By (6), to prove the basic result of this section we need to show that $v^s = v$. By virtue of (5) and (11) it suffices to verify that *under conditions* (A)—(D)

$$v^{\varphi} \leq v^s \tag{14}$$

in fact, in view of (5) we have equality both here and in (11).

By the results of §7 of Chapter 6, in a semicontinuous model Z_β with finite state space X and coefficient $\beta < 1$, there exists a stationary optimal strategy $\varphi_\beta = \psi_\beta^\infty$. By (6.5)

$$v_\beta = w_\beta(\varphi_\beta) = N_\beta q_{\psi_\beta} \tag{15}$$

where

$$N_\beta = \sum_{t=0}^{\infty} \beta^t P_{\psi_\beta}^t \tag{16}$$

All the elements of the matrix $(1 - \beta)N_\beta$ lie in the segment $[0,1]$. Therefore, from conditions (A) and (B), there is a sequence $\beta \uparrow 1$ such that the limits

$$\psi = \lim \psi_\beta, \qquad N = \lim(1 - \beta)N_\beta \geq 0 \tag{17}$$

exist (throughout what follows β takes on values only from this sequence). It follows from (C) and (D) that

$$P_\psi = \lim P_{\psi_\beta}, \qquad q_\psi \geq \overline{\lim}\, q_{\psi_\beta} \tag{18}$$

It follows from (12), (15), and (18) that

$$v^{\varphi} \leq N q_\psi. \tag{19}$$

It is clear from (16) that

$$N_\beta P_{\psi_\beta} = \frac{1}{\beta}(N_\beta - E),$$

E being the identity matrix. Multiplying both sides by $(1 - \beta)$ and taking account of (17) and (18), we find in the limit that $NP_\psi = N$. Hence $NP_\psi^t = N$ for any integer t, so that $NM_\psi = N$ (see (3)). Using this equation and formulas (2) and (19), we get

$$v^{\varphi} \leq NM_\psi q_\psi = Nw(\psi^\infty) \leq Nv^s. \tag{20}$$

Finally, it follows from (7) that $P_\chi v^s \leq v^s$ for any selector χ. In particular $P_{\psi_\beta} v^s \leq v^s$, so that, by induction $P_{\psi_\beta}^t v^s \leq v^s$, whence

$$N_\beta v^s = \sum_{t=0}^{\infty} \beta^t P_{\psi_\beta}^t v^s \leq \sum_{t=0}^{\infty} \beta^t v^s = \frac{v^s}{1 - \beta}.$$

It follows from this and (17) that $Nv^s \leq v^s$. Along with (20), this yields the desired inequality (14).

Part III

Some Applications

Chapter 8

Models with Incomplete Information

§1. Description of the Model

Up to this point we have been supposing that we observe completely the trajectory of the controlled process:

$$x_m \xrightarrow{a_{m+1}} x_{m+1} \xrightarrow{a_{m+2}} \cdots x_{t-1} \xrightarrow{a_t} x_t \to \cdots . \tag{1}$$

Now we suppose that *the state of the system at the time t is described by a pair* $x_t y_t$, *the first of these components becoming known to us and the second not.* Thus, the real course of the process is given by the trajectory

$$x_m y_m \xrightarrow{a_{m+1}} x_{m+1} y_{m+1} \xrightarrow{a_{m+2}} \cdots x_{t-1} y_{t-1} \xrightarrow{a_t} x_t y_t \to \cdots , \tag{2}$$

of which we observe, as before, the chain (1). The actions a_t and the observed states x_{t-1} are connected as before by a projection j.

The unobserved states y_t are elements of some sets Y_t. They influence both the transition mechanism into the next state and the resulting reward. The transition function p now gives a probability distribution for the state $x_t y_t$ of the product space $X_t \times Y_t$, depending on y_{t-1} and a_t (inasmuch as $x_{t-1} = j(a_t)$, the introduction of the additional argument x_{t-1} would not give anything new). The running reward q on the tth stage depends on the same pair $y_{t-1} a_t$. The final reward at time n depends on the pair $x_n y_n$*. Here we intend that the reward $q(y_{t-1} a_t)$ at the tth step should be paid out at the end of the control process. If we were to receive this reward immediately, then its numerical value would give us additional information on the unobserved states of the system, and the elements of the model would have to be reconstructed so as to include the number $q(y_{t-1} a_t)$ in the observed state x_t†.

* A more general case, when the reward on the tth step depends on the elements $x_{t-1} y_{t-1} a_t x_t y_t$, reduces to the case we are considering by the introduction of a new payoff equal to the mathematical expectation of $q(x_{t-1} y_{t-1} a_t x_t y_t)$ relative to the distribution $p(\cdot | y_{t-1} a_t)$.

† And then we would have a special case of the situation of which we spoke in the previous footnote: $q(x_{t-1} y_{t-1} a_t x_t y_t) = q(x_t)$.

In order to define a measure on the space of trajectories, it is necessary to give the initial distribution μ and the strategy π.

The rôle of the initial distribution μ here is somewhat different from what it was in the models with complete information. In taking μ to be known, we thus suppose known as well the probability distribution for the unobserved initial state y_m (although we are not given the value of y_m). In mathematical statistics one distinguishes the Bayesian approach, in which an "apriori" probability distribution for the unknown parameter y is introduced, and the minimax approach, in which the statistical decisions are evaluated according to the "worst" of the possible values of y. In supposing that μ is known, we choose the Bayesian approach.

The strategy π cannot depend on the unobserved values $y_m, y_{m+1}, \ldots$. However it can take account, besides the observed states $x_m, x_{m+1}, \ldots$ and the actions $a_{m+1}, a_{m+2}, \ldots$ already used, of the initial distribution as well. Inasmuch as the value of x_m becomes known to us, for the choice of the actions it is not the joint distribution μ of the initial pair $x_m y_m$ that is essential for us, but rather the conditional distribution ν_m for y_m given the observed x_m. We include the distribution ν_m into the observed history on which the next action depends. The pair $x_m \nu_m$ plays the rôle of the initial state. Here x_m is any point of the space X_m and ν_m is any probability measure on Y_m.

The pairs x_m, μ and x_m, ν_m are closely connected, but one can not express one in terms of the other uniquely. It is more convenient to deal with the second pair.

The values $w(\mu,\pi)$ and $v(\mu)$ are defined in the usual way in terms of the measure P_μ^π. The statement of the problem of optimal control does not change.

We note that if each of the spaces Y_t consists of one point, then we obtain the complete information model presented in the preceding chapters.

§2. Reduction to a Model with Complete Information. The Finite Case

For each model with incomplete information we shall construct a model with complete information in such a way that the values of the corresponding strategies coincide. Then we may apply the results of the preceding chapters and obtain theorems on the existence of optimal strategies in models with incomplete information.

Our plan consists in introducing new state spaces, taking as the state at the time t all the information at our disposal necessary for the further action. At the initial time m this information is described by the observed state x_m and the *apriori* distribution ν_m for the unobserved state y_m. At any time $t > m$, it is natural to describe it by the pair $x_t \nu_t$, where ν_t is the "*a posteriori*" probability distribution for the state y_t, taking account of the entire observed history.

We begin with the case when all the spaces X_t, Y_t are finite. In this case the probability of a chain $l = x_m y_m a_{m+1} x_{m+1} y_{m+1} \cdots a_n x_n y_n$ is defined by the formula*

$$P_\mu^\pi(l) = \mu(x_m y_m)\pi(a_{m+1}|x_m \nu_m)p(x_{m+1} y_{m+1}|y_m a_{m+1}) \cdots \pi(a_n|x_m \nu_m a_{m+1} x_{m+1} \cdots a_{n-1} x_{n-1})p(x_n y_n|y_{n-1} a_n), \quad (1)$$

where μ is the initial distribution, π is the applied strategy, and the distribution ν_m is computed from the formula

$$\nu_m(y_m) = \nu_m(y_m|x_m) = \frac{\mu(x_m y_m)}{\sum_{z \in Y_m} \mu(x_m z)} \quad (2)$$

if the denominator is not 0, and ν_m is an arbitrary probability measure on Y_m, if the denominator is zero.

