Chapter 4

Discrete Models

§1. Passage to an Infinite Interval of Control

When there is no natural moment for stopping a process, it is appropriate to consider control over an infinite time interval.

The problem of optimal control over an infinite time interval may be posed in various ways. One may seek a maximization of the average gain per unit time; to this we have devoted Chapter 7. In this chapter we maximize the total mean value of the reward I across an infinite time. Such an approach is interesting in the first place when the values of I are bounded above.

In this chapter we deal with discrete, i.e. finite and countable, models. The general case, requiring a more thorough acquaintance with measurability problems, and making use of the material of Chapters 2 and 3, will be taken up in Chapter 5.

§2. Summable Models

The definitions of a controlled Markov process and of a model do not change for an infinite control interval $[m,\infty)$, except that now the state spaces $X_m, X_{m+1}, \ldots$ and the action spaces $A_{m+1}, A_{m+2}, \ldots$ form infinite sequences, and there is no terminal payoff. It is also necessary to give strategies for histories of arbitrarily long length.

In Chapter 1 the value of a strategy π with the initial distribution μ was defined by the formula

$$w(\mu,\pi) = P\left[\sum_{m+1}^{n} q(a_t) + r(x_n)\right] = \sum_{m+1}^{n} Pq(a_t) + Pr(x_n)$$

where P is the measure in the space of paths defined by equation (1.3.2). In the case of an infinite interval it is natural to put

$$w(\mu,\pi) = \sum_{m+1}^{\infty} Pq(a_t). \tag{1}$$

Here $Pq(a_t)$ can be calculated according to formulas (1.3.2)–(1.3.3), breaking off the trajectory $x_m a_{m+1} x_{m+1} \cdots a_n x_n \cdots$ at x_n for an arbitrary $n \geq t$ (one easily sees that the value of $Pq(a_t)$ does not depend on n).

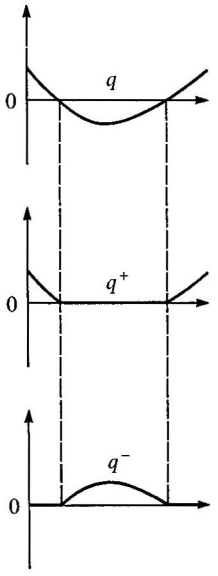

Figure 4.1

Let us put $q^+ = \max\{q,0\}$, $q^- = \max\{-q,0\}$ (see figure 4.1). The sum of the series (1) may fail to exist. But the series

$$\sum_{m+1}^{\infty} Pq^+(a_t) = w^+(\mu,\pi) \tag{2}$$

and

$$\sum_{m+1}^{\infty} Pq^-(a_t) = w^-(\mu,\pi), \tag{3}$$

always have definite sums, finite or $+\infty$. We will say that the model is *μ-summable above* if $w^+(\mu,\pi) < +\infty$ for all π, and that it is *μ-summable below* if $w^-(\mu,\pi) < +\infty$ for all π. The formulations for both cases often turn out to be quite symmetric. In such situations we shall frequently speak of "μ-summability", dropping the qualification "above" or "below". Formulations of this kind may be understood in two ways, either always with the word "above", or always with the word "below"

If the model is μ-summable, then

$$\sum_{m+1}^{\infty} Pq^+(a_t) - \sum_{m+1}^{\infty} Pq^-(a_t) = \sum_{m+1}^{\infty} [Pq^+(a_t) - Pq^-(a_t)]$$

$$= \sum_{m+1}^{\infty} P[q^+(a_t) - q^-(a_t)]. \tag{4}$$

The validity of these equalities is a consequence of the following general property of numerical series.

Property S. *If the sum of the positive terms, or the sum of the negative terms in a series, is finite, then the sum of the series has a meaning (finite or equal to* $-\infty$ *or* $+\infty$*) and does not change if the terms are rearranged, or grouped in parentheses in any manner (the number of parentheses and the number of terms in each parenthesis may be infinite).*

The right hand sides of (1) and (4) coincide because $q = q^+ - q^-$. Thus, for a μ-summable model formula (1) has a meaning and

$$w(\mu,\pi) = w^+(\mu,\pi) - w^-(\mu,\pi). \tag{5}$$

Models on a finite interval $[m,n]$ may be considered as special cases of models on the infinite interval $[m,+\infty)$; it suffices to put

$$q(a) = \begin{cases} r(ja) & \text{if } a \in A_{n+1}, \\ 0 & \text{if } a \in A_{n+2} \cup A_{n+3} \cup \cdots. \end{cases}$$

For models on a finite interval the μ-summability above follows from the fact that the functions q and r are bounded above (see condition α of Chapter 1, §12). On an infinite interval this is already not so. Therefore it makes no sense to introduce that condition, and *we will exclude it from our initial premises.* As a consequence, along with the passage to an infinite interval of control we will obtain some strengthened results for the finite interval.

§3. The Fundamental Equation

We shall show that the formulas

$$w(\mu,\pi) = \sum_{x \in X_m} \mu(x) w(x,\pi) \tag{1}$$

and

$$w(x,\pi) = \sum_{a \in A(x)} \pi(a|x)[q(a) + w'(p_a,\pi_a)] \tag{2}$$

(the fundamental equation) established in Chapter 1, §§4,5 and 12 hold now as well. Precisely:

a) *If the model is μ-summable, then it is x-summable* for all x with $\mu(x) > 0$, and equation (1) is satisfied.*

b) *If the model is x-summable, then the derived model is p_a-summable for all $a \in A(x)$, and equation (2) is satisfied.*

* We will say that the model is *x-summable* when it is summable relative to the μ-distribution concentrated at the point x.

