*Proof.* Let $T$ be the first entry time in the set $E_0$. Of course, $\boldsymbol{Q}(T < \infty) = 1$. Let $\Omega_T = \{T < \infty\}$ and define a random sequence $(Y_n)_{n \geqslant 0}$ on $\Omega_T$ by

$$Y_n(\omega) = X_{T+n}(\omega).$$

By $\boldsymbol{Q}^Y$ we denote the restriction of $\boldsymbol{Q}$ on $\mathscr{F}^Y$ (which is a $\sigma$-algebra of subsets of $\Omega_T$). Set

$$\boldsymbol{P}^Y(A) = \frac{\boldsymbol{P}(A)}{\boldsymbol{P}(T < \infty)}$$

for all $A \in \mathscr{F}^Y$. (Notice that in view of Corollary 9 we have $\boldsymbol{P}(T < \infty) > 0$.) By the strong Markov property $(Y_n)_{n \geqslant 0}$ is Markov with respect to both $\boldsymbol{Q}^Y$ and $\boldsymbol{P}^Y$, admitting the transition probabilities $Q(x, A)$ and $P(x, A)$, respectively. Now we observe $\boldsymbol{Q}^Y(Y_0 \in E_0) = 1$. In view of Corollary 9 we find $\boldsymbol{Q}_n^Y \ll \boldsymbol{P}_n^Y$ for all $n \geqslant 0$. According to Theorem 17, $\boldsymbol{Q}^Y \ll \boldsymbol{P}^Y$ or $\boldsymbol{Q}^Y \perp \boldsymbol{P}^Y$. Using Theorem 5, we get $\boldsymbol{Q}(Z_\infty^Y = \infty) = 0$ or 1, where $Z_n^Y = d\boldsymbol{Q}_n^Y/d\boldsymbol{P}_n^Y$ and $Z_\infty^Y = \limsup_n Z_n^Y$. From formula (2) we derive

$$Z_{T+n} = Z_T(Z_0^Y)^{-1}Z_n^Y \quad \text{on } \{T < \infty\} \quad \boldsymbol{P}\text{-a.s. and } \boldsymbol{Q}\text{-a.s.}$$

Consequently, $\boldsymbol{Q}(Z_\infty = \infty) = 0$ or 1. Applying Theorem 5, the assertion now follows.

## References


[1] J. Feldman, *Equivalence and Perpendicularity of Gaussian Processes*, Pacific J. Math. 8 (1958), pp. 699–708.

[2] J. Hajek, *On a Property of Normal Distributions of an Arbitrary Stochastic Process*, Czechosl. Math. J. 8 (1958), pp. 610–618.

[3] S. Kakutani, *On Equivalence of Infinite Product Measures*, Ann. of Math. 49 (1948), pp. 214–226.

[4] R. S. Lipcer and A. N. Shiryaev, *Statistics of random processes*, Nauka, Moscow 1974 (in Russian).

[5] A. A. Lodkin, *Absolute continuity of measures corresponding to Markov processes with discrete time parameter*, Teorija Verojatn. i Primen. 16 (1971), pp 703–707 (in Russian).



*Presented to the semester*
*MATHEMATICAL STATISTICS*
*September 15–December 18, 1976*



MATHEMATICAL STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 6
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1980


# ON MARKOVIAN DECISION PROCESSES WITH UNBOUNDED REWARDS


H. J. GIRLICH

*University of Leipzig, Leipzig, G.D.R.*


## 1. Introduction

The concept of Markovian decision processes was first introduced by R. Bellman in 1957.

With the important contributions of R. A. Howard, D. Blackwell and others the mathematical foundations and the applications of this part of dynamic programming developed rapidly. It is very interesting to note how special applied topics make it necessary to extend the standard decision model.

If we consider e.g. queueing systems, it is natural to choose, for modelling, a countable state space — the number of customers waiting to be served — and an arbitrary action space. Furthermore, the queueing process carries rewards, where the negative rewards sometimes have a component that increases without bound with the state of the system.

