The Annals of Statistics
1974, Vol. 2, No. 6, 1327-1334

INCOMPLETE INFORMATION IN MARKOVIAN DECISION MODELS

BY DETLEF RHENIUS

Universität Hamburg

If a set of states is given in a problem of dynamic programming in which each state can be observed only partially, the given model is generally transformed into a new model with completely observed states. In this article a method is introduced with which Markov models of dynamic programming can be transformed and which preserves the Markov property. The method applies to relatively general sets of states.

0. Introduction. Stochastic decision models in which the state space can be observed only partially (in the sense of Hinderer [5]: models with incomplete information) have been examined by several authors. Dynkin [3] gives examples of such decision models, among which the well-known "two-armed-bandit" problem is also listed (Feldman [4]). Further examples can be taken from learning theory (Karush and Dear [8]) in which in each of the consecutive trials a learning state $z' \in Z'$ is present which cannot be observed. On the other hand, an event $z \in Z$ can be observed which is made up of the behaviour of the learning subject and the reinforcement which it contains (see Iosifescu and Theodorescu [7]). The experimenter influences learning by the selection of a learning-method (action) $a \in A$ in each trial and this selection depends on the previous course of learning; the action is to be selected in such a way that a reward function is maximized.

Many decision models which are used in practice (e.g. in learning theory) are Markovian. In these it is desirable to select an optimal action which depends only on the present state. This is frequently possible in models with complete information, but not in models with incomplete information since the present state can be observed only partially. Dynkin [3] and Sawaragi, Yoshikawa [11] solve the problem by using a set of probability measures instead of that part of the state space which cannot be observed.

Thus Dynkin gets a model with observed history and he only works with this model. He does not mention that there are other strategies in the starting model than in the modified one. Hence one has to show that both models are equivalent with regard to optimality. This will be done in Section 4 of this article. In addition the reward functions in Dynkin's paper are more special and his results concerning Markov plans are confined to ε-optimality.

With respect to the points just listed the approach of Sawaragi, Yoshikawa is similar to that of the present article. However, the authors limit their

Received April 1973; revised December 1973.
AMS 1970 *subject classifications.* Primary 49C15; Secondary 49A05.
Key words and phrases. Decision model, concealed state space, standard Borel space.

considerations to stationarity and to more special transition probabilities. Moreover, in the present article both state spaces and the action space are assumed to be more general, namely Borel-spaces. In the papers of Dynkin and Sawaragi, Yoshikawa, respectively, the spaces are assumed to be denumerable. In the following we mainly use the notation of Hinderer [5].

1. Definitions. In a Markovian decision model (MDM) with incomplete information there is an observed state space $(Z, \mathcal{Z})$, a concealed state space $(Z', \mathcal{Z}')$ and an action space $(A, \mathcal{A})$. Here $(Z, \mathcal{Z})$, $(Z', \mathcal{Z}')$ and $(A, \mathcal{A})$ are measurable spaces.

At time n let the system be in the state $(z_n, z_n') \in Z \times Z'$ and let the action $a_n \in A$ be selected. Which a_n can occur at all, is determined by the correspondence

$$D_n(z_n) \subset A .$$

D_n is a function from Z into the family of all subsets of A. When z_n is observed, $D_n(z_n)$ is the set of available actions after the observation. (The function $D_n(z_n)$ was introduced into the theory by Dynkin.) We make the assumption that

$$\{(z_n, a_n) : a_n \in D_n(z_n)\}$$

is measurable. In models with incomplete information the functions D_n are independent of the concealed states.

The transition from a state (z_n, z_n') at time n to a state (z_{n+1}, z_{n+1}') at time $n + 1$ while action a_n is being carried out is described by means of the transition probability (abbreviated t.p.)

$$q_n(z_n, z_n', a_n; \cdot)$$

from $Z \times Z' \times A$ to $Z \times Z'$. Here the point behind the semicolon stands for a probability measure (abbreviated p.m.) which depends on (z_n, z_n', a_n).

The p.m.

$$q_0(\cdot) ,$$

defined on $(Z \times Z', \mathcal{Z} \otimes \mathcal{Z}')$, describes the start of this process at time 1.

The state (z_n, z_n') and the action a_n give the real reward

$$r_n(z_n, z_n', a_n) .$$

The MDM with incomplete information is then written in the form

$$((Z \times Z', \mathcal{Z} \otimes \mathcal{Z}'), (A, \mathcal{A}), (D_n), (q_n), (r_n)) \tag{1}$$

(see Hinderer [5]). A MDM with complete information is a special case of this in which Z' consists of only one element.