The construction of the auxiliary model with complete information begins with the choice of the state spaces $\tilde{X}_t$. Put $\tilde{X}_t = X_t \times N_t$, where N_t is the set of all probability measures on the set Y_t.

The actions in the new model remain as before. Now the same action a_t is possible for different states $\tilde{x}_{t-1} = x_{t-1} \nu_{t-1}$, differing in the distributions ν_{t-1}. If we wish the fibres $\tilde{A}(x)$ not to intersect, then we have to consider a pair $\nu_{t-1} a_t = \tilde{a}_t$ as an action (see the analogous remark in §2 of Chapter 1).

In order to construct the new transition function $\tilde{p}$, we need to assign to each pair $\nu_{t-1} a_t$ a probability distribution on the space $X_t \times N_t$. The original transition function defines a distribution on the space $X_t \times Y_t$, as a function of $y_{t-1} a_t$. It is natural to assign to this pair a distribution on the space $X_t \times Y_t$ given by the formula

$$\tilde{p}(x_t y_t|\nu_{t-1} a_t) = \sum_{y_{t-1} \in Y_{t-1}} p(x_t y_t|y_{t-1} a_t)\nu_{t-1}(y_{t-1}). \quad (3)$$

This distribution splits into a distribution on X_t and a conditional distribution Y_t:

$$\tilde{p}(x_t y_t|\nu_{t-1} a_t) = \tilde{p}(x_t|\nu_{t-1} a_t)\nu_t(y_t|\nu_{t-1} a_t x_t); \quad (4)$$

here

$$\tilde{p}(x_t|\nu_{t-1} a_t) = \sum_{y_t \in Y_t} \tilde{p}(x_t y_t|\nu_{t-1} a_t) \quad (5)$$

and

$$\nu_t(y_t|\nu_{t-1} a_t x_t) = \frac{\tilde{p}(x_t y_t|\nu_{t-1} a_t)}{\tilde{p}(x_t|\nu_{t-1} a_t)}; \quad (6)$$

if the denominator in (6) vanishes, we take as ν_t the arbitrary fixed measure ν_t^0 on Y_t. For fixed $\nu_{t-1} a_t$ formula (6) defines a mapping $\nu_t = F(x_t)$ of X_t into N_t. Consider the probability distribution of the point $x_t F(x_t) \in X_t \times N_t$ which corresponds to the distribution of x_t given by (5). We thus have a probability distribution on the

* In the definition of the strategy π we have to add the requirement of measurability in the argument of ν_m.

space $X_t \times N_t = \tilde{X}_t$, depending on $v_{t-1}a_t = \tilde{a}_t$, i.e. a transition function $\tilde{p}$ of $\tilde{A}_t$ into $\tilde{X}_t$.

According to our plan, the distributions v_t ought to be "*a posteriori*" distributions for y_t, taking account of all the observations made up to the time t. In other words, the formula

$$v_t(y) = P_\mu^\pi\{y_t | x_m a_{m+1} x_{m+1} \cdots a_t x_t\} = \frac{P_\mu^\pi(x_m a_{m+1} x_{m+1} \cdots a_t x_t y_t)}{\sum_{z \in Y_t} P_\mu^\pi(x_m a_{m+1} x_{m+1} \cdots a_t x_t z)}$$

has to be satisfied. When $t = m$ this is true in view of formula (2), and if $t > m$ it is verified by induction using (1) and (3)–(6).

The new running reward is given by the formula

$$\tilde{q}(v_{t-1}a_t) = \sum_{y_{t-1} \in Y_{t-1}} q(y_{t-1}a_t)v_{t-1}(y_{t-1}), \tag{7}$$

and the new final reward by the formula

$$\tilde{r}(x_n v_n) = \sum_{y_n \in Y_n} r(x_n y_n)v_n(y_n). \tag{8}$$

* * *

Starting from model Z with incomplete information, we have constructed a new model $\tilde{Z}$ with complete information, in which the state and action spaces are uncountable. Let us show that the model $\tilde{Z}$ is semicontinuous (see Chapter 2, §4).

A probability measure ν on a space of s points is described by the choice of s nonnegative numbers adding to 1. This is a bounded closed set in s-dimensional coordinate space and is accordingly compact. Therefore all the spaces $\tilde{X}_t = X_t \times N_t$ and $\tilde{A}_t = N_{t-1} \times A_t$ are compact, hence condition 2.4.A is satisfied.

Now we shall verify the quasi-continuity of the correspondence $\tilde{x} \to \tilde{A}(\tilde{x})$ (condition 2.4.B). Suppose that the sequence of states $\tilde{x}_n = x_n v_n$ converges to a state $\tilde{x} = xv$ and that the actions $\tilde{a}_n$ belong to the fibres $\tilde{A}(\tilde{x}_n)$. Then $x_n \to x$ and $v_n \to v$. As to the $\tilde{a}_n$, we have $\tilde{a}_n = v_n a_n$, where $a_n \in A(x_n)$. Because the whole action space A is finite, infinitely many of the a_n must be equal to the same element $a \in A$, i.e. $a_{n_k} = a$ on some subsequence $\{n_k\}$. Evidently the sequence $\tilde{a}_{n_k} = v_{n_k}a$ converges to $va \in \tilde{A}(\tilde{x})$.

Condition 2.4.C requires that the transition function $\tilde{p}$ carry continuous bounded functions f on $\tilde{X}_t$ into continuous functions g on $\tilde{A}_t$ (see also 2.4.C'). For the transition function $\tilde{p}$ constructed above we have

$$g(v_{t-1}a_t) = \sum_{x_t \in X_t} f(x_t v_t)\tilde{p}(x_t | v_{t-1}a_t), \tag{9}$$

where the measures v_t are calculated according to formula (6). Since the sets A_t and X_t are finite, we need only verify that each term of the sum (9) is continuous in v_{t-1}. What we have is the product of two functions, of which the second, $\tilde{p}(x_t | v_{t-1}a_t)$, is continuous everywhere (see (5) and (3)), and the first, $f(x_t v_t)$, is bounded and continuous everywhere where the second is different from 0 (see (6)). Clearly such a product is a continuous function.

The continuity and boundedness of the reward functions (condition 2.4.D) are clear from formulas (7) and (8).

If the model Z is time homogeneous, then the model $\tilde{Z}$ is also homogeneous.

* * *

The initial distribution μ in the model Z splits into a distribution in X_m and a conditional distribution in Y_m:

$$\mu(x_m y_m) = \mu(x_m)v_m(y_m | x_m), \tag{10}$$

where

$$\mu(x_m) = \sum_{y_m \in Y_m} \mu(x_m y_m), \tag{11}$$

and v_m is found according to formula (2). Formula (2) establishes a mapping $v_m = F(x_m)$ of X_m into N_m. Denote by $\tilde{\mu}$ the probability distribution of the point $x_m F(x_m) \in \tilde{X}_m = X_m \times N_m$ which corresponds to the distribution of x_m given by (11). The measure $\tilde{\mu}$ may be considered as the initial distribution in the model $\tilde{Z}$ corresponding to the initial distribution μ in the model Z.

