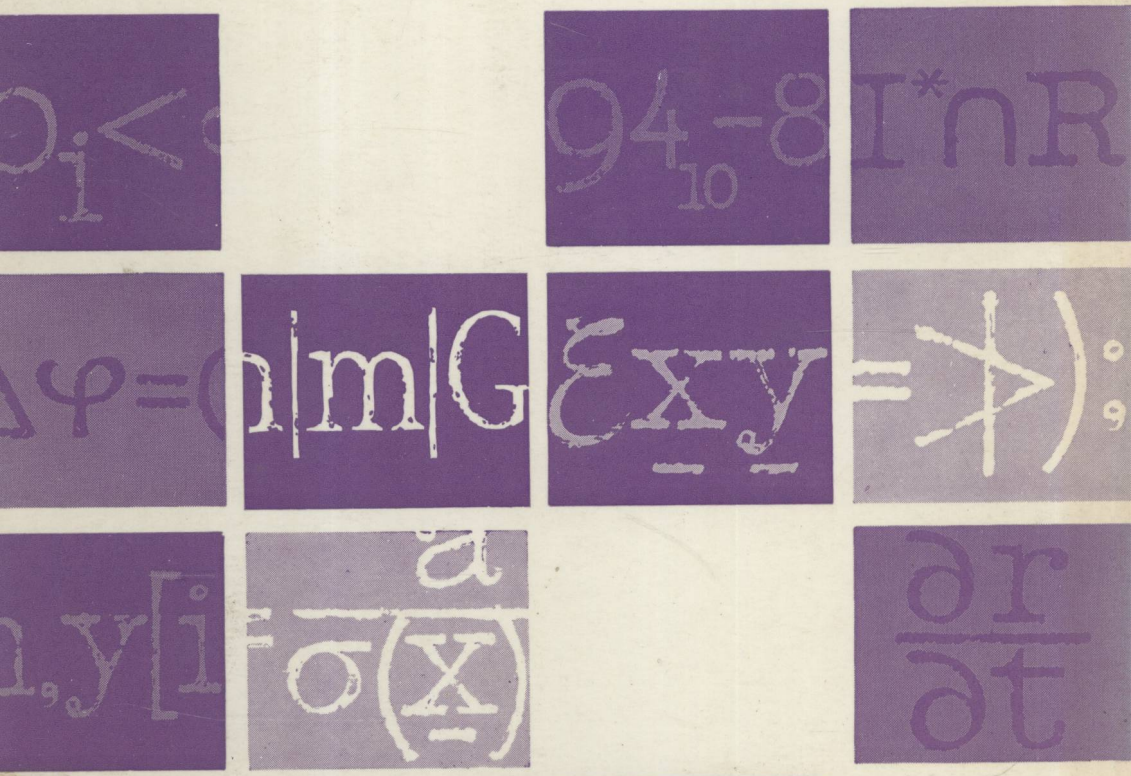
DYNAMIC PROGRAMMING AND MARKOV POTENTIAL THEORY

A. HORDIJK



HEMATICAL CENTRE TRACTS

MC

51

MATHEMATICAL CENTRE TRACTS 51

A.HORDIJK

DYNAMIC PROGRAMMING AND MARKOV POTENTIAL THEORY

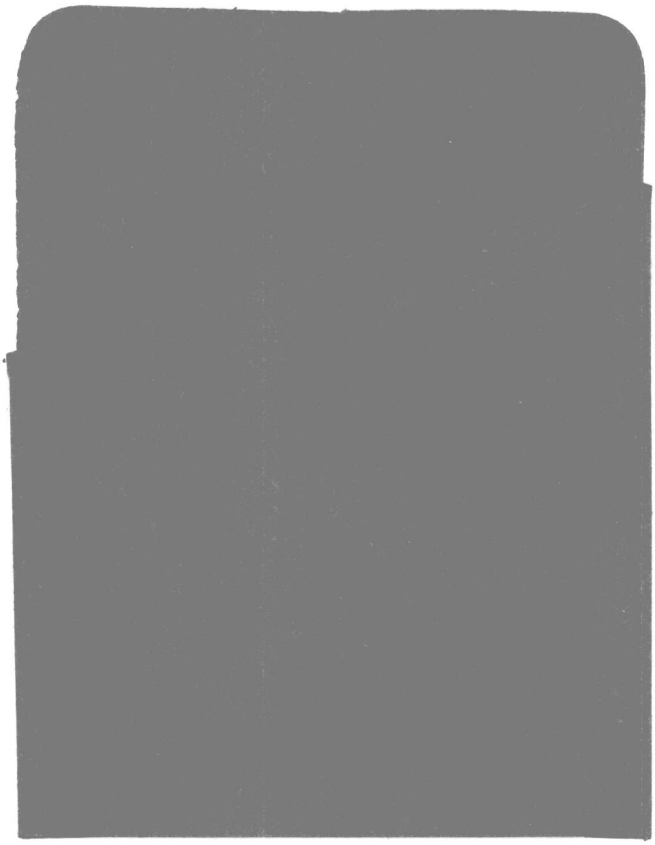
MATHEMATISCH CENTRUM AMSTERDAM 1974

AMS (MOS) subject classification scheme (1970): 60J05, 60J45, 62L15, 90C40

ISBN 90 6196 095 9

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.



Note to Mathematical Centre Tracts 51

The monograph *Dynamic programming and Markov potential theory* is an unaltered printing of the author's thesis.

At the beginning of 1975 a list of errata and addenda will be available.

A.H.

ACKNOWLEDGEMENTS

The research leading to this monograph was carried out at the Mathematical Centre. The author is grateful for the splendid opportunity given to him there and wishes to thank in particular Prof.dr. J. Hemelrijk for encouraging him to study statistics and Prof.dr. G. de Leve for introducing him to the subject of Markov programming.

This book owes much to the valuable suggestions of professors dr. J.A. Bather and dr. J.Th. Runnenburg. The stimulating interest of professor Bather and of the members of the 1973-Colloquium on Probability Theory, organized by professor Runnenburg, proved most helpful in doing the hard job of writing things down.

I also wish to thank my colleague Henk Tijms, who shares with me an interest in dynamic programming, for many discussions on this subject.

The author's sincere thanks go to Mr. J. Hillebrand and to Mrs. S.M.T. Hillebrand-Snijders for editing and typing the manuscript, to Mr. K.M. van Hee for proofreading, to Messrs. D. Zwarst, J. Suiker and J. Schipper for the reproduction.

SUMMARY

This monograph contains the material presented in 1973 in the Colloquium on Probability Theory organized jointly by the Mathematical Centre and the Institute for Applications of Mathematics of the University of Amsterdam.

The central theme is the investigation of the existence of optimal policies or optimal strategies in various discrete time dynamic programming problems.

In section 2 some well-known theorems in Markov potential theory are generalized to collections of Markov chains. Most of the definitions and results in this section also play an important role in the sequel.

In sections 3 and 4 a discrete time optimal control problem is investigated. It is proved that the value function is the minimum of the c_p - excessive functions that majorize the reward function. Further it is shown that a strategy is optimal if and only if it is thrifty and equalizing.