ASSUMPTIONS. Additional to those listed above we assume

(i) The σ-algebra $\mathcal{Z}$ possesses a countable basis.

(ii) The σ-algebra $\mathcal{Z}'$ is standard Borel (see Mackey [9]).

(iii) The functions r_n satisfy one of the two assumptions C^+ or C^- (see below Section 2).

2. Optimal plans. A plan $\Pi = (\Pi_n)$ is a sequence of t.p.'s

$$\Pi_n(z_1, a_1, z_2, a_2, \cdots, z_n; \cdot)$$

from $(Z \times A)^{n-1} \times Z$ to A with the property

$$\Pi_n(z_1, a_1, z_2, a_2, \cdots, z_n; C) = 1$$

for every $C \in \mathfrak{A}$ which contains $D_n(z_n)$.

Let the set of all the plans of the MDM (1) be Δ. It does not seem sensible to consider any plans which depend also on the concealed history since one cannot work with them in practice. For the same reason the functions D_n also depend only on the observed state. The decisive quality of a MDM with incomplete information lies in this.

According to the theorem of Kolmogoroff the p.m.

$$P_\Pi \cdot = q_0 \otimes \bigotimes_{n=1}^{\infty} (\Pi_i \otimes q_i)$$

can now be defined on the measurable space

$$((Z \times Z' \times A)^{\mathbf{N}}, (\mathfrak{Z} \otimes \mathfrak{Z}' \otimes \mathfrak{A})^{\otimes \mathbf{N}})$$

for every $\Pi \in \Delta$.

Now

C^+: $\quad \sup_{\Pi \in \Delta} \int \sum_{n=1}^{\infty} r_n^+ \, dP_\Pi < \infty$ or

C^-: $\quad \sup_{\Pi \in \Delta} \int \sum_{n=1}^{\infty} r_n^- \, dP_\Pi < \infty$

is used as an assumption for the reward functions. These assumptions seem to be the most general ones found in the literature, containing e.g. the case of discounting factor 1 if the (stationary) reward is of constant sign (see Hinderer [6]).

Let the total reward on using a plan Π be

$$G_\Pi \cdot = \int \sum_{n=1}^{\infty} r_n \, dP_\Pi \, .$$

We look for plans for which G_Π will be maximized.

3. Construction of a new model. In MDM's with incomplete information it is not possible to limit oneself to Markovian plans instead of all the plans out of Δ. The derivations of e.g. Blackwell [1], [2], Strauch [12] and Hinderer [5] are not applicable here because they would lead in this case to plans which also depend on the concealed history.

Instead, a MDM

$$((Z \times V, \mathfrak{Z} \otimes \mathfrak{V}), (A, \mathfrak{A}), (D_n), (p_n), s_n) \tag{2}$$

can be defined which is now a MDM with complete information because $z \in Z$ can be observed and $v \in V$ can be calculated from the observations.

In this Z, $\mathfrak{Z}$, A, $\mathfrak{A}$ and (D_n) are to be understood as in (1), whereas the remaining terms are to be defined as follows:

(α) V is the set of all the p.m.'s on $(Z', \mathfrak{Z}')$, and $\mathfrak{V}$ is the smallest σ-algebra on V referring to which all maps φ_B (for $B \in \mathfrak{Z}'$) from V into $[0, 1]$, defined as

$$\varphi_B(\mu) = \mu(B)$$

are measurable.

(β) For the definition of p_n a theorem of Rhenius [10] (Hauptsatz 4.11) is required. From it follows, namely, that the t.p.

$$u_n(z_n, v_n, a_n; \cdot) \mathrel{.}= \int q_n(z_n, z_n', a_n; \cdot) v_n(dz_n')$$

from $Z \times V \times A$ to $Z \times Z'$ can be factorized in the form

$$u_n(z_n, v_n, a_n; \cdot) = \bar{u}_n(z_n, v_n, a_n; \cdot) \otimes w_n(z_n, v_n, a_n, z_{n+1}; \cdot),$$

with the t.p. $\bar{u}_n$ from $Z \times V \times A$ to Z and the t.p. w_n from $Z \times V \times A \times Z$ to Z'. Thus

$$g_n : Z \times V \times A \times Z \to V,$$

defined by

$$g_n(z_n, v_n, a_n, z_{n+1}) \mathrel{.}= w_n(z_n, v_n, a_n, z_{n+1}; \cdot),$$

is a measurable map (see Hinderer [5], page 85). For this reason

$$\begin{aligned} t_n(z_n, v_n, a_n, z_{n+1}; W) &\mathrel{.}= 1, && \text{in the case of } g_n(z_n, v_n, a_n, z_{n+1}) \in W, \\ &\mathrel{.}= 0, && \text{in the case of } g_n(z_n, v_n, a_n, z_{n+1}) \notin W, \end{aligned}$$

