary conditions. Consider the following:

- A. The control aim is attainable, in principle, with respect to the considered class of the CRP.
- B. A class of the CRP must not contain the “very” non-homogeneous in time (non-stationary) processes.
- C. In the course of time the influence of the distant controls on the process vanishes.

Here are some intuitive arguments in their favour.

Condition **A** is self-evident but non-trivial. In optimization problems for Markov chains the accessibility of the limiting average reward maximum is the necessary and sufficient condition for an adaptive strategy to exist.

Condition **B** arises in conjunction with the fact that a non-stationary CRP usually has non-stationary optimal strategies. Then in the course of control it has to search not only the optimal choice rules of controls but also to guess a law of their change (or “readjusting” to be able to pass from one rule to the other.) It is easy to give examples of classes of very simple CRP for which adaptive strategies do not exist.

Condition **C** rules out the possibility that incorrect controls chosen on the initial interval of control make the control aim unattainable. We shall illustrate this by an example. Let a class \mathcal{K} of CRP have the set of controls $U = \{u', u''\}$. For any CRP from \mathcal{K} it is required to maximize the average reward $W(u_0, u_1, \dots, u_t)$ calculated with respect to the measure generated by a program strategy. Let the class \mathcal{K} also have the property that for each CRP from \mathcal{K} the functions $W(u', u_1, \dots, u_t)$, $t \geq 1$ have the same unknown sign but the functions $W(u'', u_1, \dots, u_t)$ have the opposite one. It is clear that if the CRP is unknown then we cannot choose the proper initial control and, hence, the given aim is not always achieved.

1.4. Learning Systems

The essence of the definition of adaptive strategy given in the previous section is very broad. For a CRP in discrete time we would like to consider the adaptive strategies from another point of view, namely, to represent them in the form of automata and, hence, to produce the universal constructive definition. We shall begin from a notion of the learning system represented, in a general case, by an infinite automaton.

Let $\sigma^{(l)}$ be the control choice rule, i.e. a probability measure defined on U such that $(z^l, u^{(l-1)}) \in Z^{l+1} \times U^l$, where l is an integer that denotes the memory depth. In the singular case each deterministic rule is just a function $h(z^l)$ since the controls may be successively eliminated.

Definition 1. An object

$$\mathcal{E}_\sigma = (Z, \sigma^{(l)}, U)$$

which is represented as the stationary strategy generated either by the distribution $\sigma^{(l)}$ or by the function h is called an *elementary controlling system* \mathcal{E}_σ (or \mathcal{E}_h).

The observations z_t of the controlled process x_t and the controls u_t are considered as the input and output of \mathcal{E}_σ respectively. At times $t-1$ and t the controls are calculated by the pre-histories

$$(z_{t-l}, \dots, z_{t-1}; u_{t-l}, \dots, u_{t-2}) \quad \text{and} \quad (z_{t-l+1}, \dots, z_t; u_{t-l+1}, \dots, u_{t-1})$$

respectively.

In controlling the CRP the system \mathcal{E}_σ generates the next value of the process x_t (it is defined by the conditional distribution of the model μ_t) by using the control u_{t-1} . Then this value is transformed into the observable value z_t . It is the input signal of the system \mathcal{E}_σ and, further, according to the rule $\sigma^{(l)}$ the next control u_t is generated and so on.

Let \mathcal{D}_l denote the set of all rules with the memory depth l , i.e. the conditional distributions on U at the pre-histories belonging to $Z^l \times U^{l-1}$. Let us also put $\mathcal{D}_\infty = \cup_l \mathcal{D}_l$. We shall choose a non-empty set of admissible rules $\mathcal{D} \subseteq \mathcal{D}_\infty$. Let us now consider the family $\tilde{\mathcal{E}}$ of all elementary control systems corresponding in a one-to-one fashion to the rules from \mathcal{D} , i.e.

