Journal of Engineering Mathematics, Vol. 10, No. 2, April 1976
Noordhoff International Publishing–Leyden
Printed in The Netherlands

On the Bellman principle for decision problems with random decision policies

R. P. VAN DER VET

Koninklijke/Shell-Laboratorium, (Shell Research B.V.), Amsterdam, The Netherlands.

(Received November 4, 1974)

SUMMARY
In this paper we give two new results in the field of discrete time-dynamic decision problems. Firstly we prove the validity of the Bellman principle for the class of random decision policies; and secondly we give the effect on the objective function resulting from the decision maker being able to make use of an advisor with more information.

1. Introduction

We consider a decision process with decision steps at N discrete time instants. There are two ways to attack such a problem:

(a) Determine the optimal sequence of decisions simultaneously. This is the case, for instance, in multiperiod linear programming, where the whole sequence is generated simultaneously by solving one linear program; and this is what we mean, in fact, by the optimal sequence of decisions maximizing or minimizing a multiperiod objective.

(b) Use a decomposition in time. This can be performed by the application of the Bellman principle [1], which reduces the optimization procedure by determining an optimal solution at each time instant separately.

The value of the latter approach will increase when it can be shown that it leads to the same optimal solution as obtained by the simultaneous optimization process. Now, the two methods obviously give the same solution for deterministic problems, but to show this rigorously for decision problems with random parameters is less trivial. Agreement has been shown for deterministic decision policies [2–6].

The main purpose of this paper is to prove this agreement for a more extended class of random decision policies, *i.e.* for policies which are mappings into the set of distribution functions over the admissible decisions. No attention has been paid to this up to now. The result is given in section 5. In section 6 we give another new result, which takes account of the information aspect. Here we give the effect on the objective function resulting from the decision maker being able to make use of an adviser who has more information at his disposal.

For elucidation purposes we will represent the decision problem as an N-step decision tree in which each path is a chain of such distributions. This representation makes possible an explicit formulation of the Bellman principle for the problem under consideration. For this we use a presentation due to Clément [7]. In fact it will be shown that the validity of this principle is based on a simple and fundamental isotony condition for the concatenation of sub-paths in the tree.

2. Problem statement

Consider a discrete dynamic system whose state is a random variable with a given distribution. At an arbitrary instant in time k, we have the following situation:

(a) The system is in a certain state, x_k;

(b) The decision maker has at his disposal an amount of information, z_k, and a set of admissible decision actions, u_k.

The information z_k consists of two elements: z_{k-1}, the information from the preceding time instant; and s_k, the additional information obtained at k. We assume this additional information to consist of:

1. u_{k-1}, the last decision action taken;
2. y_k, other relevant, mostly incomplete, information about the system which becomes available at k (this may be, for instance, an incomplete observation of the state x_k).

Thus, concerning the information aspect of the problem, we define

$$z_k = \{z_{k-1}, s_k\}, \text{ the observable history,}$$

where

$$s_k = \{u_{k-1}, y_k\} \text{ is the additional information obtained at instant } k.$$

The decision consists in choosing a mapping from the set of available information, z_k, into the set of all probability distributions of u_k. In fact, a decision at time instant k is an element out of the class.

$$S = \{\Phi_k \mid \Phi_k : z_k \to \Phi_k(\cdot \mid z_k)\}$$

where

$$\Phi_k(\cdot \mid z_k) : u_k \to \Phi_k(u_k \mid z_k)$$

is a density function of u_k given z_k.

We will refer to the elements of S as the random decision policies or, more simply, as the policies.

Note that a random decision policy can be interpreted as follows: if the decision maker could repeatedly arrive at time instant k, every time with the same information z_k, then the random decision policy is the distribution over the decision actions he would take. It should be noted that the class D of deterministic policies

$$D = \{C_k \mid C_k : z_k \to u_k = C_k(z_k)\}$$

is contained in S, since the δ-function

$$\Phi_k(u_k \mid z_k) = \delta[u_k - C_k(z_k)]$$

is an element of S.

At each time instant k we define the valuation $V_k[x_{k+1}, u_k]$, which represents a measure of the behaviour of the system. The overall objective is defined as the expectation EJ_0, where

$$J_0 = \sum_{i=0}^{N-1} V_i[x_{i+1}, u_i].$$

It is the task of the decision maker to choose the overall strategy $\Phi_0, \ldots, \Phi_k, \ldots, \Phi_{N-1}$ (where $\Phi_k \in S$ and $k = 0, 1, \ldots, N-1$) which makes the objective EJ_0 maximal. This problem will be treated in the next sections.

