SEQUENTIAL METHODS IN STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 16
PWN-POLISH SCIENTIFIC PUBLISHERS
WARSAW 1985

OPTIMAL DECISION RULES

H.-J. GIRLICH

Section of Mathematics, University of Leipzig,
Leipzig, G.D.R.

1. Introduction

This paper deals with the optimal control of dynamic systems such as a machine replacement, inventory control or production planning when the probabilistic and the reward structure of the system are not completely known.

Usually, the Bayesian approach or adaptive control (cf. [23]) is used to solve such problems. A careful analysis of the kind of uncertainty shows that in some cases the unloved mini-max criterion may be a useful alternative. The emphasis is on finding optimality conditions in terms of decision rules in Markov games, which constitute a natural model for the irregular case. The advantage of this setup will become clear in solving an inventory problem in an exemplary manner.

2. A classification

We introduce a simple classification of models for the control of a dynamic system with incomplete information about the transition law. Here, the state space has two components, an observable part X and a concealed part B. If only one element is in B, i.e., we know the transition law, we have the usual Markov decision model in the sense of Blackwell (cf. [2]). If B contains exactly two elements:

$$B = \{0, 1\},$$

then we distinguish three types of sequential patterns

$$0-0-0-\ldots-0-0-0-\ldots, \tag{1}$$

$$0-0-0-\ldots-0-1-1-\ldots, \tag{2}$$

$$1-0-1-\ldots-0-0-1-\ldots, \tag{3}$$

i.e., the sequence of noncontrollable concealed states of the process may be constant (1), have a jump (2), or be irregular (3).

The classical example of the constant-pattern case is the two-armed bandit problem, first considered by Bellman ([1]). Waldmann ([21]) gives a remarkable insight into the numerical effort, which is even required when B consists of just two elements. Recently, some further interesting results have been obtained for the more general case of a finite set B and a constant pattern, cf. [7] for the multi-armed bandit problem and [8] for Markov chains. Rieder ([16]) has provided the probably most general Bayesian decision model with a constant pattern (B infinite).

The quickest-detection problem is an example of the jump-pattern case. This problem can be reduced to finding the optimal stopping time for specific Markov random sequences (cf. [18]).

When the transition law changes with running time in an unpredictable manner, then we have the irregular case. Here, it is expedient to choose a game against nature as a model for this situation. Clearly, this is a pessimistic choice, but there is no simple alternative to it. But contrary to the constant-pattern case (cf. [8], Appendix B), now the minimax criterion has the nice property of the method of successive approximation being applicable to computing optimal decision rules.

Before we proceed to applying this approach to an inventory model, let us introduce a formal framework containing games, such as the terminating game in the sense of [17] and the perfect-information game in the sense of [3].

3. Markov games

We consider a tuple $M = (X, A, B, q, E, F, k)$ of the following objects: the state space X, the action spaces A and B for players I and II, respectively, a Markovian transition law q from $X \times A \times B$ to X; E and F are sets of admissible strategies of I and II respectively, which we shall explain at once, k is the bounded cost function defined on $X \times A \times B \times X$.

Let $\mathscr{A}$ and $\mathscr{B}$ be maps with $\mathscr{A}(x) \subseteq A$, $\mathscr{B}(x, a) \subseteq B$ for all $a \in \mathscr{A}(x)$ and $x \in X$.

For $H_n := X \times A \times B \times \ldots \times X \times A \times B \times X$, the transition probabilities π_n: $H_n \to A$, ϱ_n: $H_n \times A \to B$ with

$$\pi_n\big(\mathscr{A}(x_n) \mid x_0, a_0, b_0, \ldots, x_n\big) = 1,$$

$$\varrho_n\big(\mathscr{B}(x_n, a_n) \mid x_0, a_0, b_0, \ldots, x_n; a_n\big) = 1$$

will be called the *admissible decision rules* at time n for I or II respectively. Now, let

$$E := \mathop{\times}_{n \in N} E_n, \qquad F := \mathop{\times}_{n \in N} F_n,$$

where E_n, F_n are given sets of admissible decision rules at time n.