When studying stochastic systems of this kind, it is not possible to apply directly the model of Blackwell [1], which assumes a uniformly bounded reward. Thus, we have two options: one is to transform the queueing model into a model of the uniformly bounded case, cf. [12], the other is to weaken Blackwell's assumption. The papers given by Harrison [4], [5], Lippman [8], [9], van Nunen [14], [15] and Wessels [16] about models whose state or action space is countable and which are sufficient for treating queueing systems lead in this direction.

This paper aims at giving a further generalization of Blackwell's model necessary for inventory systems.

An essential property of some inventory models is that their state and action space have the same structure.

Therefore, we consider Markovian decision processes with both state and action space being uncountable and rewards unbounded.

In Section 2 we shall outline fundamental definitions and results of the standard model (cf. [1]), which is to be modified in Section 3 and applied to an inventory system in Section 4. An elaborate discussion of problems and results of this paper may be found in [2], [3], [7] and [10].

## 2. The standard model

Markovian decision processes are characterized by three structures: a probabilistic, a decision and a reward structure. These structures are specified by four objects:

— a non-empty Borel subset of a Euclidean space, the so-called *state space* $X$,

— a non-empty Borel subset of a Euclidean space, the so-called *action space* $A$,

— a transition probability $q = q(\cdot/\cdot, a)$ from $X$ into itself, depending on some parameter $a \in A$, the so-called *law of motion*,

— a real-valued Baire function $r$ on $X \times A$, the so-called *reward* (*function*).

We describe the evolution of a system defined by $X$, $A$, $q$, and $r$ roughly as follows: After observing the current state $x_n$ we choose an action $a_n$ from the set $A$ of possible actions and receive an immediate reward $r(x_n, a_n)$, and the system moves to a new state $x_{n+1}$ according to the distribution $q(\cdot/x_n, a_n)$.

Now, we need the notion of a decision rule to define a Markovian decision process and a criterion of optimality to define a decision problem.

By a *decision rule* (or a *plan*) we mean a prescription for taking actions at each point in time, based on the knowledge of the whole history of the process which is described by $h_n := (x_1, a_1, \ldots, a_{n-1}, x_n)$ at the $n$th stage.

A plan $\pi$ specifies for each $h$ a probability distribution over $A$: $\pi(h_n) =: \pi_n$. A *stationary plan* is defined by a single Baire function $f$ mapping $X$ into $A$: we then write $\pi = f^\infty$. If the system is in the state $x_n$, we choose action $a_n := f(x_n) \in A$.

We take the infinite Cartesian product $\Omega := X \times A \times X \times A \times \ldots$ as a sample space. With the usual Borel $\sigma$-field $\mathcal{F} := \sigma(X) \otimes \sigma(A) \otimes \ldots$ we obtain the measurable space $(\Omega, \mathcal{F})$.

The application of any plan $\pi$ generates a probability measure $P^x := \delta_x \pi_1 q \pi_2 q \pi_3 \ldots$ and thus a stochastic process on $(\Omega, \mathcal{F}, P_\pi^x)$, the so-called *Markovian decision process* $(x_n, a_n)$, $n = 1, 2, \ldots$, which is determined by $\pi$ and starts in the state $x$.

The Markovian decision process is linked with the discounted expected total reward over the infinite future

$$v_\pi(x) := \sum_{n=1}^{\infty} \beta^{n-1} E_\pi^x r(x_n, a_n) \tag{1}$$

with the discount factor $\beta \in (0, 1)$.

Without some assumptions concerning the reward function $r$ or the law of motion $q$ there is no guarantee that under an arbitrary plan $v_\pi(x)$ exists for all $x \in X$.

Blackwell's assumption:

$$\text{(B1):} \qquad \sup_{(x,a) \in X \times A} |r(x, a)| \leqslant M < \infty,$$

ensures of course (in this case, called *uniformly bounded*) that, for each $x \in X$, $v_\pi(x)$ is now bounded over all plans $\pi$. The function $v_\pi$ with $v_\pi = v[\pi]$ is called the *return function of* $\pi$.



