A THEORY FOR SEMI-MARKOV DECISION PROCESSES WITH UNBOUNDED COSTS AND ITS APPLICATION TO THE OPTIMAL CONTROL OF QUEUEING SYSTEMS

BY

PETER ORKENYI

TECHNICAL REPORT NO. 64

AUGUST 1976

PREPARED UNDER CONTRACT

NC0014-76-C-0418 (NR-047-061)

FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF OPERATIONS RESEARCH

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

A THEORY FOR SEMI-MARKOV DECISION PROCESSES WITH UNBOUNDED COSTS, AND ITS APPLICATION TO THE OPTIMAL CONTROL OF QUEUEING SYSTEMS

by

PETER ORKENYI

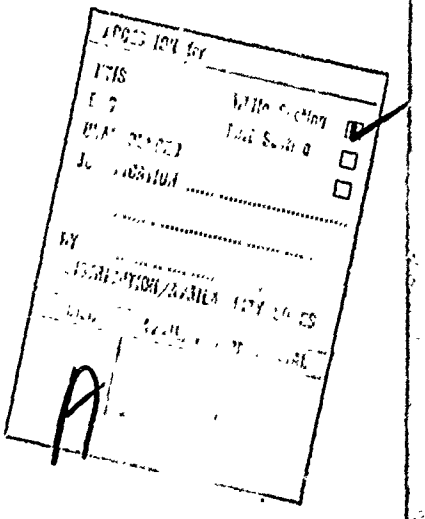
TECHNICAL REPORT NO. 64
AUGUST 1976

PREPARED UNDER CONTRACT
N00014-76-C-0418 (NR-047-061)
FOR THE OFFICE OF NAVAL RESEARCH

Frederick S. Hillier, Project Director

This research was supported in part by
NATIONAL SCIENCE FOUNDATION GRANT ENG 75-14847

DEPARTMENT OF OPERATIONS RESEARCH
STANFORD UNIVERSITY
STANFORD, CALIFORNIA



CHAPTER 1

INTRODUCTION

Markov and semi-Markov decision processes have been studied extensively since their initial development in the late 1950's and early 1960's. They provide the natural framework for the study of a plethora of problems arising in the areas of queueing, inventory, maintenance and replacement, etc. Many useful results about Markov and semi-Markov decision processes are available now under a variety of assumptions. A common assumption has been the assumption of bounded costs. Although bounded costs is an appropriate assumption for many problems, there are also many situations, especially in the context of queueing and inventory, for which it is not appropriate. Thus, there is a need for developing a theory for Markov and semi-Markov decision processes with unbounded costs. Although there have been some efforts in this direction earlier, stronger results need to be developed. That is the objective of this report. Specifically, results are obtained for semi-Markov decision processes both when the costs are discounted and when they are not. Application to the optimal control of queueing systems is also considered.

The terminology of semi-Markov decision processes is summarized in Section 1. Section 2 then presents some examples of semi-Markov decision processes both with and without unbounded costs. Section 3 reviews the literature on semi-Markov decision processes. An overview of the study is presented in Section 4.

1. Terminology of Semi-Markov Decision Processes.

The semi-Markov decision process is a stochastic process which requires certain decisions to be made at certain points in time. These points in time are the decision epochs. At each decision epoch, the system under consideration is observed and found to be in a certain state. The set of all conceivable states is the state space. The decision consists of choosing an action from a set of permissible actions. This set depends on the state of the system when the decision has to be made. The set of permissible actions for a given state is an action space. The union of all action spaces is referred to as the action space. Once an action has been chosen, the probabilistic aspects of the evolution of the system until the next decision epoch occurs (including the time elapsed and the state of the system at the next decision epoch) is completely determined by the state of the system when the action was chosen and the action itself.

A policy for a semi-Markov decision process is a rule which selects an action at each decision epoch by considering only the history of the process up to that point in time. An interesting class of policies is the class of stationary policies. A stationary policy selects the action at each decision epoch solely on the basis of the state of the system at the decision epoch. A stationary policy is deterministic if it selects the actions according to a fixed mapping from the state space into the action space; otherwise it is randomized.

A part of the process is the costs incurred. The objective is to minimize these costs. They are, however, incurred in a random fashion and at different times, so a further specification of the objective is needed. There are several alternatives. If the time factor is not

important, one may choose to minimize the total expected cost, or if this is not finite, the long-run expected average cost. If the time factor is important, one may discount the costs and minimize the total expected discounted cost.

For our purposes, a semi-Markov decision process is completely specified by four objects, the state space, S, the action spaces $\{A_s\}_{s\in S}$, the law of motion q, and the cost function c. Let $A = \bigcup_{s\in S} A_s$, and let R be the set of real numbers. The law of motion, q, is a mapping from $S \times A \times S \times R$ into R, and the cost function, c, is a mapping from $S \times A \times R$ into R. Consider a decision epoch. Suppose the state there is s and suppose the action chosen there is a. Then, for $s' \in S$ and $t \in R$, $q(s,a,s',t)$ is the joint probability that the time until the next decision epoch is less than or equal to t and that the state at the next decision epoch is s'. If the times between the decision epochs are constant, then we have a Markov decision process. Also, for $t \in R$, $c(s,a,t)$ is the expected cost accumulated until time t. The formulation of a problem in the framework of semi-Markov decision processes consists of specifying S, $\{A_s\}_{s\in S}$, q and c. Some examples of semi-Markov decision processes are now presented.

2. Examples of Semi-Markov Decision Processes With and .thout Unbounded Costs.

Selling an asset (Ross (1970)):

Consider a person who wants to sell his house. Offers arrive according to a stationary Poisson process. The sizes of the offers are independent, identically distributed random variables. When an offer arrives, it

must either be accepted or rejected. Rejected offers are lost. A maintenance cost is incurred at a constant non-negative rate until the house is sold. The problem is to decide when an offer should be accepted. This problem can be formulated within the framework of a semi-Markov decision process as follows.

Let the decision epochs be the same as the epochs when offers arrive, let the actions be to accept or reject the current offer, and let the state of the system be the size of the offer at the most recent decision epoch.

A job shop model (Lippman and Ross (1968)):

Consider a factory which is only able to handle one job at a time. Jobs arrive according to a stationary Poisson process. When a job arrives it is classified to be of a certain type. Jobs of the same type have an identical probabilistic structure for their cost and completion time. The classification of arriving jobs are independent, identically distributed random variables. Each job must either be accepted or rejected. Jobs arriving when the factory is busy are rejected automatically. The problem is to determine when a job should be accepted (rejected) when the factory is not busy. This problem can be formulated within the framework of semi-Markov decision processes as follows.

Let the decision epochs be the same as the epochs of job arrivals (neglect jobs which arrive when the factory is busy), let the available actions be to accept or reject the job that just arrived, and let the state of the system be the type of job present.

The M/G/1 queueing system with removable server (Heyman (1968)):

Consider a queueing system having one server which can be turned

on and off. Customers arrive according to a stationary Poisson process. They are served one by one on a first-come-first-served basis. The service times are independent, identically distributed random variables. There is a cost associated with the service of each customer. These costs are independent, identically distributed random variables. There are fixed charges for turning the server on and off. There is a cost for having the server on when there are no customers in the system. This cost is incurred at a constant rate at such times. Finally, there is a cost for holding customers in the system. This cost is incurred at a rate which is a non-negative, non-decreasing function of the number of customers present. The problem is to determine when the server should be turned on and turned off. This problem can be formulated within the framework of semi-Markov decision processes as follows.