3. Derivation of auxiliary formulae

In this section we derive some formulae and make some observations which are relevant for the developments in the following sections.

Let z_k be fixed. This includes that

(a) The decision actions $u_0, \ldots, u_{k-1}$ are fixed, and consequently the policies $\Phi_0, \ldots, \Phi_{k-1}$ are δ-distributions;

(b) For every $\Phi_k \in S$ we have the density function $\Phi_k(\cdot \mid z_k)$.

Furthermore we derive formulae for the following density functions:

1. The density function of the additional information s_{k+1}:

$$p(s_{k+1} \mid z_k) = p(y_{k+1}, u_k \mid z_k) = p(y_{k+1} \mid u_k, z_k)\Phi_k(u_k \mid z_k); \tag{1}$$

2. The joint density function of the pair (x_{k+1}, u_k):

$$p(x_{k+1}, u_k \mid z_k) = p(x_{k+1} \mid u_k, z_k)\Phi_k(u_k \mid z_k); \tag{2}$$

3. More generally, the joint density function of the pair (x_{j+1}, u_j); $j > k$:

$$\begin{aligned} p(x_{j+1}, u_j \mid z_k) &= \int p(x_{j+1}, u_j, s_j, \ldots, s_{k+1} \mid z_k) ds_j, \ldots, ds_{k+1} \\ &= \int p(x_{j+1}, u_j \mid z_j) p(s_j \mid z_{j-1}), \ldots, p(s_{k+1} \mid z_k) ds_j, \ldots, ds_{k+1} \end{aligned} \tag{3}$$

where $p(x_{j+1}, u_j \mid z_j)$ depends on Φ_j, and $p(s_{i+1} \mid z_i)$; $i = k, \ldots, j-1$ depends on Φ_i.

Thus, the expectations $E[V_j(x_{j+1}, u_j) \mid z_k]$; $j \geqq k$ are determined as functions of the policies $\Phi_k, \ldots, \Phi_j$.

We can now make the following observations.

The tail of the objective

For this tail we can write

$$EJ_k = E\{E(J_k \mid z_k)\}, \text{ where } J_k = \sum_{i=k}^{N-1} V_i(x_{i+1}, u_i).$$

The outer expectation applies to the variable z_k. Consider the form $E(J_k \mid z_k)$ at a fixed value of z_k. This form is fully determined by the sequence of policies $\psi_k = \{\Phi_k, \Phi_{k+1}, \ldots, \Phi_{N-1}\}$ because of (3). In order to express this dependence we use the notation

$$E_{\psi_k}(J_k \mid z_k) \tag{4}$$

where E_{ψ_k} stands for the expectation under the sequence ψ_k.

The construction of a policy Φ_k

Consider the sequence of policies from the time instant $k+1$ onwards to be fixed, say $\hat{\psi}_{k+1} = \{\hat{\Phi}_{k+1}, \ldots, \hat{\Phi}_{N-1}\}$, and that a choice has to be made at k. From this time instant on we have the valuation $E_{\{\Phi_k, \hat{\psi}_{k+1}\}}(J_k | z_k)$, which, for fixed z_k, is only dependent on the density function $\Phi_k(\cdot | z_k)$. Consequently, we can choose an optimal density function $\Phi_k^*(\cdot | z_k)$ which maximizes this valuation. But this can be performed for every arbitrary, fixed value of z_k. By doing so we construct a policy

$$\Phi_k^* : z_k \to \Phi_k^*(\cdot | z_k)$$

which makes $E_{\{\Phi_k, \hat{\psi}_{k+1}\}}(J_k | z_k)$ maximal for all z_k.

We finally present a formula of a more general nature which will play a fundamental role in the following sections. For a given random variable $w = \phi(s, t)$ we have the conditional expectation

$$\begin{aligned} E(w | r) &= \iint \phi(s, t) p(s, t | r) ds\, dt \\ &= \iint \phi(s, t) p(s | t, r) p(t | r) ds\, dt \\ &= \int \left\{ \int \phi(s, t) p(s | t, r) ds \right\} p(t | r) dt. \end{aligned}$$

As a consequence we find

$$E(w | r) = \int E(w | t, r) p(t | r) dt. \tag{5}$$

4. Tree representation of the decision problem

As a representation of the decision problem we take an N-step decision tree, with decision steps at the time instants $0, 1, \ldots, N-1$ (Fig. 1). Each path in the tree is a chain of policies