If F_n is the set of all admissible decision rules at n, for all $n \in N$, we call $M = M^p$ a *Markov game with perfect information.*

If $F_n = F_n^s$ is the set of all decision rules ϱ_n for which $\varrho_n(\cdot \mid h; a)$ is independent of $a \in A$ for every $h \in H_n$, $n \in N$, we speak of a *Markov game* M^s *with simultaneous action choice* – the *usual Markov game.*

E^m denotes the set of all Markovian decision rules π_n, i.e., for rules $\pi_n(\cdot \mid x_0, a_0, b_0, \ldots, x_n)$ depending only on x_n. If $\pi_n \in E^m$ is a degenerate distribution, i.e., there is a function e with $\pi_n(\{e(x_n)\} \mid x_0, a_0, b_0, \ldots, x_n) = 1$, then we write $\pi_n = \delta_e$ and denote the set of such decision rules by E^d. Similarly we can define F^m and F^d for player II.

A sequence $\pi = (\pi_n)$, $\varrho = (\varrho_n)$ of admissible decision rules π_n, ϱ_n will be called *strategy* of I or II respectively. If $\pi_n \in E^m$ and $\pi_n = \pi_0$ for every $n \in N$, then $\pi = (\pi_n)$ is said to be stationary and we write $\pi = \pi_0^\infty$. If, in addition, $\pi_0 \in E^d$, then a function e exists and $\pi = \delta_e^\infty$ is called a *deterministic stationary strategy with the policy e.*

In order to describe the decision processes connected with our Markov game, we use the well-known Ionescu-Tulcea theorem defining for every strategy $\pi \in E$ for I and $\varrho \in F$ for II and every initial state x a probability measure $p_{\pi\varrho}^x$ according to

$$\int_\Omega u(x_0, a_0, b_0, \ldots, x_n)\, p_{\pi\varrho}^x(d\omega)$$

$$= \int_A \pi_0(da_0 \mid x) \int_B \varrho_0(db_0 \mid x, a_0) \int_X q(dx_1 \mid x, a_0, b_0) \ldots$$

$$\ldots \int_B \varrho_n(db_n \mid x, a_0, b_0, \ldots, a_{n-1}) \int_X q(dx_n \mid x_{n-1}, a_{n-1}, b_{n-1})\, u(x, \ldots, x_n)$$

for every bounded u on H_n.

We define for $\omega \in \Omega = H_\infty$ the *total discounted cost*

$$K(\omega) := \sum_{n=0}^{\infty} \alpha^n k(x_n, a_n, b_n, x_{n+1}).$$

With every pair of admissible strategies (π, ϱ) and an initial state x, we associate the expected total discounted cost due to the player I:

$$v_{\pi\varrho}(x) := \int_\Omega K(\omega)\, p_{\pi\varrho}^x(d\omega).$$

Let

$$\bar{v}(x) := \inf_{\pi \in E} \sup_{\varrho \in F} v_{\pi\varrho}(x)$$

and

$$\underline{v}(x) := \sup_{\varrho \in F} \inf_{\pi \in E} v_{\pi\varrho}(x), \qquad x \in X.$$

If $\underline{v}(x) = \bar{v}(x) =: v(x)$ for all $x \in X$, then the function v is called the *value of the game* M. Using the minimax criterion, a strategy $\pi' \in E$ is called *optimal* (*in* M) if $v_{\pi'\varrho} \leqslant v$, $\forall \varrho \in F$; $\varrho' \in F$ is called *optimal* (*in* M) if $v \leqslant v_{\pi\varrho'}$, $\forall \pi \in E$.

If we interpret M as a game against nature, then player I is the decision-maker and player II is nature. We are looking for an optimal strategy of the decision-maker and the expected cost, which he has to pay at most, given by the value of M.