$$(\mathcal{E}_\sigma \in \tilde{\mathcal{E}}) \leftrightarrow (\mathcal{E}_\sigma = (Z, \sigma, U), \sigma \in \mathcal{D})$$

or, symbolically,

$$\tilde{\mathcal{E}} = (Z, \mathcal{D}, U).$$

We introduce one more notion. Let \mathfrak{R} be a measurable space (further it will be a metric one but now this is not important) and ξ be a CRP. Let a sequence of measurable functions Ψ_t defined on the observable trajectory z_t and on the controls u_t with values from \mathfrak{R} be given, i.e. $\Psi_t = \Psi(z_0, \dots, z_t; u_0, \dots, u_{t-1}) : Z^{t+1} \times U^t \rightarrow \mathfrak{R}$. The distributions of these variables are generated by the distributions of the model (μ_t), the observations (ν_t) and the rules ($\sigma^{(t)}$) used.

Definition 2. The sequence Ψ_t is called a *statistic* of the CRP ξ .

We shall explain the sense of these notions below.

We use the notation T_{Ψ_t} , $t \geq 0$, to denote a mapping of \mathcal{D} into \mathcal{D} , i.e. $T_{\Psi_t} : \mathcal{D} \rightarrow \mathcal{D}$ for all $t \geq 0$. The set T_{Ψ_t} , $t \geq 0$ may be considered as a family of mappings depending on two parameters, namely, Ψ and t . In addition, we shall suppose that under the mapping T_{Ψ_t} a rule having the memory depth not more than $t-1$ is associated with a rule having the memory depth not more than t . At the initial moment we use the rules having memory depth 0, i.e. they do not depend on the past history of the process. Due to correspondence between the sets $\tilde{\mathcal{E}}$ and \mathcal{D} we can consider T_{Ψ_t} as a mapping of $\tilde{\mathcal{E}}$ into \mathcal{D} . At time $t=0$ we have “the initial system” $\mathcal{E}_0 = (Z, \sigma^{(0)}, U)$ with the constant rule $\sigma^{(0)}$ not depending on the past.

Definition 3. The object

$$\mathcal{L} = [\tilde{\mathcal{E}}; T_{\Psi_t}, t \geq 0]$$

is called a *learning system*.

As seen from this definition a learning system is a generalized automaton with infinite sets of input and output signals. The detailed notation of such a system is

$$\mathcal{L} = [Z \times U, S, U; \mathcal{E}_0; T_{\Psi_t}]$$

where $Z \times U$ stands for the input alphabet, U denotes the output one, $S = (\mathcal{D}, \mathfrak{R}, Z^\infty \times U^\infty)$ is the state space, \mathcal{E}_0 is the initial state. The time-varying function T_{Ψ_t} is the function of the transitions and the output function is defined by the current state.

To find the structure and the function character of the learning system we shall consider its interaction with the CRP ξ . For the sake of simplicity, we assume that our model is completely observable, i.e. $X = Z$ and $x_t = z_t$. Otherwise, describing the state x_t we have to add “which is followed by appearance of the observation z_t with the probability defined by the conditional distribution ν_t ”. Let us apply at time $t - 1$ the control choice rule $\sigma_{t-1} \in \mathcal{D}$ which has been formed by using the history of the control process from $X^{t-1} \times U^{t-2}$. Then at the next moment t in accordance with the distribution $\mu(\cdot | x^{t-1}, u^{t-1})$ the value x_t appears and, subsequently, the statistics $\Psi_t = \Psi(x^t, u^{t-1})$ can be calculated. It generates the mapping T_{Ψ_t} and the previous rule σ_{t-1} is replaced by $\sigma_t = T_{\Psi_t} \sigma_{t-1}$ which, in turn, generates the new control u_t . Thus the new collection $(\sigma_t, x^t, u^t, \Psi_{t+1})$ is formed and at the next moment $t + 1$ the process is repeated again. The above is a repetition, practically word for word, of the process of controlling ξ under the strategy σ given in Sec. 2. Hence the notions of a learning system and a strategy coincide.