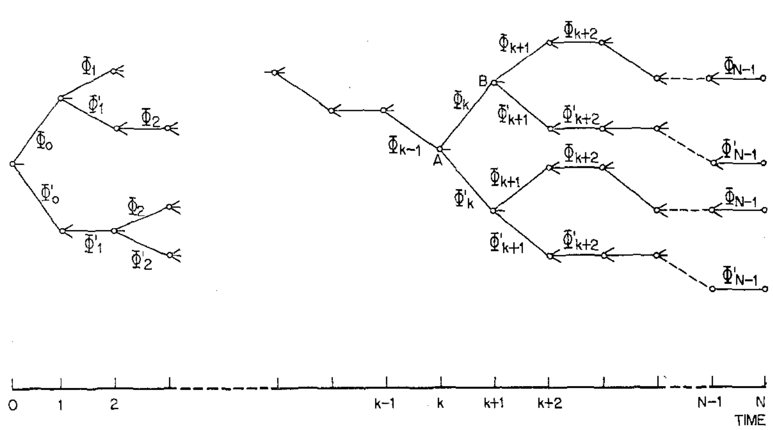

Figure 1.

Φ_k. The set of feasible policies in each node may be infinite. At an arbitrary node, at time instant k, we have the following situation:

(a) The system is in a certain state, x_k;

(b) The decision maker gets the information, z_k, at his disposal and makes a choice out of the set of feasible policies, Φ_k;

(c) This will eventually lead to a new node in the tree.

The search for an optimal path in the tree is considerably simplified if one can apply the Bellman principle to the selection procedure. The justification for this principle is contained in the following isotony condition. Let (see Fig. 1):

1. ψ_{k+1} be the path of policies $\Phi_{k+1}, \ldots, \Phi_{N-1}$ with the corresponding valuation $E_{\psi_{k+1}}(J_{k+1} | z_{k+1})$;
2. ψ'_{k+1} be an alternative path $\Phi'_{k+1}, \ldots, \Phi'_{N-1}$ with valuation $E_{\psi'_{k+1}}(J_{k+1} | z_{k+1})$;
3. Φ_k be the path from node A to node B.

Then, we can make the concatenations $\{\Phi_k, \psi_{k+1}\}$ and $\{\Phi_k, \psi'_{k+1}\}$. Furthermore, let $\psi_{k+1} \succ \psi'_{k+1}$ imply $E_{\psi_{k+1}}(J_{k+1} | z_{k+1}) \geqq E_{\psi'_{k+1}}(J_{k+1} | z_{k+1})$, for all z_{k+1}. Then, the isotony condition reads:

$$\psi_{k+1} \succ \psi'_{k+1} \Rightarrow \{\Phi_k, \psi_{k+1}\} \succ \{\Phi_k, \psi'_{k+1}\}.$$

This general setting of the Bellman principle has been presented by M. F. Clément [7].

If the system under consideration fulfils this condition, the search for an optimal path can be simplified by a backwards reduction of the set of paths. In the situation represented in Fig. 1 we can, when we have to make an optimal choice at A, delete all paths which are not better than ψ_{k+1}.

5. Verification of the isotony condition

Consider the two sub-paths $\psi_{k+1} = \{\Phi_{k+1}, \ldots, \Phi_{N-1}\}$ and $\psi'_{k+1} = \{\Phi'_{k+1}, \ldots, \Phi'_{N-1}\}$, with corresponding valuations $E_{\psi_{k+1}}(J_{k+1} | z_{k+1})$ and $E_{\psi'_{k+1}}(J_{k+1} | z_{k+1})$. Let us assume that:

$$E_{\psi_{k+1}}(J_{k+1} | z_{k+1}) \geqq E_{\psi'_{k+1}}(J_{k+1} | z_{k+1}), \text{ for all } z_{k+1}. \tag{6}$$

We now take one time step backwards and consider the concatenated subpaths $\{\Phi_k, \psi_{k+1}\}$ and $\{\Phi_k, \psi'_{k+1}\}$. Then, the isotony condition is fulfilled if

$$E_{\{\Phi_k, \psi_{k+1}\}}(J_k | z_k) \geqq E_{\{\Phi_k, \psi'_{k+1}\}}(J_k | z_k), \text{ for all } z_k.$$

In order to prove this we write:

$$E_{\{\Phi_k, \psi_{k+1}\}}(J_k | z_k) = E_{\{\Phi_k, \psi_{k+1}\}}(V_k | z_k) + E_{\{\Phi_k, \psi_{k+1}\}}(J_{k+1} | z_k).$$

Keeping z_k now fixed, we can make the following observations:

(a) The term $E_{\{\Phi_k, \psi_{k+1}\}}(V_k | z_k)$ is only dependent on the policy Φ_k, since the expectation of V_k does not depend on future policies. This term may therefore be denoted by $E_{\Phi_k}(V_k | z_k)$.