4. Simultaneous action-choice games

In his famous paper Shapley ([17]) reduced the optimal-strategy problem in a special Markov game to a study of a contracting operator U generated by decision rules. In our model, the operator has the form

$$Uw := \inf_{p \in E^m} \sup_{r \in F^m} I(p, r) w,$$

where

$$[I(p, r)u](x) := \int_A p(da|x) \int_B r(db|x, a)\, Tu(x, a, b)$$

with

$$Tu(x, a, b) := \int_X q(dy|x, a, b)\big(k(x, a, b, y) + \alpha u(y)\big).$$

Shapley showed, for a terminating Markov game with simultaneous action choice and a finite state space, the following statements:

(a) *The unique fixed point of the contracting operator U is the value v of M^s.*

(b) *If the Markovian decision rules p' and r' satisfy the inequalities*

$$I(p', r)v \leqslant v \leqslant I(p, r')v$$

for all $p \in E^m$, $r \in F^m$ then the stationary strategies p'^∞ and r'^∞ are optimal.

A straightforward generalization of these results for a contracting Markov game M^s with a countable state space is given by van der Wal ([19], [20]) and Wessels ([22]). A generalization for a metric state space is presented by Maitra and Parthasarathy ([15]), Idzik ([10]) and Couwenbergh ([4]).

All these authors proved that, under some conditions, M^s has a value and both players have optimal stationary strategies, based on two Markovian decision rules, which are not, in general, deterministic.

Iwamoto and Kai ([12]) showed that, under somewhat restrictive conditions, both players have optimal deterministic stationary strategies.

Modifying the proof of Theorem 5.1, [12], using a selection theorem given by Himmelberg, Parthasarathy and van Vleck ([9]) we obtain (cf. [6]):

PROPOSITION 1. *Let $\mathscr{A}$ and $\mathscr{B}$ be compact-convex valued, continuous and $\mathscr{B}(x, a) = \mathscr{B}(x, a')$ for all $a, a' \in \mathscr{A}(x)$. If u' is a solution of the optimality equation*

$$u(x) = \inf_{a \in \mathscr{A}(x)} \sup_{b \in \mathscr{B}(x)} Tu(x, a, b), \qquad x \in X, \tag{4}$$

with the properties

(i) *$Tu'(x, \cdot, b)$ convex on $\mathscr{A}(x)$;*

(ii) *$Tu'(x, a, \cdot)$ concave on $\mathscr{B}(x)$;*

(iii) *$Tu'(x, \cdot, \cdot)$ continuous on $\mathscr{A}(x) \times \mathscr{B}(x)$ for all $x \in X$, $a \in A$, $b \in B$,*

then there exist optimal stationary deterministic strategies for both players and u' is the value of the game M^s.

As we shall see later, in applied models the convexity conditions of Proposition 1 are often not fulfilled. Clearly, it is impossible to weaken these conditions considerably so that with the value also optimal stationary deterministic strategies for both players exist. Since in a game against nature we are interested principally in optimal strategies for the decision-maker only, we reduce our Markov game to a special Markovian decision problem with the criterion function $v_\pi = \sup_\varrho v_{\pi\varrho}$. Enlarging the set of all strategies for the opponent, we shall get weak conditions ensuring an optimal stationary deterministic strategy for the decision-maker in the next section.

5. Perfect-information games

We consider a Markov game with perfect information M^p. Contrary to the usual Markov game M^s, player II chooses his action b_n at the time n knowing the choice a_n of player I. Shapley's results are also the background of the following

PROPOSITION 2. *If u' is a bounded solution of the optimality equation* (4) *and there exists a decision rule $\delta_e \in E^d$ with the property*

$$I(\delta_e, \varrho_n) u' \leqslant u', \qquad \textit{for all} \quad \varrho_n \in F^m, \tag{5}$$

then the strategy δ_e^∞ is optimal.

For a proof see [13]. A decision rule for player I with property (5) is called an optimal decision rule.