The operation of a learning system in the course of control is included in the presence of the set of admissible rules \mathcal{D} , memorizing the past history (x^t, u^t) , calculating of statistics Ψ_t and producing of the next mapping T_{Ψ_t} . In the general case it has to remember the whole path, i.e. the point of the space $X^\infty \times U^\infty$ but often it is enough to remember the points from $X^l \times U^l$ (if the memory depth is equal to l). Hence the structure of the set S includes the admissible rules, statistics and the history of the control process. In the typical cases the statistics Ψ_t are calculated recursively but sometimes it is necessary to remember the whole past history. Among the widespread statistics we shall emphasize two.

1. The value of the process, i.e. $\Psi_t = x_t$. Then $T_{\Psi_t} = T_{x_t}$;
2. The arithmetic mean of the “rewards”, i.e. $\Psi_t = t^{-1} \sum_{i=1}^t g(x_i)$.

The task of statistics is to give a method basing on which we choose the rules σ to have desirable behavior of the model. Then appearance of a sequence x_0, x_1, \dots, x_m at the input of the system implies the transformation of the initial rule σ_0 into $\sigma_m = T_{\Psi_m} T_{\Psi_{m-1}} \dots T_{\Psi_1} \sigma_0$. We emphasize that the stochastic nature of the input sequence implies that of the output one. This means that the sequence of rules σ_t is a random process in the set \mathcal{D} or, in other words, the CRP ξ generates some random wayfaring on \mathcal{D} . The achievement of the aim means the “purposefulness” of this wayfaring and its “aspiration” to use the most profitable rules. For the deterministic models the situation becomes much more simple.

It remains to explain why a learning system is a Moor automaton. This is because its output signal u_t is determined by the state of this system, i.e. by the rule σ_t acting at that moment and by the memory contents or, more exactly, by some point (x^t, u^{t-1}) of $X^\infty \times U^\infty$. The simplest example of a learning system is

the *stochastic learning model* (SLM for short) which has appeared in mathematical modelling of the behaviorism conception in psychology. Because of this the words “behavior”, “learning”, “adaptation” are used in SLM. This model interacts with an “environment”, receiving “reactions” from it in response to “stimuli” sent to the model.

Let the sets of stimuli $X = \{x_1, \dots, x_m\}$ and reactions $U = \{u_1, \dots, u_k\}$ be finite. The appearance of a stimulus from the environment is determined by the conditional distributions $\mu(x|u)$ or, in other words, the environment is a HPIV. We now describe the operation of this SLM. The elements of the sets X and U are, strictly speaking, its input and output signals. The admissible rules are the points of the simplex $S = \{p_1, \dots, p_k : \sum_1^k p_i = 1, p_i \geq 0, i = 1, \dots, k\}$ under the given Euclidean metric. These points are called the *behavior* of the SLM. Each of them is the stochastic vector whose j th coordinate p_j is interpreted as the reaction probability u_j . The number of the last input stimulus as the statistics which defines the collection of the mappings T_{Ψ_t} is chosen. All rules have the memory depth equal to one. The transformation of the behavior \bar{p} for one step is given by the formula

$$T^{(x)}\bar{p} = \alpha_x \bar{p} + (1 - \alpha_x)\bar{q}_x, \quad 0 \leq \alpha_x < 1$$

where $\bar{q}_x = (q_1(x), \dots, q_k(x))$ is a stochastic vector. If the same stimulus x arrives at the input of the SLM during n successive steps then the mapping $T_{x_n} = (T^{(x)})^n$ will have the form