Given an arbitrary history $h = x_m v_m a_{m+1} x_{m+1} a_{m+2} x_{m+2} \cdots a_t x_t$ in the model Z, we may calculate $v_{m+1}, v_{m+2}, \ldots v_t$ recurrently from (6) and obtain the corresponding history $\tilde{h} = x_m v_m a_{m+1} x_{m+1} v_{m+1} \cdots a_t x_t v_t$ for the model $\tilde{Z}$*). This allows us to assign to each strategy π in the model $\tilde{Z}$ a strategy $\tilde{\pi}$ in the model Z in the following way: we obtain a probability distribution $\tilde{\pi}(\cdot | h)$ for the next action by substituting into $\pi(\cdot | \tilde{h})$ the value of $\tilde{h}$ constructed above starting from h. It is clear that one can obtain in this way any strategy $\tilde{\pi}$ in the model Z: it suffices to put $\pi(\cdot | x_m v_m a_{m+1} x_{m+1} v_{m+1} \cdots a_t x_t v_t) = \tilde{\pi}(x_m v_m a_{m+1} x_{m+1} \cdots a_t x_t)$, dropping the arguments $v_{m+1} \cdots v_t$ in the right hand side.

In order to reduce the control problem in the model Z with incomplete information to the analogous problem for the model $\tilde{Z}$, we need to show that the value $w(\mu,\tilde{\pi})$ of the strategy $\tilde{\pi}$ in the model Z coincides with the value $\tilde{w}(\tilde{\mu},\pi)$ of the strategy π in the model $\tilde{Z}$. For this it suffices to verify that

$$P_{\tilde{\mu}}^{\pi}\tilde{q}(v_t a_{t+1}) = P_\mu^{\tilde{\pi}} q(y_t a_{t+1}), \qquad P_{\tilde{\mu}}^{\pi}\tilde{r}(x_n v_n) = P_\mu^{\tilde{\pi}} r(x_n y_n). \tag{12}$$

* Formally we should write v_s two times when $s < t$, as a component of the state $\tilde{x}_s$ and as a component of the action $\tilde{a}_{s+1}$.

Both of these formulas follow from the following general fact: *for any function f, any initial distribution μ, in the model Z and any strategy π in the model $\tilde{Z}$*

$$P_\mu^{\tilde{\pi}} f(h_t y_t a_{t+1}) = P_{\tilde{\mu}}^{\pi} \tilde{f}(h_t v_t a_{t+1}), \tag{13}$$

where $h_t = x_m v_m a_{m+1} \cdots a_t x_t$ is the observed history at the time t and

$$\tilde{f}(h_t v_t a_{t+1}) = \sum_{y_t \in Y_t} f(h_t y_t a_{t+1}) v_t(y_t). \tag{14}$$

Let us prove (13). According to formula (1.3.2)

$$\begin{aligned} P_{\tilde{\mu}}^{\pi}(x_m v_m a_{m+1} x_{m+1} v_{m+1} a_{m+2} \cdots x_t v_t a_{t+1}) \\ = \tilde{\mu}(x_m v_m)\pi(a_{m+1}|x_m v_m)\tilde{p}(x_{m+1} v_{m+1}|v_m a_{m+1}) \\ \times \pi(a_{m+2}|x_m v_m a_{m+1} x_{m+1} v_{m+1}) \cdots \tilde{p}(x_t v_t|v_{t-1} a_t) \\ \times \pi(a_{t+1}|x_m v_m a_{m+1} x_{m+1} v_{m+1} \cdots a_t x_t v_t)^*. \end{aligned} \tag{15}$$

It follows from the definitions of $\tilde{\mu}$ and $\tilde{p}$ that this probability is equal to 0 except in the case when the measure v_m is the function of μ and x_m given by formula (2) and the v_s for $s > m$ are the functions of v_{s-1}, a_s and x_s given by formula (6). By the definition of the strategy $\tilde{\pi}$, for such "admissible" chains one can rewrite (15) in the form

$$\begin{aligned} P_{\tilde{\mu}}^{\pi}(x_m v_m a_{m+1} x_{m+1} v_{m+1} a_{m+2} \cdots x_t v_t a_{t+1}) \\ = \mu(x_m)\tilde{\pi}(a_{m+1}|x_m v_m)\tilde{p}(x_m|v_m a_{m+1})\tilde{\pi}(a_{m+2}|x_m v_m a_{m+1} x_{m+1}) \\ \cdots \tilde{p}(x_t|v_{t-1} a_t)\tilde{\pi}(a_{t+1}|x_m v_m a_{m+1} x_{m+1} \cdots a_t x_t). \end{aligned} \tag{16}$$

Compare this with the formula

$$\begin{aligned} P_\mu^{\tilde{\pi}}(x_m y_m a_{m+1} x_{m+1} y_{m+1} a_{m+2} \cdots x_t y_t a_{t+1}) \\ = \mu(x_m y_m)\tilde{\pi}(a_{m+1}|x_m v_m)p(x_{m+1} y_{m+1}|y_m a_{m+1}) \\ \times \tilde{\pi}(a_{m+2}|x_m v_m a_{m+1} x_{m+1}) \cdots p(x_t y_t|y_{t-1} a_t) \\ \times \tilde{\pi}(a_{t+1}|x_m v_m a_{m+1} x_{m+1} \cdots a_t x_t), \end{aligned} \tag{17}$$

which follows from (1) (here v_m is also the function of μ and x_m given by (2)). Making use of (16) and (17), we shall prove formula (13) by induction on t. For $t = m$ we need to show that

$$P_\mu^{\tilde{\pi}} f(x_m v_m y_m a_{m+1}) = P_{\tilde{\mu}}^{\pi} \sum_{y_m \in Y_m} f(x_m v_m y_m a_{m+1}) v_m(y_m). \tag{18}$$

* Formally we ought to have written $\pi(v_s a_{s+1}|x_m \cdots x_s v_s)$; we omit the first component v_s of the action $v_s a_{s+1}$, equal to the second component of the preceding state $x_s v_s$.

In view of (2), for any admissible chain $x_m v_m y_m a_{m+1}$ we have

$$\mu(x_m y_m)\tilde{\pi}(a_{m+1}|x_m v_m) = \mu(x_m)\tilde{\pi}(a_{m+1}|x_m v_m)v_m(y_m).$$

Multiplying both sides of this equation by $f(x_m v_m y_m a_{m+1})$, summing on x_m, y_m and a_{m+1}, and taking account of formulas (16) and (17), we get (18).

Further, in view of (17), the left side of formula (13) is equal to $P_{\tilde{\mu}}^{\pi} f_1(h_{t-1} y_{t-1} a_t)$, where

$$f_1(h_{t-1} y_{t-1} a_t) = \sum_{x_t y_t a_{t+1}} p(x_t y_t|y_{t-1} a_t)\tilde{\pi}(a_{t+1}|h_t) f(h_t y_t a_{t+1}),$$

and its right side, in view of (16), reduces to $P_{\tilde{\mu}}^{\pi} \tilde{f}_1(h_{t-1} v_{t-1} a_t)$, where

$$\begin{aligned} \tilde{f}_1(h_{t-1} v_{t-1} a_t) &= \sum_{x_t a_{t+1}} \tilde{p}(x_t|v_{t-1} a_t)\tilde{\pi}(a_{t+1}|h_t)\tilde{f}(h_t v_t a_{t+1}) \\ &= \sum_{x_t y_t a_{t+1}} \tilde{p}(x_t|v_{t-1} a_t)\tilde{\pi}(a_{t+1}|h_t)v_t(y_t) f(h_t y_t a_{t+1}). \end{aligned}$$

In order to obtain (13) from the induction hypothesis, it remains to verify that

$$\tilde{f}_1(h_{t-1} v_{t-1} a_t) = \sum_{y_{t-1}} f_1(h_{t-1} y_{t-1} a_t) v_{t-1}(y_{t-1}), \tag{19}$$

i.e. that (14) is satisfied when t is replaced by $t - 1$. Since we are dealing only with admissible chains, for which v_t is connected with v_{t-1} by formulas (3)–(4), then

$$\sum_{y_{t-1}} p(x_t y_t|y_{t-1} a_t) v_{t-1}(y_{t-1}) = \tilde{p}(x_t|v_{t-1} a_t) v_t(y_t).$$

Multiplying both sides by $\tilde{\pi}(a_{t+1}|h_t) f(h_t y_t a_{t+1})$ and summing on x_t, y_t, and a_{t+1}, we arrive at (19).