The next assertion gives us sufficient conditions to ensure the existence of a deterministic optimal decision rule. For convenience we define

$$C_1 := \{(x, a)\colon a \in \mathscr{A}(x)\}, \qquad C_2 := \{(a, b)\colon a \in \mathscr{A}(x), b \in B(x, a)\}.$$

PROPOSITION 3. *Let M^p be a Markov game with complete information and the properties:*

(i) *$\mathscr{A}(x)$ is compact for every $x \in X$,*

(ii) *$\mathscr{B}(x, a)$ is compact for every $(x, a) \in C_1$,*

(iii) *for every a and every sequence (a_n) with $a, a_n \in \mathscr{A}(x)$ and $\lim_{n\to\infty} a_n = a$ we have*

$$\bigcap_{\varepsilon>0} \bigcup_{k=0}^{\infty} \bigcap_{n=k}^{\infty} \mathscr{B}(x, a_n) \supseteq \mathscr{B}(x, a).$$

(iv) *$\int_X u(x) q(dx|\cdot, \cdot)$ is continuous for every real bounded measurable function u on X,*

(v) *$(\int kq)(x, a, \cdot)$ is continuous on $\mathscr{B}(x, a)$ for all $(x, a) \in C_1$,*

(vi) *$(\int kq)(x, \cdot, \cdot)$ is lower semi-continuous on C_2 for every $x \in X$.*

Then, there is an optimal stationary deterministic strategy for player I and the value v of the game M^p is the unique bounded measurable solution of the optimality equation:

$$u(x) = \inf_{a \in \mathscr{A}(x)} \sup_{b \in \mathscr{B}(x,a)} \mathcal{T}u(x, a, b) \tag{6}$$

for all $x \in X$.

For a proof see [14] (Theorem 4.7).

6. The optimal inventory equation

For the well-known infinite-horizon dynamic inventory model, where backlogging is allowed, Iglehart [11] has studied the optimality equation

$$f(x) = \min_{x \leqslant a} \left(K \operatorname{sgn}(a-x) + L_b(a) + \alpha \int_0^\infty f(a-\beta)\, b(d\beta)\right).$$

In the case of incomplete information about the demand distribution b, i.e., only $b \in B$ is known, we get formally the optimal min-max inventory equation

$$u(x) = \min_{x \leqslant a \leqslant C} \max_{b \in B} \left(K \operatorname{sgn}(a-x) + L_b(a) + \alpha \int_0^\infty u(a-\beta)\, b(d\beta)\right) \tag{7}$$

where C is the capacity of the facility.

Our framework leads to equation (7) for an inventory system described by a Markov game with the following objects:

(a) $X := \{x \in \boldsymbol{R}\colon x \leqslant C\}$; the state x is the stock level of the facility.

(b) $A := \{a \in \boldsymbol{R}\colon a \leqslant C\}$; the action a of the decision-maker is the stock level immediately after ordering.

(c) $\mathscr{A}(x) := \{a \in A\colon a \geqslant x\}$ – the set of all admissible actions at the state x.

(d) B – a compact set of probability measures in $\boldsymbol{R}_+$.

(e) $\mathscr{B}(x, a) = B$ for all $a \in \mathscr{A}(x)$, $x \in X$.

(f) $\int_X u(y)\, q(dy|x, a, b) := \int u(a-\beta)\, b(d\beta)$ – backlogging of the unfilled demand.

(g) $\int_X k(x, a, b, y)\, q(dy|x, a, b) := K \operatorname{sgn}(a-x) + (a-x)c + L_b(a)$, where $\operatorname{sgn} x = 1$ for $x > 0$ but equal to 0 for $x \leqslant 0$; $L_b(a) := \int_0^\infty l(a, \beta)\, b(d\beta)$ – the expected holding and shortage cost for one period; K and c are non-negative ordering cost factors.

PROPOSITION 4. *In an inventory system with the properties* (a) *to* (g) *there exists an optimal stationary deterministic strategy* δ_e^∞. *The minimal expected discounted total cost is the unique bounded solution of the optimal min-max inventory equation* (7).