Let the decision epochs be the epochs of customer arrivals and departures (neglect arrivals which occur when the server is busy). Let the available actions be to turn the server off (or have him off) and to turn him on (or have him on). Finally, let the state of the system be a vector whose first component gives the number of customers present, and whose second component shows the status of the server.

3. A Brief Survey of the Literature on Semi-Markov Decision Processes.

The first comprehensive study of Markov decision processes was done by Howard (1960). He assumed finite state and action spaces, and considered the problem both with and without discounting. He only considered stationary policies, and developed his now well-known policy improvement procedures. He proved that they would produce optimal stationary policies.

At the same time, Manne (1960) suggested solving the Markov decision

problem by using linear programming. He used the average cost criterion, and showed how to solve an inventory problem by his suggested approach. The first linear programming formulation for the problem with discounting was given by d'Epenoux (1960). Shortly afterwards, Wolfe and Dantzig (1962) proposed the use of their decomposition technique on Manne's linear programming formulation.

Blackwell (1962) considered Markov decision processes with finite state and action spaces, and proved that there is a stationary policy which is optimal among all Markov policies. He also considered the problem for arbitrarily small interest rates, and proved that there is a stationary policy which is optimal among all Markov policies for small enough interest rates. Later, Blackwell (1965) considered Markov decision processes with more general state and action spaces. He only assumed that they were Borel sets. However, he assumed that the rewards were uniformly bounded. He considered the problem with discounting, and allowed any measurable policy. His main results were the following. There is a (p,ϵ)-optimal stationary policy. If the action spaces are countable, then there is an ϵ-optimal stationary policy. If the action spaces are finite, then there is an optimal stationary policy. If there is an optimal policy, then there is one which is stationary.

Strauch (1966) considered the same problem as Blackwell, but instead of using discounting, he assumed that the rewards were negative. His main results were similar to those of Blackwell. If the action spaces are finite, then there is an optimal policy. If there is an optimal policy, then there is one which is stationary. The optimal return function is measurable and satisfies the optimality equation.

Denardo (1967) also considered the same problem as Blackwell and generalized it to include certain stochastic games. He introduced operators with certain monotonicity and contraction properties, and used the Picard-Banach fixed point theorem to prove that the functional equation of optimality has a unique solution, which is the optimal reward function.

Veinott (1966) gave a policy iteration procedure for finding a bias-optimal policy (no discounting). Later, Veinott (1969) considered a more refined optimality criterion, namely, that of finding a policy which is optimal for all sufficiently small interest rates (sensitive discount optimality). He developed a policy iteration procedure for finding a stationary policy which would be optimal according to this criterion.

Derman (1966) considered Markov decision processes with finite action spaces and a countable state space. He used the average cost criterion, and gave conditions for when a stationary, deterministic policy is optimal. Ross (1968) considered the same problem, but allowed a general state space. He derived results similar to those of Derman. He also suggested a method for converting the average cost problem to a discounted cost problem.

One of the first to consider semi-Markov processes was Pyke (1961). Shortly afterwards, Howard's results for Markov decision processes were extended to semi-Markov decision processes independently by Jewell (1963) and Howard (1964). When they considered the average cost criterion, they assumed that all states belong to one positive recurrent class. They also gave linear programming formulations.

Denardo and Fox (1968) considered the multi-chain case (i.e., the case of several positive recurrent classes), using the average cost criterion. They gave a linear programming formulation and a policy

improvement procedure. Later Denardo (1970a) developed a solution method which used Manne's linear programming formulation to solve a sequence of subproblems. This solution method has the advantage that several small linear programming problems are solved instead of one big one. Denardo (1971) also considered the problem when small interest rates are used. His results are similar to those of Veinott for the discrete-time Markov decision process. He gives a sequence of linear programming problems for finding an optimal policy.

All of these authors have assumed that the immediate rewards or costs are bounded uniformly. After Strauch, Harrison (1972) was the first one to relax the condition of bounded costs. He assumed that the expected absolute reward in one period minus the expected absolute reward in the period before it, given the state at the beginning of that period, is uniformly bounded. He then showed that the expected discounted reward is finite for each policy and that there exists a stationary policy which is optimal. He proved this by using the Picard-Banach fixed point theorem. He also extended his results from Markov decision processes to semi-Markov decision processes.

The problem with unbounded costs was also considered by Reed (1973). He investigated the problem both with and without discounting. He assumed finite action spaces and countable state space. He gave sufficient conditions for a stationary policy to be optimal.

Hordijk (1974a), (1974b) also considered the problem with unbounded costs. He introduced the notion of convergent dynamic programming, which is just to say that the expectation of the sum of the absolute rewards is finite. He proved that a policy is optimal if it is unimprovable and if another condition is satisfied.

Most recently, Lippman (1973), (1975a) considered the problem with unbounded costs. His approach is to use a norm such that the norm of the costs is finite even though the costs are unbounded. In order to obtain the usual results, he then has to make assumptions about the law of motion of the system. By doing that, he showed that Denardo's N-stage contraction assumption is satisfied, and the results follow.

4. Overview of the Study.

The emphasis of this report is on determining necessary and sufficient conditions for a stationary policy to be optimal. It is not assumed that the costs are bounded. The problem is considered both with and without discounting.

Chapter 2 treats the problem without discounting. Two closely related optimality criteria are used, namely, the average cost criterion and the undiscounted cost criterion. After introducing the important concept of an unimprovable policy, sufficient conditions are given for an unimprovable policy to be optimal. Both the special case where the optimal expected average cost is independent of the start-state and the general case when the average cost is not necessarily constant are considered.

Chapter 3 treats the problem with discounting. After formulating the problem and introducing the operators Q_π and T_π, the optimality equation is proven. The existence of stationary optimal and stationary ϵ-optimal policies are then investigated. Policy improvement is considered, and some necessary and sufficient conditions for optimality are given.

Chapter 4 is devoted to the optimal control of queueing systems.

Solution methods are explored, and four different ways of solving the problem of unbounded costs are presented.

Some general notation and conventions are best introduced here. R denotes the set of real numbers, R_+ denotes the set of non-negative real numbers, N denotes the set of natural numbers (starting with one) and N_0 denotes the non-negative integers. The Kroenecker delta function δ is defined by

$$\delta(x,y) = \begin{cases} 1 & \text{if } x = y \ , \\ 0 & \text{if } x \neq y \ . \end{cases}$$

If x is a real number, then x^+ is $\max(0,x)$ and x^- is $\max(0,-x)$. Finally, we use the convention that

$$x + y = \begin{cases} \infty & \text{if } x = \infty,\ y > -\infty \ , \\ -\infty & \text{if } x < \infty,\ y = -\infty \ , \\ \text{undefined} & \text{if } x = -y = \pm\infty \ . \end{cases}$$

CHAPTER 2

SEMI-MARKOV DECISION PROCESSES WITHOUT DISCOUNTING

This chapter presents an investigation of semi-Markov decision processes without discounting the costs. Thus, costs of equal size incurred at different times count the same. Two optimality criteria are used. The first one is the average cost criterion, according to which a policy is optimal if the long-run expected average cost is minimized by this policy. This criterion has been considered recently by Hordijk (1974a). The other criterion is the undiscounted cost criterion. A policy is optimal under this criterion if it minimizes the long-run (total) expected cost for the process which is derived from the original one by incurring an additional cost at a rate equal to the negative of the minimum average cost. This criterion has been considered by Denardo (1970). He called a policy which is optimal for this criterion a bias-optimal policy.