* * *

Let us review the situation. We have a mapping $\pi \to \tilde{\pi}$ of the strategy set for the model $\tilde{Z}$ onto the strategy set for Z such that

$$w(\mu, \tilde{\pi}) = \tilde{w}(\tilde{\mu}, \pi) \tag{20}$$

for any initial distribution μ and corresponding initial distribution $\tilde{\mu}$. Hence it follows that $v(\mu) = \tilde{v}(\tilde{\mu})$, and that the strategy $\tilde{\pi}$ is optimal for the process Z_μ if and only if the strategy π is optimal for the process $\tilde{Z}_\mu$. Accordingly, for the strategy $\tilde{\pi}$ to be uniformly optimal in the model Z it suffices that the strategy π be optimal relative to the model $\tilde{Z}$*.

* The converse is not always true since not every probability distribution on $X_m \times N_m$ can be obtained from some distribution μ on $X_m \times Y_m$. It holds if $\tilde{v}(\tilde{\mu}) = \tilde{\mu}\tilde{v}$.

We have verified that the model $\tilde{Z}$ is semicontinuous. If the interval $[m,n]$ of control is finite, then, according to the results of Chapter 2, the model $\tilde{Z}$ has a simple uniformly optimal strategy $\varphi = \psi_{m+1}\psi_{m+2}\cdots\psi_n$, where the ψ_t are (measurable) mappings of the pairs $x_{t-1}v_{t-1}$ into the $A(x_{t-1})$. In other words, there exist measurable functions

$$a_{t+1} = \psi_{t+1}(x_t v_t) \tag{21}$$

assigning, an action a_{t+1}, to each observed state x_t and to any probability distribution v_t for the nonobserved state y_t (independently of all the other information about the preceding history), and such that the strategy $\varphi = \psi_{m+1}\psi_{m+2}\cdots\psi_n$ is uniformly optimal in the model $\tilde{Z}$. This provides for a model with incomplete information the following method of construction of a strategy which is optimal for all initial distributions μ: we need only *at each stage to choose the action* $a_{t+1} = \psi_{t+1}(x_t v_t)$ *where* x_t *is the observed state and* v_t *the probability distribution for the nonobserved state* y_t, *calculated from* v_{t-1} *by* (6) *in the case* $t > m$ *and from* μ *by* (2) *in the case* $t = m$.

Now consider the control interval $[0,\infty)$. For the existence of a simple optimal strategy $\varphi = \psi_1\psi_2\cdots\psi_t\cdots$ in the model $\tilde{Z}$ it suffices to require in addition, for example, that the series

$$\sum_{t=1}^{\infty} \max_{y_{t-1}a_t} |q(y_{t-1}a_t)| \tag{22}$$

converge (see Chapter 5, §6).

This series converges in particular, *if the model* Z *is homogeneous and the discount coefficient* β *is less than* 1. According to Chapter 6, §6, in the case *the model* $\tilde{Z}$ *has a stationary optimal strategy* (i.e. the selector ψ_t is the same for all the times t).

* * *

In concrete problems it is often necessary to deal with the case when the fibres $A(x)$ intersect for different x, and the transition function and running reward function at the step t depend on x_{t-1} as well as on y_{t-1} and a_t. This case reduces to the one already investigated by the introduction of the new actions $a'_t = x_{t-1}a_t$ (cf. Chapter 1, §2).

Now suppose that the fibre $A(x)$ does not depend on x, the transition function $p(x_t y_t | x_{t-1} y_{t-1} a_t)$ and the running reward $q(x_{t-1} y_{t-1} a_t)$ do not depend on x_{t-1}, i.e. that the observed state does not affect either the possibility of control nor the further evolution of the system and the future reward. In this case the operator T in the model $\tilde{Z}$ carries any function of x, v into a function of v alone. Therefore the value $\tilde{v}(x,v)$ of the model $\tilde{Z}$ does not depend on x. It is easy to see that the selectors ψ_t in the formula (21), yielding the optimal control, may also be chosen independently of x_t.

We make use of these remarks in the next section.

§3. The Two-armed Bandit Problem

One of the simplest examples of control with incomplete data is known in the literature as *the two-armed bandit problem*. This is the name for a slot machine having two levers. After putting his money in the slot, the player pulls one of the levers. The money is either lost, or is returned with a definite gain, not depending on the lever. The levers have different gain probabilities, which we shall denote by p_1 and p_2, supposing $p_1 > p_2$. The problem is that the player does not know which of the levers is the "good" one, i.e. the one with the higher probability p_1. All he learns, at each stage, is whether he has been paid off at that stage or not.

One supposes that at the beginning of the game there is an *a priori* probability distribution for the better lever. After each try, the player calculates the *a posteriori* distribution for that lever. The fundamental result for this problem, remarkable for its simplicity and clarity, is the following: *Independently of the duration of the game, the player should at each stage pull the lever whose probability of being the good one appears at that moment to be the higher.*

In order to obtain this result, we construct a corresponding homogeneous model with incomplete information corresponding to our problem. The unobserved state y_t does not depend on t. We give it the value 1 if the left lever is the good one, and the value 2 if the right lever is the good one. We shall regard the observed state x_t at the tth stage as equal to 1 in the case of a gain and equal to 2 in the case of a loss. The action at each time consists in the choice of the left or right lever. The choice of the left lever will be denoted by $a_t = 1$, and that of the right lever by $a_t = 2$. Thus the observed and unobserved state spaces, and the action space, consist each of two elements: $X = Y = A = \{1,2\}$.

The transition function p defines the probability distribution for $x_t y_t$ depending on $y_{t-1}a_t$*. It is convenient to denote by $p_1(x)$ the probability of the outcome x for the good lever and by $p_2(x)$ the same for the bad lever, so that

$$p_i(1) = p_i, \qquad p_i(2) = 1 - p_i, \qquad i = 1, 2. \tag{1}$$

The transition function is expressed in terms of the $p_i(x)$ by the formula

$$p(xy' | ya) = \begin{cases} p_1(x) & \text{if } y' = y = a, \\ p_2(x) & \text{if } y' = y \neq a, \\ 0 & \text{if } y' \neq y. \end{cases} \tag{2}$$

The gain at each play may take on two values, depending on the construction of the machine. We denote them by d_1 and d_2, supposing that $d_1 > d_2$. In accordance with the footnote on page 201, one may replace the gain at the tth step by

* Since the control spaces (fibres) $A(x)$ intersect, in fact coincide, at different states x, then the probability distribution for $x_t y_t$ could have depended not only on $y_{t-1}a_t$, but also on x_{t-1} (see the corresponding remark in §2). In our case the value of x_{t-1} obviously does not affect this distribution.

its mathematical expectation relative to the distribution $p(\cdot|y_{t-1}a_t)$, and introduce the running reward function

$$q(ya) = \sum p_i(x)d_x, \quad \text{where } i = \begin{cases} 1 & \text{if } y = a \\ 2 & \text{if } y \neq a \end{cases} \tag{3}$$

The exact values of d_1 and d_2 are inessential for the analysis. One obtains the most compact formulas if one chooses them in such a way that

$$q(ya) = \begin{cases} p_1 - p_2 & \text{if } y = a, \\ p_2 - p_1 & \text{if } y \neq a; \end{cases} \tag{4}$$

it suffices to put $d_1 = 2 - p_1 - p_2$, $d_2 = -p_1 - p_2$. The terminal reward payoff is equal to zero.