Section 5 deals with a semi-Markov decision process having at least one state for which the expected cost until the system enters this state is uniformly bounded over all policies. Using results from the foregoing sections, we obtain a rather general condition guaranteeing the existence of optimal policies with respect to the average return criterion.

In section 6 some theorems on dynamic programming problems with total return criterion are collected.

Using results from section 6, we answer in section 7 some questions raised in connection with the notions introduced in section 2. The section is concluded with a theorem on the existence of optimal strategies for problems with a finite state space.

In section 8 the notions communicating and recurrent system are introduced. Similar to the notions communicating and recurrent class for one Markov chain, they play a basic role in Markov decision processes.

It is proved in section 9 for a wide class of sequential decision problems that the optimal stopping time is exponentially bounded under the optimal policy.

In section 10 we investigate again the discrete time dynamic programming problem with the supremum of the expected return per unit time as optimality criterion. If the invariant probability measures depend continuously on the decision rule or if they form a tight collection and the system is recurrent then there exists a stationary optimal policy.

A simultaneous Doeblincondition is investigated in section 11.

In section 12 it is pointed out that this notion provides the connection between conditions given in the literature and those of the sections 10 and 11.

In section 13 we collect several results announced in the foregoing sections. It is proved there that randomization does not increase the value function. Finally some theorems on the existence of weak and strong nearly optimal policies are given.

CONTENTS

1. INTRODUCTION

In this monograph we are mainly concerned with a dynamic system which at times $t = 0,1,\ldots$ is observed to be in one of a possible number of states. Let E denote the countable space of all possible states. If at time t the system is observed in state i then a decision must be chosen from a given set $P(i)$. The probability that the system moves to a new state j (the so-called transition probability) is a function only of the last observed state i and the subsequently taken decision. In order to avoid an overburdened notation we shall identify the decision to be taken with the probability measure on E that is induced by it. Thus for each $i \in E$ the set $P(i)$ consists of probability measures $p(i,.)$.*) Let $\mathcal{P}$ be the set of all stochastic matrices P with $p(i,.) \in P(i)$ for each $i \in E$. Hence $\mathcal{P}$ has the *product property*: with P_1 and P_2 the set $\mathcal{P}$ also contains all those P with for every $i \in E$ in the i^{th} row of P either the i^{th} row of P_1, or the i^{th} row of P_2.