Our decision problem is to maximize the return function. But does there exist an optimal plan with

$$v_\pi^* \geqslant v_\pi \quad \text{for all } \pi\,?$$

In his famous paper [1] Blackwell gave an example showing that this is not always so. However, he showed that

THEOREM 1. *In the uniformly bounded case, for any probability distribution $p$ on $X$ and any $\varepsilon > 0$, there is a stationary plan $\pi^*$ such that, for every $\pi$,*

$$p(\{x: v_{\pi^*}(x) \geqslant v_\pi(x) - \varepsilon\}) = 1. \tag{2}$$

A plan $\pi^*$ with property (2) is called $(p, \varepsilon)$-*optimal*. We will carry over the statement of Theorem 1 to the unbounded case.

## 3. A model with unbounded rewards

In this section we shall modify the standard model replacing (B1) by weaker assumptions on $r$ but under a restriction on $q$.

(A0) *There exists a real-valued function $w$ on $X$ with $w \geqslant 1$ which is $q$-integrable and satisfies*

$$\sup_{a \in A} \int_X w(x') q(dx'/x, a) \leqslant \alpha w(x)$$

*for all $x \in X$ and a real number $\alpha < 1/\beta$.*

$$\text{(A1)} \qquad \sup_{a \in A} \frac{r(x, a)}{w(x)} \leqslant M < \infty.$$

Under these assumptions it may easily be shown that $v_\pi$ exists for all $\pi$ and the components are finite but not uniformly bounded.

THEOREM 2. *If* (A0), (A1) *hold, then, for any $\varepsilon > 0$ and a probability distribution $p$ on $X$ such that $w$ is $p$-integrable, there is a $(p, \varepsilon)$-optimal plan $\pi$. Furthermore, if $w$ is bounded $p$-almost everywhere, there is a stationary $(p, \varepsilon)$-optimal plan $f^\infty$.*

An outline of the proof may be given as follow: The first part of the statement is proved in an analogous way as in [1]. Using a selection theorem, we find for such a $\pi$ a plan $\pi'$ where each $\pi_n'$ is degenerated, i.e. $\pi_n'(x_n, f_n(x_n)) = 1$, and the return of $\pi'$ is only slightly smaller than that of $\pi$. We have yet to show that under these so-called Markov plans $\pi' = (f_1, f_2, \ldots)$, there is a stationary $(p, \varepsilon)$-optimal plan.

This is possible by [1], where Banach's fixed-point theorem is used. However, our return is unbounded and belongs to a certain linear space which is not complete under the metric induced by the normal supremum norm but complete with regard to a weighted supremum norm.

We define $\boldsymbol{B}_0$ to be the set of all real-valued Baire functions on $X$. With a function $w$ satisfying (A0) we introduce the norm

$$||u|| := \sup_{x \in X} \frac{|u(x)|}{w(x)}.$$

Now, we take the set

$$\boldsymbol{B} := \{u\colon u \in \boldsymbol{B}_0,\ ||u|| < \infty\}$$

and the metric

$$d(u, v) := ||u - v||$$

and show that $(\boldsymbol{B}, d)$ is a Banach space.

Let $T_f$ be a mapping associated with a particular stationary plan $f^\infty$ given by

$$[T_f u](x) := r(x, f(x)) + \beta \int_X u(x') q(dx'/x, f(x)) \tag{3}$$

and $U_{\pi'}$ a mapping associated with a Markov plan $\pi' = (f_1, f_2, \ldots)$ given by

$$U_{\pi'} u := \sup_n T_{f_n} u, \tag{4}$$

where for $f_m$ holds $r(\cdot, f_m(\cdot)) \in \boldsymbol{B}$; then $T_{f_n}$ and $U_{\pi'}$ are contracting mappings on $(\boldsymbol{B}, d)$ with unique fixed points on $\boldsymbol{B}$. The condition for our $(p, \varepsilon)$-optimal $\pi'$ is fulfilled and we may conclude the proof as in [1].

As a by-product we obtain the useful

THEOREM 3. *Let* (A0), (A1) *be valid.*

(i) *If* $\pi$ *and* $f^\infty$ *are plans with* $v_\pi \in \boldsymbol{B}$ *and* $T_f v_\pi \geqslant v_\pi$, *then*

$$v_{f^\infty} \geqslant T_f v \geqslant v_\pi. \tag{5}$$

(ii) *A plan* $\pi$ *is optimal if and only if its return* $v_\pi \in \boldsymbol{B}$ *satisfies the optimality equation*

$$u = \sup_{a \in A} T_a u, \tag{6}$$

*where* $T_a := T_f$ *with* $f(x) = a$ *for all* $x \in X$.