($W \in \mathfrak{V}$), defines a t.p. from $Z \times V \times A \times Z$ to V.

With this we define

$$p_n(z_n, v_n, a_n; \cdot) = \bar{u}_n(z_n, v_n, a_n; \cdot) \otimes t_n(z_n, v_n, a_n, z_{n+1}; \cdot).$$

This derivation is valid for $n > 0$. A corresponding one is valid for $n = 0$.

(γ) The reward functions are defined as

$$s_n(z_n, v_n, a_n) \mathrel{.}= \int r_n(z_n, z_n', a_n) v_n(dz_n').$$

As the history in (2) can be observed completely, the plans depend on the whole history:

$$\Pi_n(z_1, v_1, a_1, z_2, \cdots, z_n, v_n; \cdot).$$

Let the set of these plans be Γ; we then have

$$\Delta \subset \Gamma.$$

For every $\Pi \in \Gamma$ we can now define the p.m.

$$Q_\Pi \mathrel{.}= p_0 \otimes \bigotimes_{n=1}^{\infty} (\Pi_n \otimes p_n)$$

on $((Z \times V \times A)^{\mathbb{N}}, (\mathfrak{Z} \otimes \mathfrak{V} \otimes \mathfrak{A})^{\otimes \mathbb{N}})$.

LEMMA 1. *Let $h(z_i, a_i, z_{i+1}, z_{i+1}')$ be a real-valued measurable function. Then*

$$\begin{aligned} &\int h(z_i, a_i, z_{i+1}, z_{i+1}') q(z_i, z_i', a_i; d(z_{i+1}, z_{i+1}')) v_i(dz_i') \\ &\qquad = \int h(z_i, a_i, z_{i+1}, z_{i+1}') v_{i+1}(dz_{i+1}') p_i(z_i, v_i, a_i; d(z_{i+1}, v_{i+1})) \end{aligned}$$

is valid according to the terms defined above and on condition that the left side is defined.

PROOF. According to the definition of u_i the left side is

$$\begin{aligned}&\int h(z_i, a_i, z_{i+1}, z'_{i+1})u_i(z_i, v_i, a_i; d(z_{i+1}, z'_{i+1}))\\ &\quad = \int h(z_i, a_i, z_{i+1}, z'_{i+1})\, w_i(z_i, v_i, a_i, z_{i+1}; dz'_{i+1})\bar{u}_i(z_i, v_i, a_i; dz_{i+1})\\ &\quad = \int h(z_i, a_i, z_{i+1}, z'_{i+1})v_{i+1}(dz'_{i+1})t_i(z_i, v_i, a_i, z_{i+1}; dv_{i+1})\bar{u}_i(z_i, v_i, a_i; dz_{i+1}),\end{aligned}$$

and that is the right side of the assertion.
A consequence of Lemma 1 is

LEMMA 2. *For the functions r_n from* (1) *and s_n from* (2), *and* $\Pi \in \Delta$,

(a) $$\int s_n^+ \, dQ_\Pi \leqq \int r_n^+ \, dP_\Pi ,$$

(b) $$\int s_n^- \, dQ_\Pi \leqq \int r_n^- \, dP_\Pi$$

are valid.