A policy R for controlling the system is a sequence of decision rules for the times $t = 0,1,\ldots$, where the decision rule for time t is the instruction at time t which prescribes the decision to be taken. This instruction may depend on the history i.e. the states and decisions at times $0,1,\ldots,t-1$ and the state at time t. When the decision rule is independent of the past history except for the present state then it can be identified with a $P \in \mathcal{P}$. A memoryless or Markov policy R is a sequence $P_0,P_1,\ldots \in \mathcal{P}$, where P_t denotes the decision rule at time t. P_t also gives the transition probabilities at time t.

In this monograph there are only a few places where non-memoryless policies are used. We need them to show that the value function is $c_{\mathcal{P}}$-superharmonic (see theorem 3.1). Theorem 13.2 says that when $\mathcal{P}$ contains all randomizations then the supremum over all memoryless policies equals the supremum over all policies. Hence in this case the value function may be defined as the supremum over the memoryless policies.

Since the law of motion of the dynamic system can be described by a non-stationary Markov chain when a memoryless policy is used, we prefer to

---

*) We allow that with positive probability the system "breaks down" or "disappears", so $p(i,j) \geq 0$, $i,j \in E$ and $p(i,E) := \sum_{j \in E} p(i,j) \leq 1$, $i \in E$.

introduce a decision process as a collection of non-stationary Markov chains (for a more general foundation of decision processes see [Hinderer]). A memoryless policy which takes at all times the same decision rule i.e. $P^\infty := (P,P,\ldots)$, $P \in \mathcal{P}$ is called a stationary policy and induces a stationary Markov chain.

One of the features of this monograph is the generalization of well-known results for one Markov chain to a collection of Markov chains. We give some examples. In theorem 8.6 it is proved that the maximal average expected reward does not depend on the initial state given that the system is recurrent. This is a direct generalization of the well-known theorem that each excessive function on a recurrent chain is constant.

The main assumption in theorem 5.1 (relation 5.1.1) is nothing else than a condition guaranteeing that all Markov chains are uniformly positive recurrent. This condition is a direct generalization to a collection of Markov chains of a so-called Foster criterion or a Liapunov function criterion as it is called elsewhere (see subsection 2.7).

Finally the simultaneous Doeblin condition (see section 11) is a straightforward extension to a collection of Markov chains of the well-known Doeblin condition.

Nowadays potential theory for Markov chains is well developed. A systematic treatment of potential theory for dynamic systems would in our opinion be desirable. Although the second part of the title of this monograph suggests more, our contribution to potential theory for dynamic systems consists only in the introduction of some useful terminology and the derivation of some interesting results (sections 2 and 7). The reason is that we were mainly interested in dynamic programming. It seems that many interesting questions were left untouched.

When in state i decision p(i,.) is taken then an immediate cost depending on i and p(i,.) is incurred *). Let $c_p(i)$ be the immediate cost when taking decision p(i,.) (the i^{th} row of matrix P) in state i and write c_p for the vector with i^{th} component $c_p(i)$. Note that if $P,Q \in \mathcal{P}$ with p(i,.) = q(i,.) then $c_p(i) = c_Q(i)$.

The expectation of the cost at time n when starting in state i at time

*) It is common to minimize when speaking of costs. We shall always maximize. The reason is that along with a cost structure also a reward function shall be used (see section 3).

zero and using policy $R = (P_0, P_1, \ldots)$ will be denoted by $\mathbb{E}_{i,R}\, c(\underline{x}_n)$, where $\underline{x}_n$ *) is the state at time n. $\mathbb{E}_R\, c(\underline{x}_n)$ denotes the vector with i^{th} component $\mathbb{E}_{i,R}\, c(\underline{x}_n)$ (for stationary policy P^∞ we write $\mathbb{E}_P[\ldots]$ instead of $\mathbb{E}_{P^\infty}[\ldots]$). It is easily seen that

$$\mathbb{E}_R\, c(\underline{x}_n) = P_0\, P_1 \ldots P_{n-1}\, c_{P_n}.$$

In some of the following sections it is assumed that the cost function is a charge structure (see definition 2.12). In dynamic programming a weaker assumption like "all relevant expectations do not attain the value plus infinity" could be used. Our gain is a greater simplicity in the statements of the results. Also a nice implication is that the well-known theorem in optimal stopping remains valid: the value function is the minimum of the excessive functions that majorize the reward function.