In accordance with the general results of §2, we must pass to the model $\tilde{Z}$ with complete information. Here we are dealing with a case when the fibre $A(x)$, the transition function $p(\cdot|xya)$, and the reward function payoff $q(xya)$, all do not depend on x. The remark at the end of §2 is therefore applicable, and in the construction of the optimal controls in the model $\tilde{Z}$ we may consider the action of the operator T on functions in the space N. In accordance with the formulas of §2 and of Chapter 1, §6, we have

$$Tf(v) = \max[U_1 f(v), U_2 f(v)], \tag{5}$$

where

$$U_a f(v) = \tilde{q}(va) + \sum_{x=1}^{2} \tilde{p}(x|va) f(v') \qquad (a = 1,2), \tag{6}$$

and v' is the distribution for the unobserved parameter $y_t = y_0$, into which the distribution v transforms after applying the action a in the observed state x. By formulas (2.3)–(2.6) and (2)–(4) we have

$$\begin{aligned} \tilde{p}(x|v1) &= p_1(x)v(1) + p_2(x)v(2), \\ \tilde{p}(x|v2) &= p_2(x)v(1) + p_1(x)v(1), \\ v'(y|v1x) &= \frac{p_y(x)v(y)}{\tilde{p}(x|v1)}, \\ v'(y|v2x) &= \frac{p_y(x)v(y)}{\tilde{p}(x|v2)}, \\ \tilde{q}(va) &= (p_1 - p_2)[v(a) - v(\bar{a})], \end{aligned} \tag{7}$$

where $\bar{a} = 3 - a$.

The distribution v is defined entirely by the number

$$\delta = v(2) - v(1); \tag{8}$$

in fact,

$$v(1) = \frac{1 - \delta}{2}, \qquad v(2) = \frac{1 + \delta}{2}. \tag{9}$$

Therefore the space N of distributions on Y may be identified with the segment $\delta \in [-1,1]$.

Taking account of formulas (7)–(9), we may rewrite formula (6), defining the operators U_a, in the form

$$\begin{aligned} U_1 f(\delta) = &-2R\delta + (Q_1 - R\delta) f\left(\frac{-R + Q_1\delta}{Q_1 - R\delta}\right) \\ &+ (Q_2 + R\delta) f\left(\frac{R + Q_2\delta}{Q_2 + R\delta}\right), \\ U_2 f(\delta) = &\ 2R\delta + (Q_1 + R\delta) f\left(\frac{R + Q_1\delta}{Q_1 + R\delta}\right) \\ &+ (Q_2 - R\delta) f\left(\frac{-R + Q_2\delta}{Q_2 - R\delta}\right), \end{aligned} \tag{10}$$

where

$$\begin{aligned} Q_x &= \frac{p_1(x) + p_2(x)}{2}, \\ R &= \frac{p_1(1) - p_2(1)}{2} = \frac{p_2(2) - p_1(2)}{2} > 0. \end{aligned} \tag{11}$$

(Here, in the right sides of the expressions for the $U_a f$, the argument v' is replaced by $\delta' = v'(2) - v'(1)$, calculated using (7)).

We wish to prove the optimality of the stationary strategy defined by the selector

$$\psi(\delta) = \begin{cases} 1 & \text{if } \delta < 0, \\ 2 & \text{if } \delta \geq 0. \end{cases} \tag{12}$$

To this end we need to verify that for any n

$$T_\psi^n 0 = T^n 0. \tag{13}$$

Note that in view of (12)

$$T_\psi f(\delta) = \begin{cases} U_1 f(\delta) & \text{if } \delta < 0, \\ U_2 f(\delta) & \text{if } \delta \geq 0. \end{cases} \tag{14}$$

Put

$$f_n = T^n 0. \tag{15}$$

For the proof of (12) it suffices to verify that for any n

$$T_\psi f_n = T f_n. \tag{16}$$

Put

$$g_n = U_2 f_n - U_1 f_n. \tag{17}$$

In view of (14) and (5) formula (16) will be proved if we establish that

$$\delta g_n(\delta) \geq 0, \qquad \delta \in [-1,1]. \tag{18}$$

This last is established quite easily in the special case when $p_2 = 1 - p_1$, i.e. the probability of gain for the good lever is equal to the probability of loss for the bad lever. Indeed, it is clear from (1) and (11) that in this case $Q_1 = Q_2$, and it follows from (10) that $\delta g_n(\delta) = 4R\delta^2$. In the general case it is more convenient to establish by induction on n the following somewhat stronger statement:
(*A*): $g_n(\delta)$ *is nondecreasing, and*

$$g_n(0) = 0. \tag{19}$$

In order to carry out the induction we will need the following properties of the operators U_a:

a) $U_1 U_2 = U_2 U_1$;
b) The operator U_2 carries a nondecreasing function into a nondecreasing function.

Assertion a) is verified by an elementary calculation using formula (10). It has the following interpretation: if we play twice, pulling first one lever and then the other, the result does not depend on the order of these choices.

Since the function $2R\delta$ is nondecreasing, it suffices to verify assertion b) for the operator

$$Sf(\delta) = U_2 f(\delta) - 2R\delta = \lambda(\delta) f[\alpha(\delta)] + \mu(\delta) f[\beta(\delta)],$$

where

$$\lambda(\delta) = Q_1 + R\delta, \qquad \mu(\delta) = Q_2 - R\delta,$$

$$\alpha(\delta) = \frac{R + Q_1\delta}{\lambda(\delta)}, \qquad \beta(\delta) = \frac{-R + Q_2\delta}{\mu(\delta)}.$$

We verify directly that on the segment $[-1,1]$ the functions λ, α, and β are increasing, that $\alpha \geq \beta$ and $\lambda + \mu = 1$. We have depicted the graphs of α and β in figure 8.1. If $-1 \leq \delta_1 < \delta_2 \leq 1$ we have $\lambda(\delta_2) - \lambda(\delta_1) = \mu(\delta_1) - \mu(\delta_2)$, so that

$$\begin{aligned} Sf(\delta_2) - Sf(\delta_1) = {} & [\lambda(\delta_2) - \lambda(\delta_1)]\{f[\alpha(\delta_2)] - f[\beta(\delta_2)]\} \\ & + \lambda(\delta_1)\{f[\alpha(\delta_2)] - f[\alpha(\delta_1)]\} \\ & + \mu(\delta_1)\{f[\beta(\delta_2)] - f[\beta(\delta_1)]\}. \end{aligned}$$

If f is a nondecreasing function, then all the terms in square brackets are nonnegative, so that the function Sf is also nondecreasing.

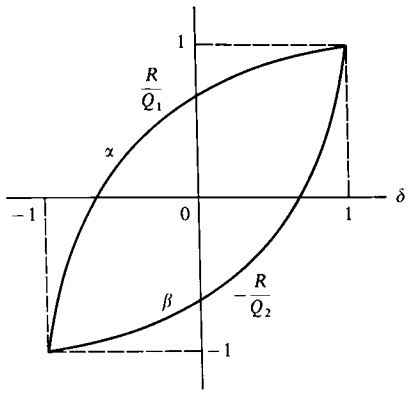

Figure 8.1

Now we turn to the proof of proposition (A). For $n = 0$ we have $f_0 = 0$ and $g_0 = U_2 0 - U_1 0 = 4R\delta$; this function satisfies (A).