If, in addition, $L_b(\cdot)$ *is convex for every* $b \in B$, *then the optimal decision rule* δ_e *is of* (s, S) *type, i.e., there exist numbers* s, S *with* $s \leqslant S$ *and*

$$e(x) = \begin{cases} x; & x > s, \\ S; & x < s. \end{cases}$$

Proof. The assumptions (i) to (vi) of Proposition 3 are fulfilled, thus the first part holds. Since in the case of the n-period problem the optimal decision rule in each period is always of the (s, S) type (cf. [5]), the second part follows straightforwardly as in the proof of Theorem 3 in [11].

We note that in our inventory system with convex L_b and $K = 0$, Proposition 1 is applicable, too. This means that even in the case of simultaneous action choice there is no better strategy for the decision-maker than δ_e^∞. But for $K > 0$, condition (i) of Proposition 1 does not hold.

7. References

[1] R. Bellman, *A problem in the sequential design of experiments*, Sankhyā Ser. A. **16** (1956), 221–229.

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

[3] D. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*, New York 1954.

[4] H. A. M. Couwenbergh, *Stochastic games with metric state space*, Internat. J. Game Theory **9** (1980), 25–36.

[5] V. Dietzsch, *Dynamische Minimax-Entscheidungsmodelle*, Theses, Karl-Marx-Universität, Leipzig 1977.

[6] H.-J. Girlich and H.-U. Küenle, *On dynamic min-max decision models*, in *Trans. 9th. Prague Conference 1982*, Vol. A, Prague 1983, 257–262.

[7] J. C. Gittins, *Bandit processes and dynamic allocation indices*, J. Roy. Statist. Soc. Ser. B **41** (1979), 148–177.

[8] K. van Hee, *Bayesian Control of Markov Chains*, Mathematisch Centrum, Amsterdam 1978.

[9] C. Himmelberg, T. Parthasarathy, and F. van Vleck, *Optimal plans for dynamic programming problems*, Math. Oper. Res. **1** (1976), 390–394.

[10] A. Idzik, *Remarks on discounted stochastic games*, in *Trans. 8th. Prague Conference 1978*, Vol. C, Prague 1979, 165–174.

[11] D. L. Iglehart, *Optimality of (s, S) policies in the infinite horizon dynamic inventory problem*, Management Sci. **9** (1963), 259–267.

[12] S. Iwamoto and Y. Kai, *On deterministic stationary strategies for Markov games*, Bull. Math. Statist. Res. **16** (1974), 71–82.

[13] H.-U. Küenle, *Über die Optimalität von Strategien in stochastischen dynamischen Minimax-Entscheidungsmodellen, I*, Math. Operationsforsch. Statist. Ser. Optim. **12** (1981), 421–435.

[14] —, *On ε-optimal strategies in discounted Markov games*, in: J. Zabczyk (ed.), *Optimal Control Theory*, Banach Center Publ., vol. 14, PWN–Polish Scientific Publishers, Warsaw 1984.

[15] A. Maitra and T. Parthasarathy, *On stochastic games*, J. Optim. Theory Appl. **5** (1970), 289–300.

[16] M. Rieder, *Bayesian dynamic programming*, Ad. in Appl. Probab. **7** (1975), 330–348.

[17] L. Shapley, *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A. **39** (1953), 1095–1100.

[18] A. N. Shiryayev, *Optimal Stopping Rules*, Springer-Verlag, New York 1978.

[19] J. van der Wal, *Discounted Markov games*, Internat. J. Game Theory **6** (1977), 11–22.

[20] —, *Stochastic Dynamic Programming*, Mathematisch Centrum, Amsterdam 1980.

[21] K.-H. Waldmann, *Numerical aspects in Bayesian inventory control*, Z. Oper. Res. **23** (1979), 49–60.

[22] J. Wessels, *Markov games with unbounded rewards*, Bonner Math. Schriften **98**, Univ. Bonn, Bonn 1977, 133–147.

[23] J. S. Zypkin, *Adaption und Lernen in kybernetischen Systemen*, Verlag Technik, Berlin 1970.

Presented to the semester
Sequential Methods in Statistics
September 7–December 11, 1981