Proposition (i) extends Howard's plan improvement and (ii) Bellman's criterion of optimality.

## 4. An inventory model

The development of dynamic programming is closely connected with the development of the inventory theory beginning in the early 1950's with papers by Arrow, Harris, Marschak and Dvoretzky, Kiefer, Wolfowitz. Of course, the study of inventory processes as special cases of Markovian decision processes began only after its foundations had been provided at the end of the 1960's (cf. [6], [2]).

Before modelling we shall explain an inventory problem in a heuristic manner. We assume we have a facility for stocking several products. The stock levels are reduced by demand, which occurs with random size, and are reviewed periodically, with the manager having to decide each time whether additions to stock from an exogenous source by ordering, or reductions by selling are to be made. An optimum decision, in this discussion, is one which minimizes the sum of the cost associated with the inventory under the possibility that the facility offers.



Now, we consider a multi-product inventory model with periodic review. We denote the stock level of the product number $i$ by $x^{(i)}$. If $m$ different products are held, we have the state variable $x := (x^{(1)}, \ldots, x^{(m)})$, $x_n$ is the state at the beginning of the period number $n$, $X' := \boldsymbol{R}^m$ is the state space. We take $a := (a^{(1)}, \ldots, a^{(m)})$ as an action variable, where $a^{(i)}$ is the stock level of the product number $i$ immediately after decision-making, $A' := \boldsymbol{R}^m$ is the action space. The demands in each period are independent of previous periods and identically distributed random variables $\xi_n$ with a continuous probability density $\varphi$ and $\xi_n := (\xi_n^{(1)}, \ldots, \xi_n^{(m)})$, where each $\xi_n^{(i)}$ is non-negative with $E(\xi_n^{(i)}) = \mu^{(i)}$.

The law of motion $q'$ is then(1) given for every Borel set $B \subset R^m$ by

$$q'(B/x, a) := \int_{a - \xi \in B} \varphi(\xi)\, d\xi. \tag{7}$$

We have costs of the following types:

$l(a)$ — cost of stocking and shortage for being in stock position $a$ for one period, $l$ is a convex function on $A'$ with $l(a) + c \cdot a \to \infty$ with $a^{(i)} \to \pm\infty$;

$K$ — a fixed cost per order or selling;

$c \cdot (a - y)$ — linear ordering cost or selling reward

$$c := (c^{(1)}, \ldots, c^{(m)}) \quad \text{with} \quad c^{(i)} \geqslant 0.$$

Thus, we obtain the reward function

$$r'(x, a) := -[l(a) + K\Delta(x, a) + c \cdot (a - x)], \tag{8}$$

where $\Delta(x, a) = 0$, if $x = a$, and 1 else.

The objective is to choose a reorder/selling-policy, which minimizes the expected infinite horizon discounted costs. This is an admissible plan, which maximizes the discounted expected total reward over infinite future.

An *admissible plan* is a plan with restricted action-alternatives. Let $S_c := (S_c^{(1)}, \ldots, S_c^{(m)})$ describe the capacity of the facility and $S_s := (S_s, \ldots, S_s^{(m)})$ the maximal selling set in any period.

When the system is in the state $x$, we have only the possibility to choose an action $a$ from

$$A_x = \{a\colon x - S_s \leqslant a \leqslant S_c\}. \tag{9}$$

Thus, a slight modification yields the unrestricted model:

---

(1) backlogging of unfilled demand.

(10) $X := \{x: x \in X',\ x \leqslant S_c\}, \quad A := \bigcup_{x \in X} A_x,$

$q(\cdot/x, a) := q'(\cdot/x, S_c)$ and $r(x, a) := r'(x, S_c)$

if $a \in A \setminus A_x$, $q := q'$ and $r := r'$ if $a \in A_x$ for all $x \in X$.