(b) For the second term we apply equation (5):

$$E_{\{\Phi_k, \psi_{k+1}\}}(J_{k+1} | z_k) = \int E_{\{\Phi_k, \psi_{k+1}\}}(J_{k+1} | z_{k+1}) p(s_{k+1} | z_k) ds_{k+1}$$

where $z_{k+1} \stackrel{\text{def}}{=} (z_k, s_{k+1})$. The first factor under the integral sign is only dependent on ψ_{k+1} and becomes $E_{\psi_{k+1}}(J_{k+1} | z_{k+1})$.

Thus, we obtain:

$$E_{\{\Phi_k, \psi_{k+1}\}}(J_k | z_k) = E_{\Phi_k}(V_k | z_k) + \int E_{\psi_{k+1}}(J_{k+1} | z_{k+1}) p(s_{k+1} | z_k) ds_{k+1}. \qquad (7)$$

A similar equation holds for

$$E_{\{\Phi_k, \psi'_{k+1}\}}(J_k | z_k). \qquad (8)$$

Note that, because of equation (1), the term $p(s_{k+1} | z_k)$ is only dependent on the policy Φ_k, and so this term has the same value in equations (7) and (8). Further, from equation (6) $E_{\psi_{k+1}}(J_{k+1} | z_{k+1}) \geqq E_{\psi'_{k+1}}(J_{k+1} | z_{k+1})$, and we can therefore conclude from (7) and (8) that

$$E_{\{\Phi_k, \psi_{k+1}\}}(J_k | z_k) \geqq E_{\{\Phi_k, \psi'_{k+1}\}}(J_k | z_k).$$

Clearly this result holds for every arbitrary, fixed value of z_k, and thus the isotony condition is fulfilled.

6. The effect of the information on the decision making

The information vector z_k, as defined in section 2, is of a very general form. The only assumption made was that it contains the previous decision actions. The information y_k may contain either complete knowledge about the behaviour of the system, or none at all. In order to investigate the effect of this information, we can compare two situations which differ in that the information in one situation includes the information in the other. We feel that the description becomes more intuitive if we consider the situations from the point of view of two persons: (a) the *decision maker*, and (b) the *advisor*. The decision maker has the information z_k at his disposal. He chooses policies from the class $S = \{\Phi_k\}$ which make use of this information. The advisor has more information at his disposal, say $\{z_k, \eta_k\}$, where η_k might reflect the fact that the advisor has more information about the behaviour of the system. The advisor chooses policies from the class $T = \{\Theta_k\}$ which make use of the information $\{z_k, \eta_k\}$. The elements Θ_k are defined in the same way as Φ_k (see section 2).

We note that the advisor can also choose a policy Φ_k by neglecting the additional information η_k. Thus, summing up, we have:
(a) S, the class of the decision maker's coarse strategies, Φ_k, using the information z_k; and
(b) T, the class of the advisor's fine strategies, Θ_k, using the information $\{z_k, \eta_k\}$.
Here, $S \subset T$.

We will now explain and give an interpretation of the quantities $E_{\psi_k}(J_k | z_k, \eta_k)$ and $E_{\pi_k}(J_k | z_k)$, where $\psi_k = \{\Phi_k, \ldots, \Phi_{N-1}\}$ and $\pi_k = \{\Theta_k, \ldots, \Theta_{N-1}\}$.

$E_{\psi_k}(J_k | z_k, \eta_k)$ is the advisor's expectation under fixed $\{z_k, \eta_k\}$ if he applies the coarse strategy ψ_k, *i.e.* without making use of the additional information η_k.

$E_{\pi_k}(J_k | z_k)$ is the decision maker's expectation under fixed z_k if he follows the advisor's fine strategy π_k.