Let us consider first the case of a nonnegative reward function q. Note that if $q = 0$ on all the sets A_t with $t > n$, then the choice of the control after the time n plays no rôle, and the situation reduces to control on the segment $[m,n]$ (for a reward bounded below). Therefore for the reward function q_n defined by

$$q_n = \begin{cases} q & \text{on } A_{m+1} \cup A_{m+2} \cup \cdots \cup A_n \\ 0 & \text{on } A_{n+1} \cup A_{n+2} \cup \cdots \end{cases}$$

relations (1) and (2) follow from the results of §§4,5, and §12 of Chapter 1. As $n \to \infty$ the nonnegative function q_n tends monotonically to q. Under such convergence it is legitimate to pass to the limit both under the expectation sign P and under the summation sign in the series (2.1). This means that for arbitrary μ and π the value $w_n(\mu,\pi)$ with the reward q_n converges nondecreasingly to the value $w(\mu,\pi)$ for the reward q, and the same is true for the derived model. With this convergence, termwise passage to the limit under the summation sign in formulas (1) and (2) is legitimate, and we find that these formulas are valid for any nonnegative reward q.

Now suppose that q can take on values of any sign. We have proved that formula (1) is satisfied for the values w^+ and w^- corresponding to the nonnegative rewards q^+ and q^-:

$$w^+(\mu,\pi) = \sum_{X_m} \mu(x) w^+(x,\pi), \tag{3}$$

$$w^-(\mu,\pi) = \sum_{X_m} \mu(x) w^-(x,\pi). \tag{4}$$

Accordingly, if $w^+(\mu,\pi) < +\infty$, then $w^+(x,\pi) < +\infty$ for all x with $\mu(x) > 0$, and the same is valid for w^-. Taking account of (4) and (3) and employing property S, we obtain equation (1).

Further, suppose that $w^+(x,\pi) < +\infty$ for all strategies π. Fix some $a \in A(x)$ and any strategy π' in the derived model. Let ψ_a be a selector of the correspondence $A(y)$, $y \in X_m$, which assigns to the point x the fixed action a. Applying formula (2) to the strategy $\pi = \psi_a \pi'$ and the nonnegative reward function q^+, we get

$$w^+(x,\pi) = q^+(a) + w^{+\prime}(p_a,\pi_a).$$

But $\pi_a = \pi'$, and hence $w^{+\prime}(p_a,\pi') < \infty$. Applying the analogous arguments to w^-, we obtain the first half of assertion b).

Now suppose that π is any strategy. Applying formula (2) to the reward functions q^+ and q^-, and subtracting one from the other with the aid of property S, we conclude that (2) is satisfied for the function q as well.

* * *

Now that we have extended the fundamental equation to the general case, we may use all the consequences of that equation. In particular, if the model is μ-summable and if ψ_t is a selector of the correspondence $A(x)$, $x \in X_{t-1}$ $(t = m +$

$1, \ldots, n)$, and π is any strategy in the derived model of order $n - m$, then

$$w(x,\psi_{m+1}\psi_{m+2}\cdots\psi_n\pi) = T_{\psi_{m+1}}T_{\psi_{m+2}}\cdots T_{\psi_n}w_n(x,\pi). \tag{5}$$

(cf. formula (1.7.5)). Here the quantity (5) is less than $+\infty$ for either q^+ or q^-.

For a μ-summable model on the finite interval $[m,n]$ with reward functions q and r, formula (5) takes the form

$$w(x,\psi_{m+1}\psi_{m+2}\cdots\psi_n) = T_{\psi_{m+1}}T_{\psi_{m+2}}\cdots T_{\psi_n}r(x)$$

or, if we make use of the formula preceding (2.1),

$$T_{\psi_{m+1}}T_{\psi_{m+2}}\cdots T_{\psi_n}r(x) = \sum_{m+1}^{n} P_x^{\varphi}q(a_t) + P_x^{\varphi}r(x_n), \tag{6}$$

where φ is the simple strategy given by $\varphi = \psi_{m+1}\psi_{m+2}\cdots\psi_n$. Obviously formula (6) may be applied as well to a model Z on the infinite interval $[m,\infty)$ for any $n > m$ and any function r on X_n, if the "truncated" model $Z^n(r)$ is μ-summable ($Z^n(r)$ is obtained by breaking Z at time n and taking the terminal reward to be r).

Further, for any strategy ρ on the segment $[m,n]$,

$$w(\mu,\rho\pi) = \sum_{m+1}^{n} P_{\mu}^{\rho}q(a_t) + P_{\mu}^{\rho}w(x_n,\pi). \tag{7}$$

For models on a finite interval this formula was proved in Chapter 1, §8. The passage to an infinite time interval in the case $q \geq 0$ is carried out in the same way as in the proof of assertions a) and b). Finally, for a μ-summable model and an arbitrary reward function, formula (7) is obtained by subtracting the corresponding formulas for q^+ and q^-.

Applying formula (7) to the reward q^+, we note that if the model Z is μ-summable then the quantity

$$P_{\mu}^{\rho}w^+(x_n,\pi) = w^+(\nu,\pi),$$

is finite, where

$$\nu(y) = P_{\mu}^{\rho}\{x_n = y\} \qquad (y \in X_n).$$

Therefore the corresponding derived model is ν-summable.