The basic reason for taking the state space a countable set was that many of the problems which arise in general state spaces already appear in the countable state space. The countable state space does not have the "compactness" properties of the finite state space and with the countable state space one avoids the "measurability" questions of more general state spaces. As to the generalization of the results of this monograph, some can be generalized in a straightforward way, some results cannot be generalized and for the other results we do not know.

In an important part of the literature on Markovian decision processes it is assumed that for each state the set of available decisions in that state is a finite set. Usually randomized decisions i.e. convex combinations of the available decisions with a corresponding convex combination of the costs as the immediate cost, are allowed. We prefer to start with general sets of decisions $P(i)$, $i \in E$, which may contain all randomizations. As long as there are no constraints introduced the distinction between randomized and non-randomized decisions is in our opinion not very important (cf. section 13).

In several places we need a notion of convergence on P. A sequence

*) Random variables are underlined.

P_n, $n = 1,2,\ldots$ is convergent to P if $\lim_{n\to\infty} p_n(i,j) = p(i,j)$ for all i and j. In this case, we shall say that $\lim_{n\to\infty} P_n = P$. $\mathcal{P}$ with this topology is a metric space (see section 13).

The identification of the set of actions with the set of probability measures and several notations are adopted from [Bather].

The number of papers on dynamic programming is overwhelming. Only those books or papers referred to in this monograph, or those that proved important for the author's study of these topics are included in the bibliography.

It is difficult to provide a readable and consequent notation for the topics studied. The list of notations may be helpful to overcome possible notational shortcomings.

2. POTENTIALS AND EXCESSIVE FUNCTIONS

The aim of this section is twofold. First to generalize some well-known theorems in Markov potential theory (theorems 2.9 and 2.20 to 2.23). The second intention of this section is to introduce notions which, in our opinion, are basic in the study of discrete time dynamic programming problems. Further we collect in this section definitions and results which play an important role throughout this monograph.

Each function used in this monograph is assumed to be a finite and real valued function. Moreover when writing $\mathbb{E}_P f(\underline{x}_n)$ or $P^n f$ it is tacitly assumed that

$$\sum_j p^n(i,j)|f(j)| < \infty \text{ for all } i \in E.$$

2.1. DEFINITION. *Function* w *is a charge with respect to* P *if*

$$\mathbb{E}_P \sum_{n=0}^{\infty} |w(\underline{x}_n)| = \sum_{n=0}^{\infty} P^n|w| < \infty.$$

2.2. DEFINITION. *Function* f *is a potential w.r.t.* P *if there exists a charge* w *w.r.t.* P *such that*

$$f = \sum_{n=0}^{\infty} P^n w.$$

So function w is called a charge if the sum $\sum_{n=0}^{\infty} P^n w$ is well-defined. This sum is then a potential.

2.3. DEFINITION. *Function* f *is a*

$$\begin{array}{lll} c\text{ - }\textit{super} & & \geq \\ c\text{ - } & \textit{harmonic function w.r.t. P if} & f = c + Pf. \\ c\text{ - }\textit{sub} & & \leq \end{array}$$

2.4. DEFINITION. *Function* f *is a* c*-excessive function w.r.t.* P *if*

(2.4.1) c *is a charge w.r.t.* P

(2.4.2) $\sum_{n=0}^{\infty} P^n c \leq f$

(2.4.3) $c + Pf \leq f$.

So a c-superharmonic function with c a charge satisfying relation (2.4.2)

is a c-excessive function. To see that c-excessive functions form an interesting class one should realize that when f is the value function of a stopping problem for a Markov chain with matrix of transition probabilities P and "cost" function c then relations (2.4.2) and (2.4.3) are fulfilled. This can be seen by noting that the left-hand side of (2.4.2) denotes the "return" in case we will never stop which is less than the value function. The left-hand side of (2.4.3) denotes the "return" if we wait one period and then continue in an optimal way. This may be a sub-optimal policy.

2.5. THEOREM. *Function* f *is a potential w.r.t.* P *iff* $w_P := f-Pf$ *is a charge w.r.t.* P *and* $\lim_{n\to\infty} P^n f = 0$.