This model has unbounded rewards. Blackwell's assumption is not satisfied by our $r$. If we take

$$w(x) := b - c \cdot x \quad \text{for all } x \in X \text{ and } b > 0,$$

then we have

$$\sup_{a \in A} \int_X w(x') q(dx'/x, a) = \max\bigl(w(x) + c \cdot S_s,\ w(S_c) + c \cdot \mu\bigr). \tag{11}$$

Furthermore, from (8) we obtain

$$\sup_{a \in A} r(x, a) \leqslant c \cdot S_s. \tag{12}$$

Now, we choose $b$ large enough for (A0) and (A1) to hold with $M = 1$ and a real $\alpha$ with $\alpha > 1$ and $\alpha\beta < 1$.

Therefore, we may apply to our case the results of the last section.(2) In particular, we are able to show that a stationary $(p, \varepsilon)$-optimal plan exists. However, in our special case this assertion can be strengthened by

THEOREM 4. *For the inventory model* (7), (8), (9), (10), *there is a region* $\sigma \subset X$ *and a point* $S \in X \setminus \sigma$ *such that* $f^\infty$ *is an optimal plan, where*

$$f(x) := \begin{cases} S & \text{for} \quad x \in \sigma, \\ x & \text{for} \quad x \in X \setminus \sigma. \end{cases} \tag{13}$$

This statement is proved by Johnson [6] for $S_s = 0$ and a discrete distribution of demand. Using Theorem 3, Küenle [7] extends Johnson's method for demand having a continuous distribution.

In practice, it is advisable to optimize in the class of $(\sigma, S)$-plans with simple regions $\sigma$, characterized by a few parameters (cf. [11], [3], [10]).

## References


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

[2] H.-J. Girlich, *Diskrete stochastische Entscheidungsprozesse*, B. G. Teubner Verlagsgesellschaft, Leipzig 1973.

[3] —, *Zur Theorie stochastischer Lagerhaltungsmodelle*, Diss. (B), University of Leipzig, 1974.

[4] J. M. Harrison, *Countable state discounted Markovian decision processes with unbounded rewards*, Techn. Report No. 17, Dep. Operations Research, Stanford University, 1970.


(2) If $S_s = 0$, then we have Strauch's negative case (cf. [13]).




[5] —, *Discrete dynamic programming with unbounded rewards*, Ann. Math. Statist. 43 (1972), pp. 636–644.

[6] E. L. Johnson, *Optimality and computation of* $(\sigma, S)$ *policies in the multi-item infinite horizon inventory problem*, Manag. Scie. 13 (1967), pp. 475–491.

[7] H.-U. Küenle, *Markovsche Entscheidungsmodelle mit allgemeinem Zustands- und Aktionsraum und ihre Anwendung in der Lagerhaltungstheorie*, Diss. (A), Depart. of Math., University of Leipzig, 1976.

[8] S. Lippman, *Semi-Markov decision processes with unbounded rewards*, Management Sci. 19 (1973), pp. 717–731.

[9] —, *On dynamic programming with unbounded rewards*, ibid. 21 (1975), pp. 1225–1233.

[10] M. Miethe, *Untersuchung von* $(\sigma, S)$*-Prozessen in der Lagerhaltungstheorie*, Diss. (A), Depart. of Math., University of Leipzig, 1976.

[11] B. D. Sivazlian, *Stationary analysis of a multicommodity inventory system with interacting set-up costs*, SIAM J. Appl. Math. 20 (1971), pp. 264–278.

[12] S. Stidham and N. U. Prabhu, *Optimal control of queueing systems*, In: *Mathematical Methods in Queueing Theory*, Lecture Notes in Economics and Math. Systems, Vol. 98, Springer-Verlag, 1974.

[13] R. E. Strauch, *Negative dynamic programming*, Ann. Math. Statist. 37 (1966), pp. 871–890.

[14] J. van Nunen, *Contracting Markov decision processes*, Diss., Technological University Eindhoven, 1976.

[15] J. van Nunen and J. Wessels, *A note on dynamic programming with unbounded rewards*, Technological University Eindhoven, Memorandum COSOR 75–13, 1975.

[16] J. Wessels, *Markov programming by successive approximations with respect to weighted supremum norms*, Technological University Eindhoven, Memorandum COSOR 74–13, 1974


*Presented to the semester*
*MATHEMATICAL STATISTICS*
*September 15–December 18, 1976*