§4. Uniformly ε-Optimal Strategies

For a μ-summable model, the value of the initial distribution μ defined by the formula

$$v(\mu) = \sup_{\pi} w(\mu,\pi)$$

has a meaning, and the definitions of optimal and of ε-optimal strategies for the process Z_{μ}, given in §§3 and 12 of Chapter 1, need no changes.

A uniformly ε-optimal strategy was defined in Chapter 1 as a strategy π satisfying the condition

$$w(\mu,\pi) \geq v(\mu) - \varepsilon \tag{1}$$

for all initial distributions μ. We proved there that this condition was equivalent to the requirement that

$$w(x,\pi) \geq v(x) - \varepsilon \tag{2}$$

for all $x \in X_m$. Now, these two conditions are no longer equivalent: if $w(x,\pi)$ has a meaning for all $x \in X_m$, this does not mean that $w(\mu,\pi)$ is defined for every μ. Therefore *we shall adopt condition* (2) *as the definition of* ε*-optimality, restricting ourselves to the class of those models which are x-summable for all* $x \in X_m$.

We shall show that:

a) *For any* $\varepsilon > 0$ *there exists a uniformly* ε*-optimal strategy* π;

b) *If the model is* μ*-summable, then the function* $v(x)$ *is also* μ*-summable* and*

$$v(\mu) = \sum_{X_m} \mu(x)v(x) \qquad (= \mu v). \tag{3}$$

c) *If the strategy* π *is uniformly* ε*-optimal, then* $w(\mu,\pi) \geq v(\mu) - \varepsilon$ *for all those* μ *for which the model is* μ*-summable.*

To prove Result a) it is sufficient to construct, for any $x \in X_m$, a strategy $\pi = \pi_x$ such that

$$w(x,\pi) \geq v(x) - \varepsilon,$$

and then to combine those π_x according to §4 of Chapter 1. Fix some x, and put, for brevity, $\pi_x = \pi$. If $v(x) < +\infty$, the existence of the required strategy π follows from the definition of $v(x)$. If $v(x) = +\infty$, then it follows from the definition of $v(x)$ that for any natural number k there exists a strategy π_k for which

$$w(x,\pi_k) \geq 2^k. \tag{4}$$

Since we may mix strategies (cf. Chapter 1, §3), there exists a strategy π such that

$$Pq(a_t) = \sum_{k=1}^{\infty} 2^{-k} P_k q(a_t) \qquad (t = m+1, m+2, \ldots), \tag{5}$$

where the measure P corresponds to the strategy π and the measures P_k to the strategies π_k; all the processes start at the state x.

The x-summability of the model, property 2.S and formulas (2.1), (4) and (5), imply that

$$w(x,\pi) = \sum_{k=1}^{\infty} \frac{1}{2^k} w(x,\pi_k) \geq \sum_{k=1}^{\infty} \frac{1}{2^k} \cdot 2^k = +\infty = v(x).$$

* We shall say that the function f is μ-summable above (below) if $\mu f^+ < \infty$ ($\mu f^- < \infty$). For such a function $\mu f = \mu f^+ - \mu f^-$. The words "above" ("below") in the formulation (b) is dropped in accordance with the remarks of §2.

Turning now to the proof of b), we denote by Q and R the subsets of X_m on which $v > 0$ and $v < 0$. Since

$$w(x,\pi) \le v(x),$$
$$-w^-(x,\pi) \le 0 \le w^+(x,\pi), \qquad (x \in X_m),$$
$$-w^-(x,\pi) \le w(x,\pi) \le w^+(x,\pi),$$

we find from formula (3.1) that for any strategy π

$$\mu v^- = -\sum_R \mu(x)v(x) \le -\sum_R \mu(x)w(x,\pi)$$
$$\le \sum_R \mu(x)w^-(x,\pi) \le \sum_{X_m} \mu(x)w^-(x,\pi) = w^-(\mu,\pi),$$

and for a strategy σ of Result a)

$$\mu v^+ = \sum_Q \mu(x)v(x) \le \sum_Q \mu(x)[w(x,\sigma) + \varepsilon]$$
$$\le \sum_Q \mu(x)[w^+(x,\sigma) + \varepsilon] \le \sum_{X_m} \mu(x)[w^+(x,\sigma) + \varepsilon]$$
$$= w^+(\mu,\sigma) + \varepsilon$$

(all the sums have a meaning since the terms are of the same sign).

From these inequalities it is clear that if the model is μ-summable, then the function $v(x)$ is μ-summable as well, so that $\mu v = \mu v^+ - \mu v^-$ has a meaning. The second half of assertion b), and assertion c), are proved just as in Chapter 1, §12. It follows from what has been proved that the two following conditions are equivalent:

1°. *The model Z is μ-summable above.*
2°. $v(\mu) < +\infty$.

Indeed, if $v(\mu) < +\infty$, then $v(\mu)$ has a meaning, which means that the model is μ-summable either above or below. In addition, for any strategy π

$$w(\mu,\pi) = w^+(\mu,\pi) - w^-(\mu,\pi) \le v(\mu) < +\infty$$

so that $w^+(\mu,\pi) < +\infty$. This means that 2° implies 1°. Conversely, if 1° is satisfied, then $v(\mu)$ has a meaning; if $v(x)$ were equal to $+\infty$, then, by a), there would be a strategy π for which $w(\mu,\pi) = w^+(\mu,\pi) - w^-(\mu,\pi) = +\infty$, so that $w^+(\mu,\pi) = +\infty$, which contradicts 1°. Thus 1° implies 2°. We shall show that conditions 1° and 2° are equivalent also to the following:

3°. $\sup_\pi w^+(\mu,\pi) < +\infty$.