$$T_{x_n}\bar{p} = \alpha_x^n \bar{p} + (1 - \alpha_x^n)\bar{q}_x.$$

The vector \bar{q}_x is the fixed point of this mapping, the parameter α_x pointing the convergence rate of $T_{x_n}\bar{p}$ to \bar{q}_x . Hence if the environment sends the SLM the same stimulus x then this system will “learn” the behavior \bar{q}_x . The interaction between the environment and the SLM consists of alternating the different stimulus and reactions. Therefore some k -dimensional random process \bar{p}_t corresponds to the input sequence x_t . Under the stated assumptions about the environment, \bar{p}_t is a Markov process and it is possible to prove that its probability distributions converge weakly to some limiting distribution. In some cases this distribution can be calculated. Notice that in the interaction between the environment and the SLM modeled above the notion of aim is absent. The problem of supplying the environment with some properties is not formulated, the scheme being intended for the modelling of some psychological phenomena only.

We introduce a constructive definition of “adaptive control”.

The CRP class \mathcal{K} is considered and an control aim is given as well.

Definition 4. A learning system leading each CRP from \mathcal{K} to the given aim is called an *adaptive control* with respect to the class \mathcal{K} under the given aim.

The difference of this formulation from the definition stated in Sec. 3 consists of pointing the universal automaton structure for CRP in discrete time. This structure has many advantages making the realization of the adaptive control system easy

to access. This realization may be done in the form of programs, special devices and so on. The control algorithm often happens to be more clear due to automaton structure. It is important that in Moor automata the output signals (the controls) depend on the current states only but not on the past history. If the learning system realizes adaptive control then the random wayfarings on the set of admissible control rules \mathcal{D} will gain a purposeful character. In the case of optimization aims either the random wayfarings lead to the absorbing subset of the optimal rules (belonging to \mathcal{D}) or the optimal rules (or almost such) are chosen with increasing frequency. In all cases when we can interpret the process of control as a random wayfaring on \mathfrak{D} it is convenient and reasonable to reformulate the original aim into the goal given for the wayfaring on the set of rules \mathcal{D} .

1.5. Bayesian Approach on a Finite Interval

Under the condition of absence or incompleteness of *a priori* information about the model, a control problem on a finite time interval is sometimes treated as an adaptive one but in another point of view than that stated above. In this connection we shall consider a problem of controlling a Markov process with discrete parameter. On the time interval $[1, T]$ a class \mathcal{K}_θ of controlled Markov processes is given with the state space X and the control space U which are supposed to be Euclidean. The transition functions of these processes are denoted by $\mu(\cdot|x, u; \theta)$ where the parameter $\theta \in \Theta$ characterizes the concrete process and it is running over, for the sake of definiteness, the real axis, i.e. $\Theta = \mathbb{R}^1$. The states x_t are supposed to be observed. At each moment t a reward $\varphi_t = \varphi(x_t, u_{t-1})$ represented by a continuous bounded function on $X \times U$ is given. The total reward per time T is $V_\theta = \sum_1^T \varphi_t(x_t, u_{t-1})$. It is required to maximize the objective function $W_\theta(\sigma) = E_\sigma V_\theta$. In the classical case, i.e. when the transition function μ is known such a problem can be solved by classical methods, for example, by the dynamic programming method without difficulty. But in the adaptive situation the values of parameter θ are unknown. Having the observations x_1, x_2, \dots we have to construct an optimal strategy maximizing $W_\theta(\sigma)$, i.e. to find the collection of control choice rules $(\sigma_0, \sigma_1, \dots, \sigma_{T-1})$ which maximize $W_\theta(\sigma)$.

To construct the optimal control we shall use the Bayesian approach to the problems with unknown parameters. The main hypothesis is the following:

H: There exists *a priori* distribution of the parameter θ denoted by $F(\theta)$ that is supposed to be known.

For simplicity, we shall assume that the distribution $F(\theta)$ has the density $f(\theta)$. Then according to Bayes' Theorem the posterior distribution densities of the parameter θ denoted by $f(\theta)^{(1)}, \dots, f(\theta)^{(x-1)}$ are calculated by using the observations of the process. We assume that the joint distributions appearing below and the rules σ_j have the densities which differ by the notations of their arguments.