There have traditionally been two approaches to the problem without discounting. The first one consists of restricting one's consideration to stationary (deterministic) policies and performing a stationary analysis. The second one consists of considering the problem with discounting and observing what happens when the interest rate goes to zero. Here, we will follow the first approach. It has been common to assume that the costs are uniformly bounded. We make no assumptions about the size of the costs. Reed (1973) conducted a similar but somewhat less complete study of the problem.

In Section 1, there is a formal statement of the problem to be considered. It also contains some preliminary results. Unimprovable policies are defined there. In Section 2, sufficient conditions for an unimprovable policy to be optimal are given. It is assumed that the long run expected average cost is constant. In Section 3, the results from Section 2 are extended to cover the general case of non-constant long-run expected average cost. In Section 4, there is a brief discussion of methods for finding an optimal policy.

1. Problem Formulation.

As before, let S be the state space, $\{A_s\}_{s\in S}$ be the action spaces, q be the law of motion and c the cost function. Let $\mathcal{D}$ be the set of stationary, deterministic policies, and let A be $\bigcup_{s\in S} A_s$. For each $n \in N$, let t_n, s_n, and a_n denote the time of the n^{th} decision epoch, the state observed there, and the action chosen there, respectively.

For each $\pi \in \mathcal{D}$, let v_π be the mapping from $S \times R^+$ into R such that, for each $s \in S$ and $t \in R^+$,

$$v_\pi(s,t) = E_{\pi,s}\{\sum_{n\in N_t} c(s_n,a_n,t-t_n)\} ,$$

where

$$N_t = \{n \in N | t_n \leq t\} .$$

E is the expectation operator, and the subscripts π and s respectively denote that the start-state is s and that the policy used is π. In words, $v_\pi(s,t)$ is the expected cost incurred until time t, given that the start-state is s and that the policy π is used.

v_π need not always be well-defined. Later, however, certain assumptions which guarantee the existence of v_π for each $\pi \in \mathcal{D}$ will be made.

The analysis here is based on the fact that under certain conditions (to be introduced when needed), $v_\pi(t,s)$ has a linear asymtote for each $s \in S$ and $\pi \in \mathcal{D}$. For each $\pi \in \mathcal{D}$, let φ_π and w_π be the mappings from S into R such that

$$\varphi_\pi(s) = \lim_{t \to \infty} v_\pi(s,t)/t \ ,$$

$$w_\pi(s) = \lim_{t \to \infty} \{v_\pi(s,t) - t \cdot \varphi_\pi(s)\} \ ,$$

for $s \in S$. φ_π is the long-run expected average cost, given that the start state is s and that the policy π is used. $w_\pi(s)$ is the long-run expected cost not accounted for by $\varphi_\pi(s)$.

Two optimality criteria will be used. The first one is the <u>average cost criterion</u>. A policy $\pi^* \in \mathcal{D}$ is optimal according to this criterion if $\varphi_{\pi^*}(s) \leq \varphi_\pi(s)$ for $s \in S$ and $\pi \in \mathcal{D}$, and the policy is called <u>average optimal</u>. The second criterion is the <u>undiscounted cost criterion</u>. A policy $\pi^* \in \mathcal{D}$ is optimal according to this criterion if it is average optimal and, in addition, $w_{\pi^*}(s) \leq w_\pi(s)$ for $s \in S$ such that $\varphi_{\pi^*}(s) = \varphi_\pi(s)$ for $\pi \in \mathcal{D}$. A policy which is optimal in this sense is called <u>undiscounted optimal</u>. This latter criterion has not received much attention in the literature. This may be due to the fact that often there is not much to gain by using this criterion instead of the average cost criterion. The main difference resulting from the use of these criteria is that the action in the transient states become more important when the undiscounted cost criterion is used. To illustrate this point further, an example is included below.

Example: Consider the following simple semi-Markov decision process. The state space is N and the action spaces are $(0,1)$. The times between the decision epochs are exponentially distributed with the same parameter. State 0 is an absorbing state. Consider states in N. If action 0 is taken, the state 0 is entered next with probability one. If action 1 is taken, the state numbered 1 higher is entered next with probability one. The cost structure is simple. Each time a state in N is reached, an immediate cost of 2 units is incurred, and each time the state 0 is entered, an immediate cost of 1 unit is incurred. Any policy which chooses action 0 in all the states above a given number is average optimal. The undiscounted optimal policy is the one which always chooses action 1. This is clearly the desired policy.

One special reason for using the undiscounted cost criterion is as follows. Under certain circumstances there may exist a sequence of average optimal policies $\pi_1, \pi_2, \ldots$ such that using π_1 for the first decision, π_2 for the second, π_3 for the third, and so on, leads to a long-run expected average cost which is higher than the optimal one. This can easily be seen from the example above. First let π_n be the policy which chooses action 1 for states numbered less than n and action 0 for states numbered n or higher. Each π_n is average optimal. But using π_n at the n^{th} decision epoch for $n = 1, 2, \ldots$, leads to a long-run expected average cost twice as high as the optimal one. Notice that since there is a unique undiscounted optimal policy, this situation cannot occur when the undiscounted cost criterion is used. In general, there is no guarantee for the existence of a unique undiscounted optimal policy, but often a unique undiscounted optimal policy does exist

and thus the undesirable situation mentioned above can be avoided by using the more refined criterion. Some useful semi-Markov process terminology will now be introduced.

A state is called transient if with probability one it will not be reentered after some time. A state is called recurrent if with probability one it will always be reentered. A recurrent state is positive recurrent if the expected time between consecutive visits of this state is finite. Otherwise, it is called negative recurrent. If there is a positive probability that a state is reached in a finite time from another state and vice versa, then the two states are said to communicate. The positive recurrent states belong to one or more positive recurrent classes of states. Each positive recurrent class is a set of positive recurrent states which communicate with each other, but not with states outside the class. We make the following assumptions.

Assumption 1: There is an $\epsilon > 0$ such that

$$q(s,a,s',\epsilon) = 0, \quad \text{for} \quad s \in S,\ a \in A_s,\ s' \in S .$$

In words, the time between two consecutive decision epochs is at least ϵ.

Assumption 2: For each $\pi \in \mathcal{D}$ and $s \in S$, the expected cost incurred and the expected time elapsed from time t until the first decision epoch after (or at) time t divided by the time t have zero as their limits as t tends to infinity, given the start-state s and policy π.

Faced with a particular semi-Markov decision process, one may have difficulties in showing that it satisfies the above assumption. However,

we have not been able to do without them. If the semi-Markov decision process is a Markov decision process, then the second assumption is trivially satisfied.

Some convenient notation will now be introduced. For each $\pi \in \mathfrak{D}$, let q_π and τ_π be the mappings from $S \times S$ into R such that

$$q_\pi(s,s') = \lim_{t \to \infty} q(s,a_\pi(s),s',t) ,$$

$$\tau_\pi(s,s') = \int_{t \in R_+} t\,dq(s,a_\pi(s),s',t) ,$$

for $s,s' \in S$. $a_\pi(s)$ is the action chosen by π in the state s. For each $\pi \in \mathfrak{D}$, also let ν_π and c_π be the mappings from S into R such that

$$\nu_\pi(s) = \sum_{s' \in S} \tau_\pi(s,s') ,$$

$$c_\pi(s) = \lim_{t \to \infty} c(s,a_\pi(s),t) ,$$

for $s \in S$. $q_\pi(s,s')$ is the probability that the next state will be s, given the present state s and policy π. $\tau_\pi(s,s')$ is $q_\pi(s,s')$ multiplied by the expected time until the next decision epoch, given that the next state is s'. $\tau_\pi(s)$ is the expected time until the next decision epoch, given the present state s and policy π. $c_\pi(s)$ is the expected cost until the next decision epoch, given the present state s and policy π. Naturally, we assume that all these quantities exist and are finite.