To this end we consider the model Z^+ gotten from Z by replacing the reward function q by q^+. Obviously the model Z is μ-summable above if and only if the model Z^+ has the same property. Applying the equivalence of conditions 1° and

2° proved above to the model Z^+, we find that the model Z is summable above if and only if the value of the initial distribution μ in the model Z^+ is finite. But this value is equal to

$$\sup_{\pi} w^+(\mu,\pi),$$

and this means that 1° is equivalent to 3°.

One establishes the equivalence of the following conditions similarly.

1a°. *The model Z is μ-summable below.*

2a°. $\inf_{\pi} w(\mu,\pi) > -\infty$.

3a°. $\sup_{\pi} w^-(\mu,\pi) < +\infty$.

§5. Optimality Equations

For a finite control interval (and for a reward function bounded above) the following results were obtained in §6 and §12 of Chapter 1.

a) *The value v of the model Z is expressed in terms of the value v' of the derived model Z' by the formulas*

$$v = Vu, \qquad u = Uv', \tag{1}$$

where

$$Vg(x) = \sup_{a \in A(x)} g(a) \qquad (x \in X), \tag{2}$$

$$Uf(a) = q(a) + \sum_{y \in X} p(y|a)f(y) \qquad (a \in A); \tag{3}$$

b) *For any $\kappa > 0$ there exists a selector ψ of the correspondence $A(x)(x \in X_m)$, such that*

$$u(\psi(x)) \geq v(x) - \kappa \tag{4}$$

for all $x \in X_m$.

c) *Suppose that ε' and κ are arbitrary nonnegative numbers. If the strategy π' is ε'-optimal for the model Z' and the selector ψ satisfies condition* (4), *then the strategy $\psi\pi'$ is $(\varepsilon' + \kappa)$-optimal for the model Z.*

In order to generalize these results, we have first of all to show that the functions v and v' have a meaning. The existence of v follows from the x-summability of the model Z for any $x \in X_m$, a condition introduced in the preceding section. Beginning at this point, and up to the end of the chapter, we will assume that *not only Z, but all the models Z', Z'', . . . derived from Z, are x-summable.* (If this is so we will say that *the model Z is summable*). One may always assure that this additional requirement is satisfied by excluding from X_t all the states in which the condition of x-summability of the corresponding derived model is violated. Such a purge of the state space does not affect the control of the model Z, since in view of 2a)–2b) the excluded states are inaccessible for any strategy. Under the hypotheses made

above, finite or infinite values v are defined for the model and all of the models derived from it.

Consider first the case when the model is summable above, so that $v < +\infty$. In this case the validity of the results a)–c) is established in the same way as in §12 of Chapter 1. The possibility of applying the operator U to the function v' and the equation

$$Uv'(a) = q(a) + v'(p_a)$$

follows from 3b) and 4b). The inequality $w'(p_a,\pi') \geq v'(p_a) - \varepsilon$, where π' is an ε'-optimal strategy for Z', follows from 4c).

Now suppose that the model is summable below. The example presented at the end of §13 of Chapter 1 shows that assertion b) can be false for points x at which $v(x) = +\infty$. However the following weakened variant of b) is true.

b') *For any $\kappa > 0$ and $K > 0$ there exists a selector ψ of the correspondence $A(x)(x \in X_m)$ such that*

$$u(\psi(x)) \geq \begin{cases} v(x) - \kappa & \text{for } v(x) < +\infty, \\ K & \text{for } v(x) = +\infty. \end{cases}$$

Indeed, if $v(x) < +\infty$ the previous arguments remain valid. If $v(x) = +\infty$, then, by virtue of the fundamental equation, formula (4.3), and the definition of $u(a)$ contained in (1) and (3),

$$+\infty = \sup_{\pi} w(x,\pi) \leq \sup_{\substack{\pi' \\ a \in A(x)}} [q(a) + w'(p_a,\pi')]$$

$$\leq \sup_{a \in A(x)} [q(a) + v'(p_a)] = \sup_{a \in A(x)} u(a).$$

a) follows from b'): if $v(x) < +\infty$ the previous proof remains valid, and if $v(x) = +\infty$ we have $Vu(x) = +\infty = v(x)$ in view of (2).

The result c) and its proof do not depend on whether the model is summable above or below.

In what follows it is convenient to rewrite equations (1) and condition (4) in terms of the operators T_ψ and T defined by the formulas

$$T_\psi f(x) = Uf(\psi(x)), \qquad (x \in X) \tag{5}$$

$$Tf(x) = \sup_{\psi} T_\psi f(x) = VUf(x) \qquad (x \in X), \tag{6}$$

see the end of §6 of Chapter 1. Moreover, as in §7 of Chapter 1, we suppose that $m = 0$ and denote the model Z and its successive derived models by $Z_0, Z_1, Z_2, \ldots$ and the corresponding model values by $v_0, v_1, v_2, \ldots$. It follows from Result a) that in a summable model the values v_t are connected by the recurrence relations

$$v_t = Tv_{t+1}, \qquad t = 0, 1, 2, \ldots. \tag{7}$$

Result b) implies that, in a model which is summable above, for any sequence of positive numbers $\kappa_1, \kappa_2, \ldots$, it is possible to choose selectors $\psi_1, \psi_2, \ldots$ of the correspondence $A(x)$ $(x \in X_{t-1})$, $t = 1, 2, \ldots$ such that

$$T_{\psi_t} v_t = v_{t-1} - \kappa_t. \tag{8}$$

Finally, it follows from Result c) that for such ψ_t and for any ε'-optimal strategy π in the model Z_n, the product $\psi_1\psi_2 \cdots \psi_n\pi$ is an ε-optimal strategy for Z with $\varepsilon = \kappa_1 + \kappa_2 + \cdots + \kappa_n + \varepsilon'$ $(n = 1,2, \ldots)$.