PROOF. It is easy to see that

$$s_n^+(z_n, v_n, a_n) \leqq \int r_n^+ (z_n, z_n', a_n)v_n(dz_n') .$$

Therefore we get, employing Lemma 1:

$$\begin{aligned}&\int s_n^+(z_n, v_n, a_n)\, dQ_\Pi\\ &\qquad \leqq \int r_n^+(z_n, z_n', a_n)v_n(dz_n')\Pi_n(z_1, a_1, \cdots, z_n; da_n)\\ &\qquad\qquad \times p_{n-1}(z_{n-1}, v_{n-1}, a_{n-1}; dz_n)\Pi_{n-1}(z_1, a_1, \cdots, z_{n-1}; da_{n-1})\\ &\qquad\qquad \times d(p_0 \otimes \textstyle\bigotimes_{i=1}^{n-2} (\Pi_i \otimes p_i))\\ &\qquad = \int r_n^+ (z_n, z_n', a_n)\Pi_n(z_1, a_1, \cdots, z_n; da_n)\\ &\qquad\qquad \times q_{n-1}(z_{n-1}, z'_{n-1}, a_{n-1}; d(z_n, z_n'))v_{n-1}(dz'_{n-1})\\ &\qquad\qquad \times \Pi_{n-1}(z_1, a_1, \cdots, z_{n-1}; a_{n-1})\, d(p_0 \otimes \textstyle\bigotimes_{i=1}^{n-2} (\Pi_i \otimes q_i)) .\end{aligned}$$

In the same way, with repeated use of Lemma 1, we continue until we have

$$\begin{aligned}&\int s_n^+(z_n, v_n, a_n)\, dQ_\Pi\\ &\qquad \leqq \int r_n^+(z_n, z_n', a_n)\Pi_n(z_1, a_1, \cdots, z_n; da_n)\, d(q_0 \otimes \textstyle\bigotimes_{i=1}^{n-1} (\Pi_i \otimes q_i))\\ &\qquad = \int r_n^+ (z_n, z_n', a_n)\, dP_\Pi .\end{aligned}$$

Part (b) follows analogously if

$$(x - y)^- \leqq y \qquad (x, y \geqq 0)$$

is taken into consideration.

If we limit ourselves to the plans out of Δ, then according to Lemma 2 C^+ and C^- are valid for (s_n) and Q_Π if the same qualities are valid for (r_n) and P_Π. Thus we can define

$$F_\Pi \cdot = \int \textstyle\sum_{n=1}^{\infty} s_n \, dQ_\Pi ,$$

for $\Pi \in \Delta$, as the total reward of the plan $\Pi \in \Delta$ in model (2).

THEOREM 3. *If the* MDM (2) *is derived from* (1) *with the help of the described transformation, then for* $\Pi \in \Delta$ *we have*

$$F_\Pi = G_\Pi .$$

PROOF. On account of C^+ and C^- and Lemma 2

$$G_\Pi = \sum_{n=1}^{\infty} \int r_n \, dP_\Pi$$

and

$$F_\Pi = \sum_{n=1}^{\infty} \int s_n \, dQ_\Pi$$

are valid. Therefore the theorem has been proved if it can be shown that

$$\int r_n \, dP_\Pi = \int s_n \, dQ_\Pi$$

for each n. This succeeds, however, immediately with Lemma 1.

4. Equivalence of the models. In order to prove that the models (1) and (2) correspond completely it must be shown that C^+ or C^- respectively are valid for s_n, Q_Π and Γ instead of for r_n, P_Π and Δ, and that with any plan from Γ we cannot obtain a greater reward than with the plans out of Δ. The total reward must also be defined first for plans out of Γ which do not lie in Δ. For this we need Lemma 4 which applies the following notations:

$$h_i \cdot = (z_1, a_1, \cdots, z_i) \qquad \text{for } i \geqq 1 ,$$
$$v_1(h_1) \cdot = g_0(z_1) ,$$
$$f_1(h_1) \cdot = (z_1, v_1(z_1)) ,$$

and for $n \geqq 1$:

$$v_{n+1}(h_n, a_n, z_{n+1}) \cdot = g_n(z_n, v_n(h_n), a_n, z_{n+1}) ,$$
$$f_{n+1}(h_n, a_n, z_{n+1}) \cdot = (f_n(h_n), a_n, z_{n+1}, v_{n+1}(h_n, a_n, z_{n+1})) .$$

LEMMA 4. *Let* $\Pi = (\Pi_n) \in \Gamma$. *If one defines* $\sigma = (\sigma_n) \in \Delta$ *by*

$$\sigma_n(h_n; \cdot) \cdot = \Pi_n(f_n(h_n); \cdot)$$

then each function $e_n(z_n, v_n, a_n)$ *which is integrable referring to* Q_σ *is also integrable referring to* Q_Π, *and*

$$\int e_n(z_n, v_n, a_n) \, dQ_\sigma = \int e_n(z_n, v_n, a_n) \, dQ_\Pi$$

is valid.