If the state space is finite, it can easily be shown that φ_π and w_π satisfy the following equations,

$$\varphi_{\pi}(s) = \sum_{s' \in S} q_{\pi}(s,s') \cdot \varphi_{\pi}(s') \ ,$$

$$w_{\pi}(s) = c_{\pi}(s) - \sum_{s' \in S} \tau_{\pi}(s,s') \cdot \varphi_{\pi}(s') + \sum_{s' \in S} q_{\pi}(s,s') \cdot w_{\pi}(s') \ ,$$

for $s \in S$ and $\pi \in \mathcal{D}$ (see Denardo and Fox (1968)). The expressions on the right-hand side are obtained by conditioning on the time of the second decision epoch and the state at that epoch. If $\pi, \pi' \in \mathcal{D}$ and $\pi'' \in P$ are such that π'' uses π' at the first decision epoch and π thereafter, then

$$\varphi_{\pi''}(s) = \sum_{s' \in S} q_{\pi'}(s,s') \cdot \varphi_{\pi}(s') \ ,$$

$$w_{\pi''}(s) = c_{\pi'}(s) - \sum_{s' \in S} \tau_{\pi'}(s,s') \cdot \varphi_{\pi}(s') + \sum_{s' \in S} q_{\pi'}(s,s') \cdot w_{\pi}(s') \ ,$$

for $s \in S$. If $\varphi_{\pi''}(s) \leq \varphi_{\pi}(s)$ and $w_{\pi''}(s) \leq w_{\pi}(s)$ for $s \in S$, and if, in addition, $\varphi_{\pi''}(s) < \varphi_{\pi}(s)$ or $w_{\pi''}(s) < w_{\pi}(s)$ for some $s \in S$, then π'' is an improvement over π. It can be shown that π' is also an improvement over π in that case (see Denardo and Fox (1968)). This motivates the following definitions.

A policy π is called <u>unimprovable</u> if

$$\varphi_{\pi}(s) \leq \sum_{s' \in S} q_{\pi'}(s,s') \cdot \varphi_{\pi}(s') \ ,$$

$$w_{\pi}(s) \leq c_{\pi'}(s) - \sum_{s' \in S} \tau_{\pi'}(s,s') \cdot \varphi_{\pi}(s') + \sum_{s' \in S} q_{\pi'}(s,s') \cdot w_{\pi}(s') \ ,$$

for $s \in S$ and $\pi' \in \mathcal{D}$, assuming that all of the expressions above are well-defined and finite. A policy π is <u>strictly</u> unimprovable if it is unimprovable and if, in addition, equalities in the above expression are achieved simultaneously only when $\pi' = \pi$.

If the state space is finite, then an unimprovable policy is average optimal (see Denardo and Fox (1968)). If the state space is not finite, an unimprovable policy is not necessarily average optimal any more (see Hordijk (1974)). Thus, some additional conditions must be satisfied in order to be guaranteed that an unimprovable policy is optimal. Such conditions are given in the next sections.

2. The Case of Constant Optimal Expected Average Cost.

For many semi-Markov decision processes, the optimal long-run expected average cost is constant (i.e., independent of the start-state). In particular, if any state can be reached from each state (by using an appropriate policy) such that the expected cost up to the time the state is reached is well defined and finite, then the optimal long-run expected average cost must be constant. For in this case, the long-run expected average cost, given any start-state s and policy π, can be obtained for any other start-state by using a policy whose actions coincide with those of π at states which are reached from s with a non-zero probability under π, and otherwise are such that the expected cost up to the time when s is reached is finite.

For each $\pi \in \mathfrak{D}$, let x_π be the mapping from $S \times S$ into R_+ such that

$$x_\pi(s,s') = \lim_{t \to \infty} E_{\pi,s}\{\sum_{n \in N_t} \delta(s_n,s')\},$$

for $s,s' \in S$. Here, δ is the Kroenecker delta function, given by

$$\delta(s,s') = \begin{cases} 1 & \text{if } s = s', \\ 0 & \text{if } s \neq s'. \end{cases}$$

The fact that x_π exists (although possibly infinite valued) follows from renewal theory (see Smith (1955)). We assume that the expected time until the second decision epoch, given any start-state and action at the first decision epoch, non-zero. This implies that x_π is always finite valued.

Lemma 1: For each $\pi \in \mathcal{D}$,

$$x_\pi(s,s') = \sum_{s'' \in S} x_\pi(s,s'') \cdot q_\pi(s'',s') ,$$

for $s,s' \in S$.

Proof: For each $\alpha > 0$, $\pi \in \mathcal{D}$, let $x_{\pi,\alpha}$ be the mapping from $S \times S$ into R_+ such that

$$x_{\pi,\alpha}(s,s') = E_{\pi,s}\{ \sum_{n \in N} e^{-\alpha t_n} \cdot \delta(s_n,s') \} ,$$

for $s,s' \in S$. Since x_π exists,

$$x_\pi(s,s') = \lim_{\alpha \to 0} \alpha \cdot x_{\pi,\alpha}(s,s') ,$$

for $s,s' \in S$. Now

$$x_{\pi,\alpha}(s,s') = \sum_{s'' \in S} x_{\pi,\alpha}(s,s'') \cdot q_{\pi,\alpha}(s'',s') + \quad (s,s')$$

for $s,s' \in S$, where

$$q_{\pi,\alpha}(s,s') = \int_{t \in R_+} e^{-\alpha t} dq(s,a_\pi(s),s',t) .$$

This implies that

$$\lim_{\alpha \to 0} \alpha x_{\pi,\alpha}(s,s') = \lim_{\alpha \to 0} \sum_{s'' \in S} \alpha x_{\pi,\alpha}(s,s'') \cdot q_{\pi,\alpha}(s'',s') ,$$

or

$$x_\pi(s,s') = \sum_{s''\in S} \kappa_\pi(s,s'')\cdot q_\pi(s'',s'), \quad \text{for} \quad s,s' \in S .$$

Lemma 2: Let ϵ (> 0) be as in Assumption 1. Then, for each $\pi \in \mathcal{D}$,

$$E_{\pi,s}\{\sum_{n\in N_t} \delta(s',s_n)\} \leq \frac{t}{\epsilon} \cdot \frac{x_\pi(s,s')}{x_\pi(s,s)} , \quad \text{for} \quad t \in R_+ ,$$

for states s and s' which are positive recurrent under π.

Proof: Let π be a policy in $\mathcal{D}$, and let s and s' be positive recurrent states under π. By Lemma 1,

$$x_\pi(s,s') = \sum_{s''\in S} x_\pi(s,s'')\cdot E_{\pi,s''}\{\delta(s',s_2)\} .$$

Using Lemma 1 repeatedly, we obtain

$$x_\pi(s,s') = \sum_{s''\in S} x_\pi(s,s'')\cdot E_{\pi,s''}\{\delta(s',s_n)\}, \quad \text{for} \quad n \in N .$$

Therefore

$$E_{\pi,s}\{\delta(s',s_n)\} \leq \frac{x_\pi(s,s')}{x_\pi(s,s)}, \quad \text{for} \quad n \in N .$$

Now

$$E_{\pi,s}\{\sum_{n\in N_t} \delta(s',s_n)\} = \sum_{n\in N} E_{\pi,s}\{\delta(s',s_n)\cdot P_{\pi,s}\{t_n \leq t | s_n\}\}$$

$$\leq \sum_{n \leq \frac{t}{\epsilon}} E_{\pi,s}\{\delta(s',s_n)\} ,$$

by Assumption 1. The lemma now follows by combining the two last results.