Suppose that (A) is valid for some $n \geq 0$, and let us prove that then it is true for $n + 1$ as well. It is clear from (10) that

$$U_a(F_2 - F_1) = (-1)^a 2R\delta + U_a F_2 - U_a F_1.$$

Therefore, using a), we may rewrite the function $g_{n+1} = U_2 f_{n+1} - U_1 f_{n+1} = U_2 T f_n - U_1 T f_n$ in the form

$$g_{n+1} = U_2 \Phi + U_1 \Psi, \tag{20}$$

where

$$\Phi = T f_n - U_1 f_n, \qquad \Psi = U_2 f_n - T f_n. \tag{21}$$

From the induction hypothesis and formulas (5) and (17), it follows that

$$\Phi(\delta) = -\Psi(-\delta) = \begin{cases} 0 & \text{for } \delta \leq 0, \\ g_n(\delta) & \text{for } \delta \geq 0, \end{cases} \tag{22}$$

and that the function Φ is nondecreasing. Formulas (22) and (10) imply that

$$U_1 \Psi(\delta) = -U_2 \Phi(-\delta),$$

so that formula (20) takes on the form

$$g_{n+1}(\delta) = U_2 \Phi(\delta) - U_2 \Phi(-\delta). \tag{23}$$

In view of b) the function $U_2\Phi$ is nondecreasing, and it follows from (23) that the function g_{n+1} has this property as well. If $\delta = 0$ we find from (23) that $g_{n+1}(\delta) = 0$. Thus proposition (A) is valid for $n + 1$ as well.

The optimality of the stationary strategy generated by the selector (12) is proved.

§4. Reduction to a Model with Complete Information. The General Case

Up to now we have been supposing that the state and action spaces are finite. This assumption is too restrictive since the most natural applications lead to more general spaces (as in most of the examples considered in the preceding chapters). The basic idea of §2 was that of considering the pair xv as a state, where x was the observed state and v was the probability distribution on the space of unobserved states. This idea is applicable also in the general case, but its realization is technically more complicated, since in place of elementary calculations with conditional probabilities in finite spaces it is necessary to deal with the more complicated theory of conditional distributions presented in Appendix 4.

So, we suppose that X_t, Y_t, and A_t are arbitrary Borel spaces, that the transition function $p_t(dx_t\, dy_t | y_{t-1}a_t)$ and the running reward function $q(y_{t-1}a_t)$ are measurable in $y_{t-1}a_t$, and that the terminal reward function $r(x_n y_n)$ is measurable in $x_n y_n$. We shall also suppose that these spaces and functions have all of the additional properties contained in requirements 2.2.α) 2.2.ε). The strategy π now depends not only on the observed history, but also on the initial distribution v_m on the set Y_m, and must be measurable in the set of all these arguments.

The distribution in the space of paths corresponding to the initial distribution μ and the strategy π, is given by the formula

$$P_\mu^\pi(dx_m\, dy_m\, da_{m+1}\, dx_{m+1}\, dy_{m+1} \cdots da_n\, dx_n\, dy_n) = \mu(dx_m\, dy_m)\pi(da_{m+1}|x_m v_m)p(dx_{m+1}\, dy_{m+1}|y_m a_{m+1}) \cdots \pi(da_n|x_m v_m a_{m+1} x_{m+1} \cdots a_{n-1}x_{n-1})p(dx_n\, dy_n|y_{n-1}a_n). \quad (1)$$

Here v_m is the conditional distribution of y_m for a given x_m. This is a measurable function of x_m, satisfying the equation

$$\mu(dx_m\, dy_m) = \mu(dx_m)v_m(dy_m|x_m) \quad (2)$$

(see Appendix 4; formula (2) does not define the function v_m uniquely, but we fix a version of this function).

The space N_t of probability measures on Y_t is also a Borel space (see Appendix 5). As in the discrete case, the new transition function assigns to each value $\tilde{a}_t = v_{t-1}a_t$ a distribution in the space $X_t \times N_t$, concentrated on the pairs $x_t v_t$, where v_t is a single valued function of x_t (whose structure depends on $\tilde{a}_t$). As in §2, we start out

from the distribution

$$\tilde{p}(dx_t\, dy_t|\tilde{a}_t) = \int_{Y_{t-1}} p(dx_t\, dy_t|y_{t-1}a_t)v_{t-1}(dy_{t-1}) \quad (3)$$

in the space $X_t \times Y_t$ (see formula (2.3)). Formulas (2.4)–(2.5) are replaced by

$$\tilde{p}(dx_t\, dy_t|\tilde{a}_t) = \tilde{p}(dx_t|\tilde{a}_t)v_t(dy_t|\tilde{a}_t x_t) \quad (4)$$

and

$$\tilde{p}(dx_t|\tilde{a}_t) = \tilde{p}(dx_t \times Y_t|\tilde{a}_t). \quad (5)$$

In place of the elementary formula (2.6) for the definition of the measures $v_t(\cdot|\tilde{a}_t x_t)$ we must now make use of the results of Appendix 4. In view of Lemma 2 of §1 of Appendix 5 the measure (3) and hence also the measure (5), depend in a measurable way on $\tilde{a}_t$. Therefore $v_t(\cdot|\tilde{a}_t x_t)$ may be regarded as measurable relative to $\tilde{a}_t$ and x_t taken together (see the footnote on page 263. Hence one easily deduces that the measure in the space $X_t \times N_t$ given by the pair $\tilde{p}(dx_t|\tilde{a}_t)$, $v_t(\cdot|\tilde{a}_t x_t)$ is also measurable relative to $\tilde{a}_t$.

Let $\mu_a(dx)$ be a measure on X, depending measurably on a, and $\varphi_a(x)$ be a measurable mapping of the product $A \times X$ into the space E. Then the image $\tilde{\mu}_a$ of the measure μ under the mapping φ_a depends measurably on a. To see this suppose that f is any measurable function on the space E. Then $f[\varphi_a(x)]$ is measurable in xa, and according to Lemma 2 of §1 of Appendix 5 the function

$$F(a,a') = \int_X f[\varphi_a(x)]\mu_{a'}(dx) \quad (6)$$

is measurable in aa'. Accordingly, $F(a,a)$ is measurable in a. But if $f = \chi_\Gamma$ we have $F(a,a) = \tilde{\mu}_a(\Gamma)$.

The running and terminal reward in the new model are given by the formulas

$$\tilde{q}(v_{t-1}a_t) = \int_{Y_{t-1}} q(y_{t-1}a_t)v_{t-1}(dy_{t-1}),$$
$$\tilde{r}(x_n v_n) = \int_{Y_n} r(x_n y_n)v_n(dy_n). \quad (7)$$

It is easy to verify that the functions $\tilde{p}$, $\tilde{q}$, and $\tilde{r}$ satisfy conditions 2.2.α)–2.2.ε). Denote the Borel model with complete information defined by them by $\tilde{Z}$. If Z is homogeneous, then $\tilde{Z}$ is homogeneous as well. We leave it to the reader to verify that if Z is nontrivial then $\tilde{Z}$ is also nontrivial.

* * *

We assign to each initial distribution μ in the model Z an initial distribution $\tilde{\mu}$ in the model $\tilde{Z}$, as follows. Suppose that $v_m(\cdot|x_m)$ is the measure in the space

Y_m introduced at formula (2). The measurable mapping

$$x_m \to \nu_m(\cdot \mid x_m)$$

of the space X_m into the space N_m induces the measurable mapping

$$x_m y_m \to x_m \nu_m(\cdot \mid x_m) \tag{8}$$

of the product $X_m \times Y_m$ into the product $X_m \times N_m$. The image of the measure μ under the mapping (8) is the desired initial distribution $\tilde{\mu}$ in $\tilde{Z}$.