It follows from (7) that for any $n > 0$

$$v_0 = T^n v_n. \tag{9}$$

In the next section we shall deduce that under some additional restrictions

$$v_0 = \lim_{n\to\infty} T^n 0.$$

Simply put, this means that the control on a finite, but sufficiently long, interval $[0,n]$ may yield just about the same gain as the control on an infinite time interval.

Now suppose further that $\varphi = \psi_1\psi_2 \cdots \psi_t \cdots$, where the selectors ψ_t satisfy condition (8), and that $\varepsilon = \kappa_1 + \kappa_2 \cdots + \kappa_t + \cdots$. In §7 we will discuss conditions under which the simple strategy φ is uniformly ε-optimal.

§6. An Expression for the Value of a Model

Consider a summable model Z. Obviously, for any n, the model Z^n, gotten from Z by replacing the reward function q in all the spaces A_t with $t > n$ by 0, is summable. We will denote the values of v and w in the model Z^n by v^n and w^n respectively. It is clear that $v^n = 0$ on X_n, and therefore in view of (4.9) $v^n = T^n 0$ on X_0.

By virtue of (1.1),

$$w(x,\pi) = \lim_{n\to\infty} w^n(x,\pi) \qquad (x \in X_0) \tag{1}$$

so that

$$v(x) = \sup_{\pi} \lim_{n\to\infty} w^n(x,\pi).$$

If exchanging the sup and lim signs were legitimate, we could find from this that

$$v = \lim_{n\to\infty} v^n = \lim_{n\to\infty} T^n 0. \tag{2}$$

But this operation is possible only under quite restrictive conditions, and (2) can be false, as is demonstrated by the following example.

EXAMPLE 1. Consider a homogeneous model, in which all the X_t (and A_t) are identical. The space X_t consists of the point x and of the points y_k, $k = 1, 2, \ldots$ (see

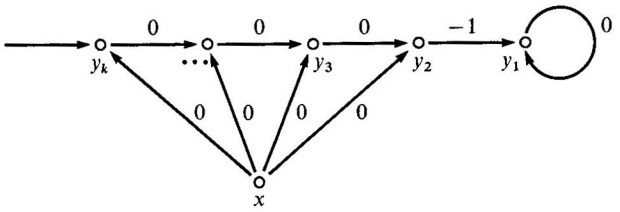

Figure 4.2

figure (4.2)). From y_{k+1} there is a deterministic transition into y_k, and the state y_1 is absorbing. By an appropriate choice of the action, we may pass from x into any of the states y_k with index $k \geq 2$. The reward function q is equal to 0 throughout, aside from the arrow leading from y_2 into y_1; here $q = -1$. Starting from x, we sooner or later pass from y_2 into y_1, so that $v(x) = -1$. However $v^n(x) = 0$ for any integer n, because one may pass from x to y_k with an index k so large that n steps do not suffice to attain y_1.

It follows from equation (1) only that

$$v \leq \varliminf_{n\to\infty} v^n.$$

Indeed, fix x, and then choose an arbitary number c less than $v(x)$. Then $w(x,\pi) > c$ for some strategy π, and, in view of (1), $w^n(x,\pi) > c$ beginning with some $n = n_0$. Thereafter, for $n \geq n_0$, $v^n(x) > c$. Therefore $\varliminf v^n(x) \geq c$.

For v to be equal to the limit v^n, we need to exclude the possibility of substantial losses in an arbitrarily far removed future. Example 1 shows that summability below is not sufficient for this.

In order for equation (2) to be satisfied it suffices that the quantity

$$w_n(x,\pi) = \sum_{t=n+1}^{\infty} P_x^{\pi} q(a_t) \tag{4}$$

should satisfy the relation

$$\varliminf_{n\to\infty} \inf_{\pi} w_n(x,\pi) \geq 0 \qquad (x \in X_0). \tag{5}$$

Indeed, put $z_n(x) = \inf_{\pi} w_n(x,\pi)$. For any strategy π

$$w^n(x,\pi) + z_n(x) \leq w^n(x,\pi) + w_n(x,\pi) = w(x,\pi) \leq v(x)$$

so that

$$v^n(x) + z_n(x) \leq v(x).$$

By (5), we have

$$0 \leq \underline{\lim}_{n\to\infty} z_n \leq \underline{\lim}_{n\to\infty} (v - v^n) = v - \overline{\lim}_{n\to\infty} v^n.$$

Jointly with inequality (3), this yields (2).