Lemma 3: If π^* is an unimprovable policy such that $\varphi_{\pi^*}(s)$ is constant and

$$\sum_{s' \in S} x_{\pi}(s,s') \cdot |w_{\pi^*}(s')| < \infty ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$, then

$$\varphi_{\pi^*}(s) \cdot \sum_{s' \in S} x_{\pi}(s,s') \cdot v_{\pi}(s') \leq \sum_{s' \in S} x_{\pi}(s,s') \cdot c_{\pi}(s') ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$.

Proof: Let φ be the constant such that $\varphi = \varphi_{\pi^*}(s)$, for $s \in S$. Since π^* is unimprovable,

$$c_{\pi}(s') \geq w_{\pi^*}(s) + \varphi \cdot v_{\pi}(s') - \sum_{s'' \in S} q_{\pi}(s',s'') \cdot w_{\pi^*}(s'') ,$$

for $s' \in S$ and $\pi \in \mathfrak{D}$. Multiplying both sides by $x_{\pi}(s,s')$ and summing up over $s' \in S$ yields

$$\sum_{s' \in S} x_{\pi}(s,s') \cdot c_{\pi}(s')$$

$$\geq \sum_{s' \in S} x_{\pi}(s,s') \{ w_{\pi^*}(s') + \varphi \cdot v_{\pi}(s') - \sum_{s'' \in S} q_{\pi}(s',s'') w_{\pi^*}(s'') \} ,$$

for $s \in S$, $\pi \in \mathfrak{D}$. The sums on both sides of the above inequality exists, since

$$\sum_{s' \in S} x_{\pi}(s,s') \{ w_{\pi^*}(s') + \varphi - v_{\pi}(s') - \sum_{s'' \in S} q_{\pi}(s',s'') w_{\pi^*}(s'') \}^-$$

$$\leq \sum_{s' \in S} x_{\pi}(s,s') w_{\pi^*}(s')^- + \sum_{s',s'' \in S} x_{\pi}(s,s') q_{\pi}(s',s'') w_{\pi^*}(s'')^+$$

$$+ \varphi^- \cdot \sum_{s' \in S} x_{\pi}(s,s') \cdot v_{\pi}(s')$$

$$= \sum_{s' \in S} x_\pi(s,s') w_{\pi^*}(s')^- + \sum_{s'' \in S} x_\pi(s,s'') w_{\pi^*}(s'')^+ + \varphi^-$$

(using Lemma 1 and Lemma 2)

$$< \infty ,$$

using the assumption of the lemma. Now

$$\sum_{s' \in S} x_\pi(s,s') \cdot \{ w_{\pi^*}(s') + \varphi \cdot \nu_\pi(s') - \sum_{s'' \in S} q_\pi(s',s'') w_{\pi^*}(s'') \}$$

$$\geq \sum_{s' \in S} x_\pi(s,s') w_{\pi^*}(s') + \varphi \cdot \sum_{s' \in S} x_\pi(s,s') \cdot \nu_\pi(s')$$

$$- \sum_{s',s'' \in S} x_\pi(s,s') q_\pi(s',s'') w_{\pi^*}(s'')$$

$$= \sum_{s' \in S} x_\pi(s,s') w_{\pi^*}(s') - \sum_{s'' \in S} x_\pi(s,s'') w_{\pi^*}(s'') + \varphi \cdot \sum_{s' \in S} x_\pi(s,s') \cdot \nu_\pi(s')$$

(using Lemma 1)

$$= \varphi \cdot \sum_{s' \in S} x_\pi(s,s') \cdot \nu_\pi(s') ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$, and the lemma follows.

Q.E.D.

For each $\pi \in \mathfrak{D}$, let $R(\pi)$ denote as before the set of positive recurrent states under π, and let $T(\pi)$ denote the set of the other states. For each $\pi \in \mathfrak{D}$, let y_π be the mapping from $S \times S$ into R_+ such that

$$y_\pi(s,s') = \begin{cases} E_{\pi,s}\{ \sum_{n \in N} \delta(s',s_n) \}, & \text{for } s' \in T(\pi),\ s \in S , \\ 1 & \text{for } s' \in R(\pi),\ s \in S . \end{cases}$$

In words, $y_\pi(s,s')$ is the expected number of times the state of the system is s' before a positive recurrent state is entered from another state, given that the start-state is s and that the policy π is used.

Theorem 4: If π^* is an unimprovable policy such that $\varphi_{\pi^*}(s)$ is constant and

$$\sum_{s' \in S} (y_\pi(s,s') + x_\pi(s,s')) \cdot |w_{\pi^*}(s)| < \infty ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$, then π^* is average optimal.

Proof: We first show that

$$\varphi_\pi(s) \geq \sum_{s' \in S} x_\pi(s,s') \cdot c_\pi(s') ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$.

For each $\pi \in \mathfrak{D}$, let $\tilde{q}_\pi$ and $\tilde{c}_\pi$ be the mappings from $S \times S$ and S into R such that

$$\tilde{q}_\pi(s,s') = \begin{cases} q_\pi(s,s'), & \text{for } s \in T(\pi) , \\ 0 , & \text{for } s \in R(\pi) , \end{cases}$$

$$\tilde{c}_\pi(s) = \begin{cases} c_\pi(s) - \varphi \cdot \nu_\pi(s), & \text{for } s \in T(\pi) , \\ w_{\pi^*}(s) , & \text{for } s \in R(\pi) , \end{cases}$$

for $s' \in S$. Since π^* is unimprovable,

$$\tilde{c}_\pi(s) \geq w_{\pi^*}(s) - \sum_{s' \in S} \tilde{q}_\pi(s,s') \cdot w_{\pi^*}(s') ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$. Now

$$\sum_{s\in S} y_\pi(s'',s)\{w_{\pi^*}(s) - \sum_{s'\in S} \tilde{q}_\pi(s,s')\cdot w_{\pi^*}(s')\}^-$$

$$\leq \sum_{s\in S} y_\pi(s'',s)\cdot w_{\pi^*}(s)^- + \sum_{s\in S} y_\pi(s'',s) \sum_{s'\in S} \tilde{q}_\pi(s,s')w_{\pi^*}(s')^+$$

$$= \sum_{s\in S} y_\pi(s'',s)w_{\pi^*}(s)^- + \sum_{s'\in S} (y_\pi(s'',s') - \delta(s'',s))\cdot w_{\pi^*}(s')^+$$

$$= \sum_{s\in S} y_\pi(s'',s)w_{\pi^*}(s)^- + \sum_{s'\in S} y_\pi(s'',s)\cdot w_{\pi^*}(s')^+ - w_{\pi^*}(s'')^+$$

$$< \infty ,$$

by the last assumption of the theorem. This implies that

$$\sum_{s\in S} y_\pi(s'',s)\cdot \tilde{c}_\pi(s)^- < \infty, \quad s'' \in S, \quad \pi \in \mathfrak{D} .$$

Thus

$$\sum_{s\in S} y_\pi(s'',s)\cdot \tilde{c}_\pi(s)$$

is well-defined and greater than minus infinity for $s'' \in S$ and $\pi \in \mathfrak{D}$.