The mapping $\pi \to \tilde{\pi}$ of strategies in the model $\tilde{Z}$ into strategies in the model Z is given as in §2, except that now ν_t is found recurrently from the decomposition (4). In order to obtain the basic equation

$$w(\mu,\tilde{\pi}) = \tilde{w}(\tilde{\mu},\pi), \tag{9}$$

we need to show that for any bounded measurable function f and any μ, π

$$P_{\mu}^{\tilde{\pi}} f(h_t y_t a_{t+1}) = P_{\tilde{\mu}}^{\pi} \tilde{f}(h_t \nu_t a_{t+1}), \tag{10}$$

where $h_t = x_m \nu_m a_{m+1} x_{m+1} \cdots a_t x_t$ and

$$\tilde{f}(h_t \nu_t a_{t+1}) = \int_{Y_t} f(h_t y_t a_{t+1}) \nu_t(dy_t). \tag{11}$$

This is proved in the same way as the analogous assertion in §2, with sums replaced by integrals.

* * *

As in §2, it follows from equation (9) that *if the strategy π is optimal in the auxiliary model $\tilde{Z}$ with complete information, then the corresponding strategy $\tilde{\pi}$ is optimal in the model Z.* The same is true for ε-optimal strategies. For the optimality (ε-optimality) of a strategy $\tilde{\pi}$ under the initial distribution μ, it is necessary and sufficient that π should be optimal (ε-optimal) under the corresponding initial distribution $\tilde{\mu}$. If π is stationary, then $\tilde{\pi}$ is stationary as well.

We have proved relation (9) under the assumption that the control interval is finite and that the reward functions q and r are bounded above. It is easy to see that (9) remains valid for nonnegative unbounded reward functions and for an infinite control interval. If the reward function takes values of different signs, then it is useful to consider its positive and negative parts. Note that for any function q

$$\tilde{q}^{+} \le \widetilde{q^{+}}, \qquad \tilde{q}^{-} \le \widetilde{q^{-}}, \tag{12}$$

the wavy line denoting the operation defined by formula (7). Therefore, if the model Z is μ-summable above (below), the model $\tilde{Z}$ is $\tilde{\mu}$-summable above (below).

An initial state in the model $\tilde{Z}$ is a pair $x\nu$, where x is the observed initial state in the model Z and ν is the apriori distribution for the unobserved initial state. Therefore, for the model $\tilde{Z}$ to be summable above (below), it is sufficient that for any x and ν the model Z should be $x\nu$-summable above (below). Taking account of inequalities (12) and the formula (2.12), we see that the boundedness above (below) of the model Z implies the analogous property for the model $\tilde{Z}$. Here, by boundedness above of a model with incomplete information, we mean the existence of positive functions $c_t(x\nu)$ such that, for sufficiently large t, for any strategy $\tilde{\pi}$

$$P_{x\nu}^{\tilde{\pi}} q^{+}(y_{t-1} a_t) \le c_t(x\nu) \qquad (x \in X_m, \nu \in N_m),$$

and the series Σc_t converges at each point $x\nu$. (Boundedness below of Z is defined analogously).

Combining these results with the results of Chapters 3–6, we may obtain various conditions for existence of optimal strategies in Z. For example, it follows from Result II'a of Chapter 5, §1 that *if the model Z is μ-summable above and bounded above, then for any $\varepsilon > 0$ there exists a simple strategy φ* (in the model $\tilde{Z}$) *such that $w(\mu,\tilde{\varphi}) \ge v(\mu) - \varepsilon$.* Or it follows from Chapter 6, §8 that *if in the homogeneous model Z the reward q is bounded and the discount coefficient β is less than* 1, *then for any $\varepsilon > 0$ and any initial distribution μ there exists a stationary strategy $\tilde{\varphi}$ such that $w(\mu,\tilde{\varphi}) \ge v(\mu) - \varepsilon$.*

§5. The Stabilization Problem

Now we turn to the stabilization problem, treated previously in Chapter 1, §2, Chapter 2, §11, Chapter 6, §12, and Chapter 7, §12. We will suppose that at each time t the state of the system is observed with some error ξ_t. As everywhere else in this chapter, we denote the observed state by x. It is connected with the true state y_t by the formula

$$x_t = y_t + \xi_t, \qquad t = 0, 1, 2, \ldots. \tag{1}$$

The actions and the random perturbations of the system are denoted as before respectively by a_t and s_t. Thus, the recurrence equation describing the actual (unobserved) evolution of the system now has the form

$$y_t = y_{t-1} - a_t + s_t, \tag{2}$$

and the running reward is equal to

$$q(y_{t-1} a_t) = -b(y_{t-1} - a_t)^2 - c a_t^2 \qquad (t = 1,2,\ldots) \tag{3}$$

(in the case of complete information we have $y_t = x_t$). It remains to assume how does the process start. We will suppose that the control begins at an instant when

the system is taken out of equilibrium by a random perturbation s_0, so that

$$y_0 = s_0. \tag{4}$$

Complete results have been obtained only under the hypothesis that *all the random variables* $s_0, \xi_0, s_1, \xi_1, \ldots$ *are normally distributed.* Suppose moreover that *they are mutually independent*, and that

$$Es_t = E\xi_t = 0, \qquad \operatorname{Var} s_t = \sigma^2, \qquad \operatorname{Var} \xi_t = \tau^2. \tag{5}$$

Without loss of generality we may suppose that $\tau = 1$ (changing the units of measurement if necessary).

The linear operations (1) and (2) do not lead out of the class of normal distributions. If (η_1,η_2) is a normally distributed random vector with the parameters

$$E\eta_i = c_i, \qquad E(\eta_i - c_i)(\eta_j - c_j) = b_{ij} \qquad (i,j = 1,2),$$

then the conditional distribution of η_2 for a known value of η_1 is also normal with the parameters*.

$$\begin{aligned} E(\eta_2|\eta_1) &= c_2 + \frac{b_{12}}{b_{11}}(\eta_2 - c_1), \\ \operatorname{Var}(\eta_2|\eta_1) &= b_{22}\left(1 - \frac{b_{12}^2}{b_{11}b_{22}}\right) \end{aligned} \tag{6}$$

Therefore we have to deal only with the normal distributions ν_t for the unobserved states y_t. A normal distribution is defined by two parameters—the mathematical expectation m and the variance D—so that the space N_t may be identified with the half-plane $N = \{(m,D): D \geq 0\}$.

Let us describe the remaining elements of the auxiliary model $\tilde{Z}$ introduced in §2 and §4. The initial distribution $\tilde{\mu}$ in the space $X \times N$ is constructed starting from the joint distribution μ of the pairs $(x_0,y_0) = (s_0 + \xi_0, s_0)$. This last is normal and in view of (5) has the parameters

$$Ex_0 = Ey_0 = 0, \qquad \operatorname{Var} x_0 = \sigma^2 + 1, \qquad \operatorname{Var} y_0 = \sigma^2, \qquad Ex_0y_0 = \sigma^2. \tag{7}$$

Therefore, according to formulas (6), the conditional distribution $\nu_0(\cdot|x_0)$ has the parameters

$$m_0 = \frac{\sigma^2}{\sigma^2 + 1}x_0, \qquad D_0 = \frac{\sigma^2}{\sigma^2 + 1}. \tag{8}$$

The transition function $\tilde{p}$ assigns to each pair $\nu_{t-1}a_t = (m_{t-1},D_{t-1})a_t$ a probability distribution for x_t and a conditional distribution $\nu_t(\cdot|x_t) = (m_t(x_t),D_t(x_t))$. In view of formulas (1)–(2) and (5) the normal distribution $\tilde{p}(\cdot|\nu_{t-1}a_t)$ of the pair

* See H. Cramér [1], Chapter 21, Section 12.