The rules which form the optimal strategy are constructed by means of dynamic programming methods. We shall start with the last rule $q_{T-1}(u_{T-1}|x^{T-1})$ that represents the density of the conditional distribution of σ_{T-1} . To that end, we shall write the last summand (taken before averaging with respect to the past history) of the objective function

$$\mathbf{E}(\varphi_T|x^{T-1}) = \int \varphi_T(x_T, u_{T-1})p(x_T, u_{T-1}|x^{T-1})dx_T du_{T-1}$$

for any past history x^{T-1} . In this integral the conditional density is defined by the following formula

$$\begin{aligned} p(x_T, u_{T-1}|x^{T-1}) &= q(u_{T-1}|x^{T-1})p(x_T|u_{T-1}, x^{T-1}) \\ &= q(u_{T-1}|x^{T-1}) \int p(x_T|u_{T-1}, x_{T-1}; \theta) f^{(T-1)}(\theta|x^{T-1})d\theta \end{aligned}$$

where the second factor under the integral symbol stands for the *a posteriori* distribution density of the unknown parameter on the step $T-1$. By using Bayes' formula this density is calculated successively from the moment $t = 1$ with the prior density $f^{(1)}(\theta)$ up to the moment $t = T-1$. And so, the function being maximized is the following

$$\mathbf{E}(\varphi_T|x^{T-1}) = \int \Psi_T(u_{T-1}, x^{T-1})q(u_{T-1}|x^{T-1})du_{T-1}$$

where

$$\Psi_T(u_{T-1}, x^{T-1}) = \int \varphi_T(x_T, u_{T-1})p(x_T|u_{T-1}, x_{T-1}, \theta) f^{(T-1)}(\theta|x^{T-1})dx_T d\theta.$$

Let u_{T-1}^* denote the value of the argument u_{T-1} where the function Ψ is maximum, i.e.

$$u_{T-1}^* = \mathbf{arg\,max} \Psi_T(u_{T-1}, x^{T-1}).$$

It is clear that the function $q(u_{T-1}|x^{T-1})$ must be the “ δ -function” concentrated at the point u_{T-1}^* . In other words, the optimal rule at the moment $T-1$ is deterministic, namely,

$$u_{T-1}^* = h_{T-1}(x^{T-1})$$

and it is determined uniquely by the past history of the control process.

We shall now consider the second rule from the end. We require that, together with the rule h_{T-1} obtained, it gives the maximum of the function $\mathbf{E}(\varphi_{T-1} + \varphi_T|x^{T-2})$ for any past history x^{T-2} . Likewise to the previous we have

$$\mathbf{E}(\varphi_{T-1}|x^{T-2}) = \int \Psi'_T(u_{T-2}, x^{T-2})q(u_{T-2}|x^{T-2})du_{T-2}$$

where

$$\begin{aligned}\Psi'_T(u_{T-1}, x^{T-2}) &= \int \varphi_{T-1}(x_{T-1}, u_{T-2})p(x_{T-1}|u_{T-2}, x^{T-2}, \theta) \\ &\quad \times f^{(T-2)}(\theta|x^{T-1})dx_{T-1} d\theta.\end{aligned}$$

We account again that the posterior distribution density $f(\theta|x^{T-2})$ was calculated. From the equalities

$$\begin{aligned}\mathbf{E}(\varphi_T|x^{T-2}) &= \mathbf{E}(\mathbf{E}(\varphi_T|x^{T-1})|x^{T-2}), \\ \mathbf{E}(v_T|x^{T-2}) &= \int v_T p(x_{T-1}|x_{T-2}, u_{T-2}; \theta)q(u_{T-2}|x^{T-2})du_{T-2} dx_{T-1}\end{aligned}$$

it follows

$$\begin{aligned}\max \mathbf{E}(\varphi_{T-1} + \varphi_T|x^{T-2}) \\ = \max \int [\Psi'_T + \int v_T p(x_{T-1}|x_{T-2}, u_{T-2})dx_{T-1}]q(u_{T-2}|x^{T-2})du_{T-2}\end{aligned}$$