Now

$$\varphi_\pi(s) = \lim_{t\to\infty} E_{\pi,s}\{\sum_{n\in N_t} c(s_n,a_n,t-t_n)\}/t$$

$$= \lim_{t\to\infty} E_{\pi,s}\{\sum_{n\in N_t} c(s_n,a_n,\infty)\}/t$$

(by Assumption 2)

$$= \lim_{t\to\infty} E_{\pi,s}\{\sum_{n\in N_t} \sum_{s'\in S} \delta(s',s_n)\cdot c_\pi(s')\}/t$$

$$= \lim_{t\to\infty} E_{\pi,s}\{\sum_{n\in N_t} \sum_{s'\in T(\pi)} \delta(s',s_n)\cdot c_\pi(s')\}/t$$

$$+ \lim_{t\to\infty} E_{\pi,s}\{\sum_{n\in N_t} \sum_{s'\in R(\pi)} \delta(s',s_n)\cdot c_\pi(s')\}/t$$

$$= \varphi + \lim_{t \to \infty} \sum_{s' \in T(\pi)} E_{\pi,s}\{\sum_{n \in N_t} \delta(s',s_n)\} \cdot \tilde{c}_\pi(s')/t$$

$$+ \lim_{t \to \infty} \sum_{s' \in R(\pi)} E_{\pi,s}\{\sum_{n \in N_t} \delta(s',s_n)\} \cdot (c_\pi(s') - \varphi \cdot \nu_\pi(s'))/t,$$

using Assumption 2. The first limit is non-negative, since

$$E_{\pi,s}\{\sum_{n \in N_t} \delta(s',s_n)\} \leq y_\pi(s,s'), \quad \text{for} \quad s' \in S ,$$

and since

$$\sum_{s' \in S} y_\pi(s,s') \cdot \tilde{c}_\pi(s')^- < \infty .$$

Therefore

$$\varphi_\pi(s) \geq \varphi + \lim_{t \to \infty} \sum_{s' \in R(\pi)} E_{\pi,s}\{\sum_{n \in N_t} \delta(s',s_n)\}(c_\pi(s') - \varphi \cdot \nu_\pi(s'))/t .$$

Using Lebesque's bounded convergence theorem, we obtain

$$\lim_{t \to \infty} \sum_{s' \in R(\pi)} E_{\pi,s}\{\sum_{n \in N_t} \delta(s',s_n)\} \cdot (c_\pi(s') - \varphi \cdot \nu_\pi(s'))/t$$

$$= \sum_{s' \in R(\pi)} x_\pi(s,s') \cdot (c_\pi(s') - \varphi \cdot \nu_\pi(s')) ,$$

since

$$\lim_{t \to \infty} E_{\pi,s}\{\sum_{n \in N_t} \delta(s',s_n)\}/t = \nu_\pi(s,s'), \quad \text{for} \quad s' \in S ,$$

$$E_{\pi,s}\{\sum_{n \in N_t} \delta(s',s_n)\}/t \leq \frac{1}{\epsilon} \cdot \frac{x_\pi(s'',s')}{x_\pi(s'',s'')} , \quad \text{for} \quad s' \in S, \; s'' \in R(\pi) ,$$

and

$$\sum_{s' \in S} x_{\pi}(s'',s')(c_{\pi}(s') - \varphi \cdot \nu_{\pi}(s'))^{-} < \infty \ .$$

Thus

$$\varphi_{\pi}(s) \geq \varphi + \sum_{s' \in S} x_{\pi}(s,s')(c_{\pi}(s') - \varphi \cdot \nu_{\pi}(s')) \ .$$

Using Lemma 3, we obtain

$$\varphi_{\pi}(s) \geq \varphi \ .$$

Q.E.D.

Corollary 5: Suppose that, for each $s \in S$ and $\pi \in \mathcal{D}$, the expected number of decision epochs occurring before reaching a state in $R(\pi)$ is finite. Then, if π^* is an unimprovable policy such that $\varphi_{\pi^*}(s)$ is constant and, in addition, $w_{\pi^*}(s)$ is bounded, then π^* is average optimal.

Proof: In view of the theorem and the fact that $w_{\pi^*}(s)$ is bounded, we only need to show that

$$\sum_{s' \in S} y_{\pi}(s,s') < \infty \ ,$$

for $s \in S$ and $\pi \in \mathcal{D}$. But this follows from the first assumption of the corollary, which completes the proof.

Theorem 6: If π^* is a strictly unimprovable policy such that $\varphi_{\pi^*}(s)$ is constant and, in addition,

$$\sum_{s' \in S} (y_\pi(s,s') + x_\pi(s,s')) \cdot |w_{\pi^*}(s')| < \infty ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$, then π^* is undiscounted optimal.

Proof: Let π be any average optimal stationary, deterministic policy. Following the proof of Lemma 3 and Theorem 4, one can easily see that $a_\pi(s) \neq a_{\pi^*}(s)$ imply that $\varphi_\pi(s) > \varphi_{\pi^*}(s)$ for $s \in R(\pi)$, since π^* is strictly unimprovable. This implies that $a_\pi(s) = a_{\pi^*}(s)$ for $s \in R(\pi)$. From the proof of Theorem 4,

$$c_\pi(s) \geq w_{\pi^*}(s) - \sum_{s' \in S} \tilde{q}_\pi(s,s') w_{\pi^*}(s') ,$$

for $s \in S$. This implies that

$$\sum_{s \in S} y_\pi(s'',s) \tilde{c}_\pi(s) \geq \sum_{s \in S} y_\pi(s'',s) \{w_{\pi^*}(s) - \sum_{s' \in S} \tilde{q}_\pi(s,s') w_{\pi^*}(s')\} .$$

It was shown in the proof of Theorem 4 that these sums are well-defined. Now

$$\begin{aligned} &\sum_{s \in S} y_\pi(s'',s) \{w_{\pi^*}(s) - \sum_{s' \in S} \tilde{q}_\pi(s,s') w_{\pi^*}(s')\} \\ &= \sum_{s \in S} y_\pi(s'',s) w_{\pi^*}(s) - \sum_{s,s' \in S} y_\pi(s'',s) \tilde{q}_\pi(s,s') w_{\pi^*}(s') \\ &= \sum_{s \in S} y_\pi(s'',s) w_{\pi^*}(s) - \sum_{s' \in S} (y_\pi(s'',s') - \delta(s'',s')) w_{\pi^*}(s') \\ &= w_{\pi^*}(s'') , \end{aligned}$$

for $s'' \in S$. Hence

$$w_{\pi^*}(s'') \leq \sum_{s \in S} y_\pi(s'',s) \cdot \tilde{c}_\pi(s) ,$$

for $s'' \in S$ and $\pi \in \mathcal{D}$. It is easy to check that

$$w_\pi(s'') = \sum_{s \in S} y_\pi(s'',s) \cdot \tilde{c}_\pi(s) ,$$

for $s'' \in S$, so

$$w_{\pi^*}(s'') \leq w_\pi(s'') ,$$

for $s'' \in S$.

Q.E.D.

Corollary 7: Suppose that for each $s \in S$ and $\pi \in \mathcal{D}$, the expected number of decision epochs occurring before reaching a state in $R(\pi)$ is finite. Then, if π^* is a strictly unimprovable policy such that $\varphi_{\pi^*}(s)$ is constant and, in addition, $w_{\pi^*}(s)$ is bounded, then π^* is undiscounted optimal.

Proof: The proof proceeds just as in the proof of Corollary 5, and so will not be repeated here.