PROOF. Suppose w is a charge such that $f = \sum_{n=0}^{\infty} P^n w$. Then by interchanging the order of summation (w is a charge) it follows that

$$f-Pf = \sum_{n=0}^{\infty} (P^n w - P^{n+1} w) = w.$$

Hence $w_P = w$ and consequently w_P is a charge. By iterating the equality

$$w_P + Pf = f$$

N times we find the equality

$$(2.5.1) \qquad w_P + Pw_P + \ldots + P^N w_P + P^{N+1} f = f.$$

Since $f = \sum_{n=0}^{\infty} P^n w_P$, it then follows that $\lim_{n\to\infty} P^n f = 0$.*)

To show the converse, we note that $\sum_{n=0}^{\infty} P^n w_P$ is a potential since w_P is a charge. Moreover, it follows from (2.5.1) and $\lim_{n\to\infty} P^n f = 0$ that this potential equals f. □

It can be seen from the above proof that a potential uniquely determines its charge (if f is a potential then f-Pf is its charge).

---

*) For f_n, n=1,2,... a sequence of functions, we write $\lim_{n\to\infty} f_n = 0$ if $\lim_{n\to\infty} f_n(i) = 0$ for all $i \in E$.

2.6. THEOREM. *If to* $c \geq 0$ *there exists a nonnegative* c*-superharmonic function* v *w.r.t.* P *then* c *is a charge w.r.t.* P *and* $\sum_{n=0}^{\infty} P^n c \leq v$.

PROOF. The definition of a c-superharmonic function gives

$$c + Pv \leq v.$$

By iterating this inequality N times we find

$$c + Pc + \ldots + P^N c + P^{N+1} v \leq v.$$

Since $v \geq 0$ it follows then

$$\sum_{n=0}^{\infty} P^n c \leq v < \infty$$

and consequently c is a charge. □

As an illustration of theorem 2.6 we shall prove that relation (2.7.1) is sufficient for a Markov chain to be positive recurrent. In this way we recover the condition for positive recurrence as can be found in [Foster, theorem 2]. For a countable state space a condition similar to (2.7.1) can be found in [Kushner, theorem 8.6.5.7, p. 211]. There the condition is called a Liapunov function criterion.

2.7. FOSTER CRITERION - LIAPUNOV FUNCTION CRITERION

The Markov chain with transition matrix P *is positive recurrent if there exists a state* i_0 *and a nonnegative solution* y *of the inequalities*

(2.7.1) $\qquad e + \tilde{P}y \leq y,$

where e *is defined by* $e(i) = 1$ *for all* i *and* $\tilde{P}$ *is the column-restriction of* P *to* $E\backslash\{i_0\}$ *i.e.*

$$\tilde{p}(i,j) := \begin{cases} 0 & \text{for } j = i_0 \\ p(i,j) & \text{for } j \neq i_0. \end{cases}$$

PROOF. Let $\underline{\tau}$ denote the reentry time of $\{i_0\}$, i.e. $\underline{\tau}$ is the least $n > 0$ if any with $\underline{x}_n = i_0$, and $\underline{\tau} = \infty$ if there is no such n. Then it is an easy check that

$$(2.7.2) \qquad \mathbb{P}_i[\underline{\tau} > n] = \tilde{P}^n e(i).$$

According to a well-known lemma

$$(2.7.3) \qquad \mathbb{E}_i[\underline{\tau}] = \sum_{n=0}^{\infty} \mathbb{P}_i[\underline{\tau} > n].$$

By (2.7.2) and (2.7.3) we have

$$(2.7.4) \qquad \mathbb{E}_i[\underline{\tau}] = \sum_{n=0}^{\infty} \tilde{P}^n e(i).$$

The Markov chain is a positive recurrent class ([Chung, p. 31]) if

$$(2.7.5) \qquad \mathbb{E}_i[\underline{\tau}] < \infty \text{ for all } i \in E.$$

To prove this it is by (2.7.4) sufficient to show that $\sum_{n=0}^{\infty} \tilde{P}^n e < \infty$ (i.e. all components are finite). Now theorem 2.6 says that relation (2.7.1) implies that e is a charge w.r.t. $\tilde{P}$. □

A Liapunov function criterion for the existence of an invariant probability measure in the case of a Markov process with a metric state space is given in [Hordijk and Van Goethem].

2.8. THEOREM. *If there exists a* c*-superharmonic function* f *w.r.t.* P, *for* c *a majorant of a charge then*

a. $h := \lim_{n\to\infty} P^n f$ *exists and* $-\infty \leq h(i) < \infty$ *for all* $i \in E$

b. *if* $h(i) > -\infty$ *for all* $i \in E$ *then* c *is a charge w.r.t.* P

c. *if* $h \geq 0$ [*)] *then* f *is* c*-excessive w.r.t.* P.

*) We write $x \geq y$ if $x(i) \geq y(i)$ for all i and denote 0 for the vector with each component equal to 0.