Let π_k^* be the optimal strategy of the advisor; *i.e.* π_k^* makes $E_{\pi_k}(J_k | z_k, \eta_k)$ maximal for all $\{z_k, \eta_k\}$. Since $S \subset T$ we evidently have, for all strategies ψ_k and all $\{z_k, \eta_k\}$:

$$E_{\pi_k^*}(J_k | z_k, \eta_k) \geqq E_{\psi_k}(J_k | z_k, \eta_k).$$

From this inequality we pass to the inequality

$$E_{\pi_k^*}(J_k | z_k) \geqq E_{\psi_k}(J_k | z_k).$$

This is done by applying equation (5):

$$E_{\pi_k^*}(J_k \mid z_k) = \int E_{\pi_k^*}(J_k \mid z_k, \eta_k) p(\eta_k \mid z_k) d\eta_k$$
$$\geqq \int E_{\psi_k}(J_k \mid z_k, \eta_k) p(\eta_k \mid z_k) d\eta_k$$
$$= E_{\psi_k}(J_k \mid z_k).$$

Since this result holds for all strategies ψ_k, it also holds for the optimal strategy ψ_k^* of the decision maker. So we have proved that

$$E_{\pi_k^*}(J_k \mid z_k) \geqq E_{\psi_k^*}(J_k \mid z_k).$$

This result may be interpreted as follows. The decision maker chooses a strategy, ψ_k^*, which makes his expectation maximal. However, he can obtain better results if he follows the strategy of the advisor. We note that this result does not contradict the definition of ψ_k^* since the advisor can choose from a more extended class of policies.

7. The determination of the optimal path

For the sake of completeness, we will show that for the problem under consideration the optimal path in the tree is deterministic. (This result has already been demonstrated in the literature, notably by Fel'dbaum [8] and Aoki [9]).

Let the sub-path $\hat{\psi}_{k+1} = \{\hat{\Phi}_{k+1}, \ldots, \hat{\Phi}_{N-1}\}$ be fixed. We must find a policy Φ_k^* for which the form

$$E_{\{\Phi_k, \hat{\psi}_{k+1}\}}(J_k \mid z_k) \tag{9}$$

is maximal for all z_k simultaneously. The way in which such a policy can be constructed has already been discussed in section 3. In order to find this policy we apply equation (5) to (9) and find:

$$E_{\{\Phi_k, \hat{\psi}_{k+1}\}}(J_k \mid z_k) = \int E_{\{\Phi_k, \hat{\psi}_{k+1}\}}(J_k \mid z_k, u_k) \Phi_k(u_k \mid z_k) du_k.$$

Note that through the conditioning with respect to u_k, the first factor under the integral sign becomes independent of Φ_k. Thus, we obtain:

$$E_{\{\Phi_k, \hat{\psi}_{k+1}\}}(J_k \mid z_k) = \int E_{\hat{\psi}_{k+1}}(J_k \mid z_k, u_k) \Phi_k(u_k \mid z_k) du_k.$$

Let, for an arbitrary value of z_k, the form $E_{\hat{\psi}_{k+1}}(J_k \mid z_k, u_k)$ be maximal for $u_k^* = C_k^*(z_k)$. Then we clearly must choose $\Phi_k^*(u_k \mid z_k) = \delta[u_k - C_k^*(z_k)]$ in order to make $E_{\{\Phi_k, \hat{\psi}_{k+1}\}}(J_k \mid z_k)$ maximal. If we take such actions in reverse order, successively resulting in the policies $\Phi_{N-1}^*, \ldots, \Phi_k^*, \ldots, \Phi_0^*$, we will construct the optimal path, which is a deterministic one.

Acknowledgment

The author is indebted to Dr. H. Bolder of Koninklijke/Shell-Laboratorium, Amsterdam for valuable discussions.

REFERENCES

[1] R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton (1957).
[2] K. J. Aström, Optimal control of Markov Processes with incomplete state information, *Journal of Mathematical Analysis and Applications*, 10 (1965) 174.
[3] E. B. Dynkin, Controlled random sequences, *Theory of probability and its applications*, Vol. X, No. 1 (1965) 1.
[4] H. Kushner, *Introduction to stochastic control*, Holt, Reinhart and Winston Inc., New York (1971).
[5] L. Meier, *Combined optimum control and estimation theory*, NASA report No. CR-426 (1966).
[6] C. Striebel, Sufficient statistics in the optimum control of stochastic systems, *Journal of Mathematical Analysis and Applications*, 12 (1965) 576.
[7] M. F. Clément, Categorical axiomatics of dynamic programming, *J. Math. Anal. Appl.*, 51 (1975) 47.
[8] A. A. Fel'dbaum, Dual Control Theory I–IV, *Automation and Remote Control*, Vol. 21 (1960) 1240, 1453; and Vol. 22 (1961) 3, 129.
[9] M. Aoki, *Optimization of Stochastic Systems*, Academic Press, New York (1967).