3. The Case of Non-Constant Optimal Expected Average Cost.

The case when the optimal long-run expected average cost varies with the start-state now will be considered. The notation is the same as in Section 2.

Lemma 8: If π^* is a policy such that

$$\varphi_{\pi^*}(s) \leq \sum_{s' \in S} q_\pi(s,s') \cdot \varphi_{\pi^*}(s') ,$$

$$\sum_{s' \in S} x_\pi(s,s') \cdot |\varphi_{\pi^*}(s')| < \infty ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$, then $\varphi_{\pi^*}(s)$ is constant in each positive recurrent class of states under each policy $\pi \in \mathfrak{D}$.

Proof: Let π be a policy in $\mathfrak{D}$, and let s be a state in $R(\pi)$. Using Lemma 1 repeatedly, we obtain

$$x_\pi(s,s'') = \sum_{s' \in S} x_\pi(s,s') \cdot E_{\pi,s'}\{\delta(s'',s_n)\} ,$$

for $n \in N$ and $s'' \in S$. This implies that

$$E_{\pi,s}\{\delta(s'',s_n)\} \leq \frac{x_\pi(s,s'')}{x_\pi(s,s)} ,$$

for $n \in N$, and $s'' \in S$, since $x_\pi(s,s) > 0$. Now

$$\sum_{s'' \in S} \frac{x_\pi(s,s'')}{x_\pi(s,s)} \cdot |\varphi_{\pi^*}(s'')| < \infty ,$$

because of the second assumption of the lemma. Using Lebesque's bounded convergence theorem, we obtain

$$\lim_{n \to \infty} \sum_{s'' \in S} E_{\pi,s}\{\delta(s'',s_n)\} \cdot \varphi_{\pi^*}(s'')$$

$$= \sum_{s'' \in S} x_\pi(s,s'') \cdot \varphi_{\pi^*}(s'') ,$$

or equivalently,

$$\lim_{n \to \infty} E_{\pi,s}\{\varphi_{\pi^*}(s_n)\} = \sum_{s'' \in S} x_\pi(s,s'') \cdot \varphi_{\pi^*}(s'') .$$

Let d_π be the mapping from S into R such that

$$d_\pi(s'') = \sum_{s' \in S} q_\pi(s'',s') \cdot \varphi_{\pi^*}(s') - \varphi_{\pi^*}(s'') ,$$

for $s'' \in S$. d_π is well-defined by the first assumption of the lemma. It can easily be shown by induction on n that

$$E_{\pi,s''}\{\sum_{i \leq n} d_\pi(s_i)\} = E_{\pi,s''}\{\varphi_{\pi *}(s_n)\} - \varphi_{\pi *}(s'') ,$$

for $n \in N$ and $s'' \in S$. Inserting s for s'' in this expression and taking the limits as n goes to infinity, we obtain

$$\lim_{n \to \infty} E_{\pi,s}\{\sum_{i \leq n} d_\pi(s_i)\} = \lim_{n \to \infty} E_{\pi,s}\{\varphi_{\pi *}(s_n)\} - \varphi_{\pi *}(s)$$

$$= \sum_{s' \in S} x_\pi(s,s') \cdot \varphi_{\pi *}(s') - \varphi_{\pi *}(s) < \infty .$$

The first assumption of the lemma implies that $d_\pi(s') \geq 0$, for $s' \in S$ and $\pi \in \mathfrak{D}$. Using this fact together with

$$\lim_{n \to \infty} E_{\pi,s}\{\sum_{i \leq n} d_\pi(s_i)\} < \infty ,$$

we obtain $d_\pi(s) = 0$. But $s \in R(\pi)$ was chosen arbitrarily, so $d_\pi(s) = 0$ for $s \in R(\pi)$. This implies that

$$\varphi_{\pi *}(s) = \sum_{s' \in R(\pi)} q_\pi(s,s') \cdot \varphi_{\pi *}(s) ,$$

for $s \in R(\pi)$. This, in turn, implies that

$$\varphi_{\pi *}(s) = \lim_{n \to \infty} E_{\pi,s}\{\varphi_{\pi *}(s_n)\}$$

$$= \sum_{s' \in S} x_\pi(s,s') \cdot \varphi_{\pi *}(s') ,$$

for $s \in R(\pi)$. Now, $x_\pi(s,s') = x_\pi(s'',s')$, for s and s'', if they belong to the same positive recurrent class under π. Thus,

$$\varphi_{\pi^*}(s) = \sum_{s' \in S} x_{\pi}(s,s') \cdot \varphi_{\pi^*}(s') = \varphi_{\pi^*}(s'') ,$$

for s,s'' in the same positive recurrent class under π.

Q.E.D.

For each $\pi \in \mathfrak{D}$, let $I(\pi)$ be the set of positive recurrent classes, and for each $s \in S$ and $z \in I(\pi)$, let $r_{\pi}(s,z)$ be the probability that class z is entered, given start-state s and policy π.

Lemma 9: If π^* is an unimprovable policy such that the conditions of the previous lemma hold, and, in addition,

$$\lim_{n \to \infty} \inf \sum_{s' \in T(\pi)} E_{\pi,s}\{\delta(s',s_n)\} \cdot \varphi_{\pi^*}(s') \leq 0 ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$, then

$$\varphi_{\pi^*}(s) \leq \sum_{z \in I(\pi)} p_{\pi}(s,z) \cdot \varphi_z ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$. Here, φ_z is the long-run expected average cost under π^*, given that the start-state is in the class z.

Proof: Let π be any policy in $\mathfrak{D}$, and let S_z be the set of states belonging to class z for each $z \in I(\pi)$. As in the proof of Lemma 8,

$$\varphi_{\pi^*}(s) \leq \lim_{n \to \infty} \inf E_{\pi,s}\{\varphi_{\pi^*}(s_n)\}$$

$$= \lim_{n \to \infty} \sum_{s' \in R(\pi)} E_{\pi,s}\{\delta(s',s_n)\} \cdot \varphi_{\pi^*}(s')$$

$$+ \lim_{n \to \infty} \inf \sum_{s' \in T(\pi)} E_{\pi,s}\{\delta(s',s_n)\} \cdot \varphi_{\pi^*}(s')$$

$$\leq \lim_{n \to \infty} \sum_{s' \in R(\pi)} E_{\pi,s}\{\delta(s',s_n)\} \cdot \varphi_{\pi^*}(s') ,$$

for $s \in S$. The last limit exists and is finite. By Lemma 2,

$$E_{\pi,s}\{\delta(s',s_n)\} \leq p_\pi(s,z) \leq \frac{1}{\epsilon} \frac{x_\pi(s'',s')}{x_\pi(s'',s'')} \quad \text{for some } s'' \in R(\pi),\ \epsilon > 0 ,$$

for $s' \in S_z$, $s \in S$ and $z \in I(\pi)$. Now

$$\sum_{z \in I(\pi)} p_\pi(s,z) \cdot |\varphi_z| \leq \frac{1}{\epsilon} \sum_{s'' \in S} \frac{x_\pi(s'',s')}{x_\pi(s'',s'')} \cdot |\varphi_{\pi^*}(s'')| < \infty$$

for $s \in S$. Therefore, by Lebesgue's bounded convergence theorem,

$$\lim_{n \to \infty} \sum_{s' \in R(\pi)} E_{\pi,s}\{\delta(s',s_n)\} \cdot \varphi_{\pi^*}(s')$$

$$= \sum_{z \in I(\pi)} p_\pi(s,z) \cdot \varphi_z ,$$

for $s \in S$. We conclude that

$$\varphi_{\pi^*}(s) \leq \sum_{z \in I(\pi)} p_\pi(s,z) \cdot \varphi_z ,$$

for $s \in S$.