$x_t y_t$ has the parameters

$$\begin{aligned} Ex_t = Ey_t &= m_{t-1} - a_t, \\ \operatorname{Var} y_t &= D_{t-1} + \sigma^2, \\ \operatorname{Var} x_t &= D_{t-1} + \sigma^2 + 1, \\ E(x_t - Ex_t)(y_t - Ey_t) &= D_{t-1} + \sigma^2. \end{aligned} \tag{9}$$

The parameters of the normal distribution $\tilde{p}(dx_t|\nu_{t-1}a_t)$ are contained in formulas (9), and the parameters of the normal distribution $\nu_t(dy_t|\nu_{t-1}a_tx_t)$ are, by formulas (6) and (9), equal to

$$m_t = m_{t-1} - a_t + \frac{D_{t-1} + \sigma^2}{D_{t-1} + \sigma^2 + 1}(x_t - m_{t-1} + a_t), \tag{10}$$

$$D_t = \frac{D_{t-1} + \sigma^2}{D_{t-1} + \sigma^2 + 1}. \tag{11}$$

Formulas (10)–(11) are valid for $t = 0$ as well, if we put

$$m_{-1} = a_0 = D_{-1} = 0. \tag{12}$$

For the running reward $\tilde{q}$, according to formulas (4.7) and (3) we have the expression

$$\tilde{q}(\nu_{t-1}a_t) = Eq(y_{t-1}a_t) = -bD_{t-1} - b(m_{t-1} - a_t)^2 - ca_t^2. \tag{13}$$

According to the general theory, for control at the instant t it is essential to know only a_s, x_s, m_s, and D_s for $s < t$. The variances D_t are computed according to formulas (8) and (11), independently of the observations. On the other hand, by formulas (8) and (10) we may express $x_0, x_1, \ldots, x_t$ in terms of $m_0, m_1, \ldots, m_t$ and $a_1, \ldots, a_t$. Therefore it suffices to watch the evolution of m_t only. It follows from formulas (1), (2), (8), (10) and (11) that

$$m_t = m_{t-1} - a_t + \tilde{s}_t \qquad (t = 0,1,2,\ldots), \tag{14}$$

where

$$\tilde{s}_t = D_t(y_{t-1} - m_{t-1} + s_t + \xi_t) \qquad (t = 0,1,2,\ldots). \tag{15}$$

(Here we are supposing that $y_{-1} = 0$.) Since the constant terms in the running reward do not affect the difference $w(x,\pi) - w(x,\rho)$ of the values of any two strategies, then in the search for an optimal strategy these terms may be dropped, and the running reward (13) replaced by

$$\tilde{\tilde{q}}(m_{t-1}a_t) = -b(m_{t-1} - a_t)^2 - ca_t^2. \tag{16}$$

Formulas (14)–(16) define a model with complete information, in which the states are the numbers m_t; this is the stabilization problem with complete information which we studied earlier, except that now the perturbations $\tilde{s}_t$ are different.

In the preceding chapters we supposed that the perturbations were independent, identically distributed, and had zero mathematical expectations. We shall show that all of these properties are satisfied for the $\tilde{s}_t$, except the property of being identically distributed.

The difference

$$z_t = y_t - m_t$$

is normally distributed with parameters $(0, D_t)$. Indeed, since ν_t is the conditional distribution for y_t under the observed history h, then

$$m_t = E(y_t|h), \qquad D_t = E[(y_t - m_t)^2|h].$$

Therefore

$$Em_t = EE(y_t|h) = Ey_t \tag{17}$$

and

$$D_t = EE[(y_t - m_t)^2|h] = E(y_t - m_t)^2 = Ez_t^2. \tag{18}$$

It is easy to deduce formulas (17)–(18) also by induction from the recurrence relation for z_t which follows from (1)–(2) and (10)–(11):

$$z_t = (1 - D_t)z_{t-1} + \zeta_t \qquad (t = 0,1,2,\ldots), \tag{19}$$

where

$$z_{-1} = 0 \tag{20}$$

and

$$\zeta_t = (1 - D_t)s_t - D_t\xi_t. \tag{21}$$

Using formulas (17)–(21), we show that the random variables $\tilde{s}_t$ are non-correlated and accordingly independent. Put

$$Q_t^s = \begin{cases} \prod_{k=s}^{t} (1 - D_k) & \text{for } s \le t, \\ 1 & \text{for } s > t. \end{cases} \tag{22}$$

One easily deduces from (19) that for $T > t$

$$z_{T-1} = Q_{T-1}^t z_{t-1} + Q_{T-1}^{t+1}\zeta_t + Q_{T-1}^{t+2}\zeta_{t+1} + \cdots + Q_{T-1}^T\zeta_{T-1}. \tag{23}$$

It follows from (15), (17) and (5) that

$$E\tilde{s}_t = 0. \tag{24}$$

From (15), (21) and (23), using the orthogonality of $z_{t-1}, s_t, \zeta_t, s_t, \ldots, s_{T-1}$ and ζ_{T-1}, and formulas (5), we find for $0 \le t < T$ that

$$\begin{aligned} \frac{1}{D_tD_T} E\tilde{s}_t\tilde{s}_T &= E(z_{t-1} + s_t + \xi_t)(z_{T-1} + s_T + \zeta_T) \\ &= Q_{T-1}^t Ez_{t-1}^2 + Q_{T-1}^{t+1}E(s_t + \xi_t)\zeta_t \\ &= Q_{T-1}^{t+1}[(1 - D_t)D_{t-1} + (1 - D_t)\sigma^2 - D_t]. \end{aligned}$$

In view of (11) the term in square brackets is equal to 0. From (15), (18) and (5) we have

$$\operatorname{Var}\tilde{s}_t = D_t^2(D_{t-1} + \sigma^2 + 1) = \frac{D_t^2}{1 - D_t}; \tag{25}$$

cf. (11).

The original problem has been reduced to the problem of control of a system given by the recurrence relation (14) *with the independent random perturbations* $\tilde{s}_t$. For the case of constant variance of the perturbations this last problem was solved in Chapter 2, §11 for a finite interval of control, in Chapter 6, §12 for an infinite interval of control and a discounted reward, and in Chapter 7, §12 for an averaged reward per unit time.

It is easy to see that in the general case when the random perturbations have nonequal distributions, the optimal strategies remain the same, and the value of the model is changed by a constant. For example, in the problem of maximization of the average reward the asymptotic value v of the model is given by the formula

$$v = -bD - \sigma^2 l, \tag{26}$$

where l is the positive root of the equation

$$l^2 + bl - bc = 0, \tag{27}$$

and $D = \lim_{t\to\infty} D_t$ is the positive root of the equation

$$D^2 + \sigma^2 D - \sigma^2 = 0. \tag{28}$$

(See formula (7.12.16), we leave the verification of this to the reader.)

* * *

The conditional mathematical expectation m_t of the random variable y_t is a natural estimate of y_t in terms of the observed history h (it is a function of h for which the quantity $E[y_t - f(h)]$ is minimal). In this section, we have, on a simple example, obtained the *separation theorem*, which asserts that under rather general conditions the optimal control of a linear Gaussian system with a quadratic loss function splits into: 1) the computation of the best estimates of the unobserved parameters in terms of the observed ones, and 2) the optimal control of the system resulting from the original one by the replacement of the unobserved parameters by their estimates.