Inequality (5) is evidently satisfied if the reward function q is nonnegative. It is satisfied also if *there exist positive functions* $b_t(x)$, $x \in X_0$, *such that*

$$\sum_1^\infty b_t(x) < +\infty \tag{6}$$

and for sufficiently large t, the estimate

$$P_x^\pi q(a_t) \geq -b_t(x) \tag{7}$$

is valid *for any* $x \in X_0$ *and any strategy* π. Indeed, it follows from (4) and (7) that for any x and π and sufficiently large n

$$w_n(x,\pi) \geq -\sum_{n+1}^\infty b_t(x) \tag{8}$$

so that (5) follows from (6).

Summable models satisfying the condition in italics above will be said to be *bounded below*. Thus, for any model which is bounded below we have $T^n 0 \to v$.

Any summable model on a finite interval is obviously bounded below on a finite interval. On the other hand, if inequalities (7) are satisfied *for all* t, then the model is x-summable for any initial state x. "Purging" the space X according to the description in §5, we may suppose that the model is summable. Evidently it is bounded below.

§7. Simple ε-Optimal Strategies

In this section we study strategies of the form $\varphi = \psi_1\psi_2 \cdots \psi_t \cdots$, where the ψ_t are selectors of the correspondences $A(x)$, $(x \in X_{t-1})$ (simple strategies). Our aim is to show that if

$$T_{\psi_t} v_t \geq v_{t-1} - \kappa_t \tag{1}$$

and $\varepsilon = \kappa_1 + \kappa_2 + \cdots$, then

$$w(x,\varphi) \geq v(x) - \varepsilon, \tag{2}$$

i.e. the strategy φ is uniformly ε-optimal. We shall see that this assertion is valid only under certain additional restrictions on the model Z.

Suppose that the model Z is summable. In view of formulas (3.5) and (3.7), for any strategy π and any model Z^n

$$w^n(x,\varphi) + P_x^\varphi w_n(x_n,\pi) = T_{\psi_1} T_{\psi_2} \cdots T_{\psi_n} w_n(x,\pi). \tag{3}$$

It is clear from the definition of the operators T_ψ that they preserve inequalities between functions, and that for any constant c we have $T_\psi(c+f) = c + T_\psi f$; (a constant term may be carried across the operator sign). For each $\varepsilon > 0$ there exists a strategy π_ε for which

$$v_n(x) - \varepsilon \le w_n(x, \pi_\varepsilon) \le v(x) \qquad (x \in X_n).$$

Therefore, since (3) holds for all strategies π_ε it follows that

$$w^n(x,\varphi) + P_x^\varphi v_n(x_n) = T_{\psi_1} T_{\psi_2} \cdots T_{\psi_n} v_n(x). \tag{4}$$

From inequalities (1) and the properties of the operators T_ψ indicated above it follows that

$$\begin{aligned} T_{\psi_1} \cdots T_{\psi_{n-1}} T_{\psi_n} v_n &\ge T_{\psi_1} \cdots T_{\psi_{n-1}} v_{n-1} - \kappa_n \\ &\ge T_{\psi_1} \cdots T_{\psi_{n-2}} v_{n-2} - \kappa_{n-1} - \kappa_n \ge \cdots \\ &\ge v - \kappa_1 - \kappa_2 - \cdots - \kappa_n \ge v - \varepsilon. \end{aligned} \tag{5}$$

Since

$$w(x,\varphi) = \lim_{n\to\infty} w^n(x,\varphi) \tag{6}$$

(see (6.1)), then for inequalities (2) to follow from (4) and (5) it suffices that

$$\overline{\lim_{n\to\infty}}\, P_x^\varphi v_n(x_n) \le 0. \tag{7}$$

* * *

We shall dwell in detail on the case when $v(x)$ is finite. It follows from formula (4) that in this case there exists a limit

$$\delta(x) = \lim_{n\to\infty} P_x^\varphi v_n(x_n). \tag{8}$$

Indeed, of the three terms entering into formula (4), the first has, from (6), a limit $w(x,\varphi)$. The third is monotonically nonincreasing in view of the inequalities

$$T_{\psi_{n+1}} v_{n+1} \le T v_{n+1} = v_n$$

and therefore also has some limit $\lambda(x) \le v_0(x) = v(x)$. Therefore the limit (8) exists and is equal to

$$\delta(x) = \lambda(x) - w(x,\varphi), \tag{9}$$

if at least one of the terms on the right side is finite. The finiteness of λ follows from the inequalities

$$v - \varepsilon \leq \lambda \leq v, \tag{10}$$

the lower estimate being obtained by a passage to the limit from (5).

In view of (9) and (10)

$$v(x) - w(x,\varphi) - \varepsilon \leq \delta(x) \leq v(x) - w(x,\varphi).$$

These inequalities have some interesting consequences:

1) *Always* $\delta \geq -\varepsilon$ (since $w(x,\varphi) \leq v(x)$).

2) *If* $\delta \leq 0$, *then the strategy* φ *is* ε*-optimal* (This follows also from relations (7) and (8)).

3) *If the strategy* φ *is* ε*-optimal, then* $\delta \leq \varepsilon$.

Applying 2) and 3) to the case $\varepsilon = 0$, we arrive at the following result.