The conclusion of Lemma 4 follows from the preceding definitions and from the definition of Q_σ and Q_Π respectively.

One of the immediate consequences of Lemma 2 and 4 is

THEOREM 5. *If the functions* (r_n) *satisfy the conditions* C^+ *or* C^-, *respectively, then the functions* (s_n) *satisfy the conditions*

$$C_s^+ : \sup_{\Pi \in \Gamma} \int \sum_{n=1}^{\infty} s_n^+ \, dQ_\Pi < \infty ,$$
$$C_s^- : \sup_{\Pi \in \Gamma} \int \sum_{n=1}^{\infty} s_n^- \, dQ_\Pi < \infty , \qquad \textit{respectively.}$$

With this we can define the total reward

$$F_\Pi = \int \sum_{n=1}^{\infty} s_n \, dQ_\Pi$$

for each plan $\Pi \in \Gamma$. Likewise Theorem 6 follows immediately from Lemma 4:

THEOREM 6. *With the above assumptions for every* $\Pi \in \Gamma$ *there is a* $\sigma \in \Delta$ *so that*

$$F_\Pi = F_\sigma = G_\sigma$$

is valid.

In this case σ is defined as in Lemma 4.

5. Concluding remarks. Theorems 3 and 4 show that the models (1) and (2) are the same with respect to their criterion of optimality. (2) has the advantage over (1) in that (2) is a MDM with a completely observed history. Thus we can apply to (2) the theorems which are known about the existence of optimal Markovian plans. See here e.g. the publications of Blackwell [1], [2], Strauch [12] and Hinderer [5], [6], for which it is important to note that $(V, \mathscr{V})$ is standard Borel if $(Z', \mathscr{Z}')$ is too (see Hinderer [5], Theorem 12.13).

A special case of model (2) arises if D_n, q_n and r_n in model (1) are independent of z_n (i.e. D_n constant, $q_n(z_n' a_n; \cdot)$, $r_n(z_n', a_n)$). This assumption is sensible in learning models, and with it one can derive the following theorem (using only the notion of a sufficient statistic; see Hinderer [5], Chapter 18):

THEOREM 7. *If the functions* D_n, q_n *and* r_n *in* (1) *are independent of* z_n *and if* C^+ *is valid, then there is a deterministic plan* $g = (g_n) \in \Gamma$ *for each plan* $\Pi \in \Gamma$ *so that*

$$g_n : V \to A$$

and

$$F_g \geqq F_\Pi$$

are valid.

REFERENCES

[1] BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

[2] BLACKWELL, D. (1967). Positive dynamic programming. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* I, **1** 415–418.

[3] DYNKIN, E. B. (1966). Controlled random sequences. *Theor. Probability Appl.* **10** 1–14.

[4] FELDMAN, D. (1962). Contributions to the "two-armed-bandit" problem. *Ann. Math. Statist.* **33** 847–856.

[5] HINDERER, K. (1970). Foundations of non-stationary dynamic programming with discrete time parameter. *Lecture Notes in Operations Research and Mathematical Systems.* Springer-Verlag, Berlin.

[6] HINDERER, K. (1971). Instationäre dynamische Optimierung bei schwachen Voraussetzungen über die Gewinnfunktionen. *Abh. Math. Sem. Univ. Hamburg* **36** 208–223.

[7] IOSIFESCU, M. and THEODORESCU, R. (1969). *Random Processes and Learning.* Springer-Verlag, Berlin.

[8] KARUSH, W. and DEAR, R. E. (1967). Optimal strategy for item presentation in a learning process. *Management Sci.* **13** 773–785.

[9] MACKEY, G. W. (1957). Borel structure in groups and their duals. *Trans. Amer. Math. Soc.* **85** 134–165.

[10] RHENIUS, D. (1971). Markoffsche Entscheidungsmodelle mit unvollständiger Information und Anwendungen in der Lerntheorie. Dissertation, Universität Hamburg.

[11] SAWARAGI, Y. and YOSHIKAWA, T. (1970). Discrete-time Markovian decision processes with incomplete state observation. *Ann. Math. Statist.* **41** 78–86.
[12] STRAUCH, R. E. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37** 871–890.

PSYCHOLOGISCHES INSTITUT
UNIVERSITAT HAMBURG
2 HAMBURG 13, VON-MELLE-PARK 6
HAMBURG, GERMANY