where $v_T = \max_{t \in [0, T]} \mathbf{E}\varphi_t$. We can find the rule at the moment $T - 2$ which provides the maximum to the function written down in the brackets. This rule is again deterministic and $u_{T-2}^* = h_{T-2}(x^{T-2})$ is an optimal control. Hence the density $q(u_{T-2}|x^{T-2})$ is again the “ δ -function” concentrated at the point u_{T-2}^* . Analogously, we can pass from $T - 2$ to $T - 3$ and so on. As a result of this we define the rules $h_{T-1}, h_{T-2}, \dots, h_1$ forming the optimal strategy. This strategy is deterministic and non-Markovian (at each moment t the control depends on the whole past history). One of the peculiarities of this approach consists of arranging the calculations: while the estimates of the parameter and the posterior densities are calculated successively from the moment $t = 1$ to $t = T$, the control choice rules are determined in the opposite order.

Let us analyze the sense of this approach. The supposition about the existence of a *priori* distribution of the parameter θ which means its stochastic nature is **not equivalent** to the absence of information about its value for the concrete controlled process. The original control aim (that is to maximize the objective function $W_\theta(\sigma)$ for any θ) is substituted by another one, namely, it is necessary to provide the maximum to the next function $\tilde{W}(\sigma) = \mathbf{E}_\theta W_\theta(\sigma)$. The additional integration which changes the aim is done with respect to the measure not having any relation to the situation under our consideration, i.e. to the type of the extreme problem and to the class of the controlled processes. Therefore the so-called “optimal Bayesian strategy” does not provide the original maximum to the function $W_\theta(\sigma)$ but only a smaller value depending on the prior density $f^{(0)}(\theta)$ chosen arbitrary. In our consideration this “imperceptible” substitution of the aim has been made at the end of the second paragraph where the function $W(\sigma)$ has appeared instead of the original function $W_\theta(\sigma)$. Next, the choice of $f^{(0)}(\theta)$ has neither reasonable nor natural grounds. Hence one cannot hope to receive even ε -optimality since it is impossible to choose the *a priori* density concentrated in a neighborhood of the true value of the unknown parameter. We have to “spread” the prior distribution

over the whole space Θ . The “maximum” value of the objective function and the “optimal” strategy depend on the “spreading” method.

We are forced to conclude that the described symbiosis of the Bayesian approach (to estimate the characteristics of the process) and dynamic programming method (to calculate the optimal strategies) **is not** the adaptive control because it does not guarantee that the chosen aim of the control will be attained. It is possible to reformulate the aim, namely, it is required to maximize $\mathbf{E}_\theta W_\theta(\sigma)$ by using the Bayesian approach. It is doubtful that such a distinct demonstration both of the limited nature and of the discrepancy of the problem to the essence of the matter could arise out of interest in it. It would be necessary to study the sensitivity of the attainable maximum with respect to the chosen prior distribution and to compare the efficiency of the Bayesian approach with the other statistical methods.

The unattainability of the aim under the Bayesian approach may be associated with the fact that the time interval of control is finite. In all known cases the stochastic adaptive control problems are posed and solved on an infinite time interval. Otherwise, there is a problem to determine the real aim of the control. In the stochastic control problems posed on an infinite time interval when the optimizational aim is connected, for example, with maximization of the limiting average reward

$$W(\sigma) = \lim_{t \rightarrow \infty} t^{-1} \sum_{n=1}^t \mathbf{E}_\sigma \varphi_n(x_n, u_{n-1})$$

the Bayesian approach under some restrictions gives in the limit the true value of the parameter, i.e. the posterior distribution converges to “ δ -function” concentrated at the true point. Then the influence of the incorrect controls chosen at first will disappear (“be smoothed”) in the course of time and, consequently, the maximum of the function $W(u)$ will be reached exactly or approximately.