Suppose that the value v *is finite and that the selectors* ψ_t *satisfy the conditions*

$$T_{\psi_t} v_t = v_{t-1},$$

$t = 1, 2, \ldots$: *Put* $\varphi = \psi_1 \psi_2 \cdots$. *Then the nonnegative limit*

$$\delta(x) = \lim_{n \to \infty} P_x^{\varphi} v_n(x_n)$$

exists. For the optimality of the simple strategy φ *it is necessary and sufficient that the limit be zero.*

* * *

In analogy to the class of models bounded below (see §6), one may introduce a class of models bounded above. We shall say that a summable model is *bounded above*, if there exist positive functions $c_t(x)$, $x \in X_0$, such that

$$\sum_1^{\infty} c_t(x) < +\infty, \tag{11}$$

and, for all sufficiently large t, the estimate

$$P_x^{\pi} q(a_t) \leq c_t(x) \tag{12}$$

is valid for all $x \in X_0$ and all strategies π. We shall show that *for such models the condition* (7) *is satisfied, so that the strategy* φ *is* ε*-optimal.*

First we prove that in a model Z which is bounded above, for sufficiently large n

$$P_x^{\varphi} w_n(x_n,\pi) \leq \sum_{n+1}^{\infty} c_t(x) \tag{13}$$

for any $x \in X_0$, any simple strategy φ in the model Z and any strategy π in the derived model Z_n.

Suppose that $\rho = \psi_1\psi_2 \cdots \psi_n\pi$, where $\psi_1, \psi_2, \ldots \psi_n$ are the first n factors of the strategy φ. Evidently the quantity in the left side of (13) does not depend on the values of the reward function q on the spaces $A_1, \ldots, A_n$. Putting $q = 0$ on these spaces, we find from (3.7) that

$$P_x^{\varphi} w_n(x_n,\pi) = \sum_{t=1}^{n} P_x^{\rho} 0(a_t) + \sum_{t=n+1}^{\infty} P_x^{\rho} q(a_t) = \sum_{n+1}^{\infty} P_x^{\rho} q(a_t).$$

Therefore (13) follows from (12).

Applying (13) to the uniformly ε-optimal strategy π_ε in the model Z_n, we find that

$$P_x^{\varphi} v_n(x_n) \leq P_x^{\varphi} w_n(x_n,\pi_\varepsilon) + \varepsilon \leq \varepsilon + \sum_{n+1}^{\infty} c_t(x);$$

Because the positive number ε is arbitrary, it therefore follows that

$$P_x^{\varphi} v_n(x_n) \leq \sum_{n+1}^{\infty} c_t(x). \tag{14}$$

Equation (7) now obviously follows from (14) and (11).

We note that all summable models on a finite time interval $[m,n]$ are summable both above and below.

For models bounded above, given a fixed initial state one may neglect positive contributions which are introduced into the mean reward at times which are far removed. The following example shows that this condition is at the heart of the matter.

EXAMPLE 1. Consider a homogeneous model with two states, as shown in figure 4.3. In the state x two actions are possible, carrying us into x and y respectively. The state y is absorbing. Obviously $v(x) = 1$ and $v(y) = 0$. The strategy φ, consisting of returning permanently to x, satisfies inequalities (1) for $\kappa_t = 0$, but is not optimal, since $w(x,\varphi) = 0$.

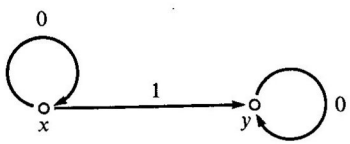

Figure 4.3

We say that the model is *uniformly bounded below* if it is bounded below and all the functions $b_t(x)$ $(t = 1,2,\ldots)$ in (6.6) and (6.7) are constants. Models uniformly bounded above are defined analogously with b_t replaced by c_t.

If the model is uniformly bounded both above and below, then in order to obtain a strategy close to an optimal one, a finite number of the conditions (1) suffice. More precisely, we shall show that *if*

$$T_{\psi_t} v_t \geq v_{t-1} - \kappa_t \qquad (t = 1,2,\ldots,n),$$

then any strategy π *coinciding on the first* n *steps with the product* $\varphi = \psi_1\psi_2\cdots\psi_n$ *is* ε*-optimal for*

$$\varepsilon = \sum_1^n \kappa_t + \sum_{t>n} (b_t + c_t) \tag{15}$$

(Under the assumptions that inequalities (6.7) and (12) are satisfied starting from n.) By taking n sufficiently large and then taking sufficiently small $\kappa_1, \kappa_2, \ldots, \kappa_n$, we may make ε arbitrarily small.

For the proof we observe that $w^n(x,\pi) = w^n(x,\varphi)$, so that, according to formula (4),

$$\begin{aligned} w(x,\pi) &= w^n(x,\pi) + w_n(x,\pi) \\ &= T_{\psi_1} T_{\psi_2} \cdots T_{\psi_n} v_n(x) - P_x^{\varphi} v_n(x_n) + w_n(x,\pi). \end{aligned}$$

It follows from (5), (14) and (6.8) that the right side is not less than $v(x) - \varepsilon$.

§8. Sufficiency of Markov and Simple Strategies

Recall the main results of §13, Chapter 1.

1. For each initial distribution μ and strategy π there exists a Markov strategy σ such that
$$w(\mu,\sigma) = w(\mu,\pi).$$

2. For any Markov strategy σ there exists a simple strategy φ such that
$$w(\mu,\varphi) \geq w(\mu,\sigma) \quad \text{for all } \mu.$$

To what extent do these results carry over to the present case?

Result 1 remains valid for any μ-summable model. Indeed, we define σ for all $t > 0$ by formula (1.13.3). We have shown that, for any t the probability distributions for a_t relative to the measures P_μ^σ and P_μ^π coincide. Therefore

$$w(\mu,\pi) = \sum_1^\infty P_\mu^\pi q(a_t) = \sum_1^\infty P_\mu^\sigma q(a_t) = w(\mu,\sigma).$$

The situation is different with Result 2. If $v = +\infty$, then this result is false; see example 2 of Chapter 1, §13. It is not known whether this result is valid for any

model which is summable above. It is possible to prove it only for models which have a nonnegative reward function q, or, somewhat more generally, for models Z which are summable above and bounded above (see §7).

For any model which is summable above, we may, as in §13 of Chapter 1, choose for each $t = 1, 2, \ldots$ a selector ψ_t of the correspondence $A(x)$, $x \in X_{t-1}$, such that

$$w_{t-1}(x,\psi_t\sigma^t) \geq w_{t-1}(x,\sigma^{t-1}) \tag{1}$$

for all $x \in X_{t-1}$; here we denote by σ^t the restriction of σ to Z^t. From these inequalities we wish to deduce that

$$w(x,\varphi) \geq w(x,\sigma) \qquad (x \in X_0) \tag{2}$$

for the simple strategy $\varphi = \psi_1\psi_2 \cdots$.

Using the operators T_ψ, we may write out the inequalities (1) in the form

$$T_{\psi_t} w_t(x,\sigma^t) \geq w_{t-1}(x,\sigma^{t-1}) \qquad (t = 1,2,\ldots).$$

Therefore, using formulas (3.5) and (3.7), we get

$$\begin{aligned} w(x,\sigma) &\leq T_{\psi_1} w_1(x,\sigma^1) \leq T_{\psi_1} T_{\psi_2} w_2(x,\sigma^2) \\ &\cdots \leq T_{\psi_1} T_{\psi_2} \cdots T_{\psi_n} w_n(x,\sigma^n) = w(x,\psi_1\psi_2\cdots\psi_n\sigma^n) \\ &= w^n(x,\varphi) + P_x^\varphi w_n(x_n,\sigma^n) \end{aligned} \tag{3}$$

for any $n > 0$.

Since $w^n(x,\varphi) \to w(x,\varphi)$, then (2) follows from (3) under the following additional assumption

$$\varlimsup_{n\to\infty} P_x^\varphi w_n(x_n,\sigma^n) \leq 0.$$

This condition is slightly weaker than (7.7), and hence it is satisfied if the model is bounded above.

A close but weaker result may be obtained for models uniformly bounded below (see §7). Indeed, in such models, for any initial distribution μ (for which the model is μ-summable above) any Markov strategy σ and any $\varepsilon > 0$, it is possible to construct a simple strategy φ such that

$$w(\mu,\varphi) > w(\mu,\sigma) - \varepsilon. \tag{4}$$

First of all in view of the results of §3, we may without loss of generality suppose that the model Z is simply summable above. Indeed we are interested only in the process Z_μ, and the model Z is μ-summable above. (cf. the analogous remark in §5).

Since

$$w^n(\mu,\sigma) \to w(\mu,\sigma) < +\infty,$$

then for sufficiently large n

$$w^n(\mu,\sigma) \geq w(\mu,\sigma) - \frac{\varepsilon}{2}. \tag{5}$$

Since the model is uniformly bounded below, for sufficiently large n

$$w(\mu,\pi) = w^n(\mu,\pi) + \sum_{n+1}^{\infty} P_\mu^\pi q(a_t) \geq w^n(\mu,\pi) - \frac{\varepsilon}{2} \tag{6}$$

for any strategy π. Starting out with a number n for which both (5) and (6) are valid, we replace the reward function q on the sets A_t with $t > n$ by zero. Then the value w^n of the previous model turns into the value w of the new model. Since the new model, being summable above, is also bounded above, then there exists in it a simple strategy φ which is uniformly no worse than σ. For such a strategy

$$w^n(\mu,\varphi) \geq w^n(\mu,\sigma), \tag{7}$$

and (4) now follows from (5), (6), and (7).

Thus we have the following results:

a) *In a μ-summable model, for each strategy π there exists a Markov strategy σ such that $w(\mu,\sigma) = w(\mu,\pi)$.*

b) *In a model which is summable above and bounded above, for any Markov strategy σ there exists a simple strategy φ such that*

$$w(x,\varphi) \geq w(x,\sigma)$$

for all $x \in X_0$.

c) *In a model which is uniformly bounded below, for any initial distribution μ for which the model is μ-summable above, any Markov strategy σ and any $\varepsilon > 0$, there exists a simple strategy φ such that $w(\mu,\varphi) \geq w(\mu,\sigma) - \varepsilon$.*

Taking account of the remark with which we began the proof of Result c), and also of formula (3.1), we obtain the following variant of Result b), analogous to Result c):

b′) *In a model which is bounded above, for any initial distribution μ for which the model is μ-summable above and any Markov strategy σ, there exists a simple strategy φ such that $w(\mu,\varphi) \geq w(\mu,\sigma)$.*

The following is a consequence of a), b′), and c).

d) *Suppose that the model is μ-summable above and that π is any strategy. If the model is bounded above, then there exists a simple strategy φ such that $w(\mu,\varphi) \geq w(\mu,\pi)$. If the model is uniformly bounded below, then for any $\varepsilon > 0$ there exists a simple strategy φ such that $w(\mu,\varphi) \geq w(\mu,\pi) - \varepsilon$.*

The question as to the possibility of extending either of the Results b), c), and along with them d), to arbitrary models which are bounded above remains open.