Q.E.D.

Let y_π be defined as in Section 2.

Theorem 10: If π^* is an unimprovable policy such that

$$\sum_{s' \in S} x_{\pi}(s,s') \cdot |\varphi_{\pi^*}(s')| < \infty ,$$

$$\sum_{s' \in S} \{x_{\pi}(s,s') + y_{\pi}(s,s')\} \cdot |w_{\pi^*}(s')| < \infty ,$$

$$\lim_{n \to \infty} \inf \sum_{s' \in T(\pi)} E_{\pi,s}\{\delta(s',s_n)\} \cdot \varphi_{\pi^*}(s') ,$$

for $s \in S$ and $\pi \in \mathfrak{D}$, then π^* is average optimal.

Proof: Let π be any policy in $\mathfrak{D}$. By Lemma 8, $\varphi_{\pi^*}(s)$ is constant for $s \in S_z$ $(z \in I(\pi))$. Therefore, by Theorem 4, $\varphi_{\pi^*}(s) \leq \varphi_{\pi}(s)$, for $s \in S_z$ $(z \in I(\pi))$. Using Lemma 9 together with this, we obtain $\varphi_{\pi}*(s) \leq \varphi_{\pi}(s)$ for $s \in S$. The costs incurred until a positive recurrent state is reached do not contribute anything to the average cost, since it can be shown (as in the proof of Theorem 4) that the expected cost until a positive recurrent state is reached is finite. Thus, π^* must be average optimal.

Q.E.D.

Corollary 11: Suppose that, for each $s \in S$ and $\pi \in \mathfrak{D}$, the expected number of decision epochs recurring before reaching a state in $R(\pi)$ is finite.

If π^* is an unimprovable policy such that $\varphi_{\pi^*}(s)$ and $w_{\pi^*}(s)$ are bounded, then π^* is average optimal.

Proof: We only need to show that

$$\sum_{s' \in S} x_{\pi}(s,s') < \infty ,$$

$$\sum_{s' \in S} y_{\pi}(s,s') < \infty ,$$

for each $s \in S$. The first sum is finite by an assumption made in Section 1, the second sum is finite by Lemma 2, and the third sum is finite by the first assumption of the corollary. Thus, the corollary follows.

Q.E.D.

Theorem 12: If π^* is a strictly unimprovable policy such that the conditions of Theorem 10 are satisfied, then π^* is undiscounted optimal.

Proof: The proof proceeds just as in the proof of Theorem 6, and so will not be repeated here.

Corollary 13: If π^* is a strictly unimprovable policy such that the conditions of Corollary 11 are satisfied, then π^* is undiscounted optimal.

See the proof of Corollary 11.

CHAPTER 3

SEMI-MARKOV DECISION PROCESSES WITH DISCOUNTING

In this chapter the optimization problem arising when the costs are discounted is investigated. From an economic viewpoint, this problem is somewhat more interesting than the problem without discounting. It has been studied by a number of investigators who have made various assumptions about the state and action spaces, the motion of the system and the costs (see Section 2 in Chapter 1). Here, the assumptions made by other authors are weakened, and more general results are obtained.

In Section 1, there is a formal statement of the problem to be considered. It also contains some preliminary results. In Section 2, some useful operators are introduced. In Section 3, the optimality equation is proven. In Section 4, there are some existence theorems. In Section 5, policy improvement is considered. In Section 6, necessary and sufficient conditions for optimality are presented. Finally, in Section 7, there is an analysis using the contraction properties of a certain operator. An alternative set of necessary and sufficient conditions for optimality are obtained.

1. Problem Formulation.

As before, let S be the state space, $(A_s)_{s\in S}$ be the set of action spaces, q be the law of motion, and c be the cost function of the SMDP. For each n in N, let s_n, a_n and t_n denote the state of the system, the action and the time of the n^{th} decision epoch, respectively. The first decision epoch is taken to occur at time zero, so

$t_1 \equiv 0$. Also, let $\mathcal{P}$, $\mathcal{S}$ and $\mathcal{D}$ denote the set of all policies, the set of stationary policies and the set of deterministic stationary policies, respectively. Let $A = \bigcup_{s \in S} A_s$.

Let α be a given positive interest rate, and let c_α be the mapping from $S \times A$ into R such that

$$c_\alpha(s,a) = \int_{t \in R^+} e^{-\alpha t} dc(s,a,t)$$

for $a \in A_s$ for $s \in S$. In other words, $c_\alpha(s,a)$ is the expected discounted cost incurred until the second decision epoch, given that the start-state is s and that the first action is a. Naturally, it is assumed that c_α exists.

For each π in $\mathcal{P}$, let v_π^+, v_π^- and v_π be the three functions from S into $R_+ \cup \{\infty\}$, $R_+ \cup \{\infty\}$ and $R \cup \{\infty\}$, respectively, such

$$v_\pi^+(s) = E_{\pi,s}\{\sum_{n \in N} e^{-\alpha t_n} \cdot c_\alpha(s_n,a_n)^+\} ,$$

$$v_\pi^-(s) = E_{\pi,s}\{\sum_{n \in N} e^{-\alpha t_n} \cdot c_\alpha(s_n,a_n)^-\} ,$$

$$v_\pi(s) = v_\pi^+(s) - v_\pi^-(s) ,$$

for s in S, where E is the expectation operator and the subscripts π and s indicate that the start-state is s and that the policy π is used. In words, $v_\pi(s)$ is the total expected discounted cost, given that the start-state is s and that the policy π is used. v_π is the <u>value function</u> of the policy π. Clearly, v_π^+ and v_π^- are well-defined (possibly infinite-valued). In order that v_π be well-defined, the following assumption is made:

Assumption 1: $v_\pi^-(s) < \infty$, for $s \in S$, $\pi \in \mathcal{P}$.

Let v_α be the function from S into $R \cup \{\infty, -\infty\}$ such that

$$v_\alpha(s) = \inf_{\pi \in \mathcal{P}} v_\pi(s) ,$$

for s in S. For purposes which will become clear later, the following assumption is made.

Assumption 2: $v_\alpha(s) > -\infty$, for $s \in S$.

If there can be an infinite number of decision epochs in a finite amount of time, some of the costs may unintentionally be ignored by the definition of v_π. In order to eliminate this problem, the following assumption is made:

Assumption 3: $P_{\pi,s}\{t_{n-1} \leq t \text{ for } n \in N\} = 0$, for $t \in R_+$, $s \in S$, $\pi \in \mathcal{P}$.

Here, P is the probability operator and the subscripts π and s indicate that the start-state is s and that the policy π is used. For purposes that will become clear later, a fourth assumption is made:

Assumption 4: Given $\epsilon > 0$, there is an m (possibly depending on s) such that

$$E_{\pi,s}\{\sum_{n \geq m} e^{-\alpha t_n} c_\alpha(s_n, a_n)^-\} \leq \epsilon$$

for π in $\mathcal{P}$.

These assumptions are satisfied trivially if $c_\alpha(s,a)$ is non-negative for each s and a. The following theorem gives some weaker conditions under which the assumptions hold.

Theorem 1: If c_α is uniformly bounded from below and there is a $\beta < 1$ such that

$$E_{\pi,s}\{e^{-\alpha t_2}\} \leq \beta ,$$

for s in S and π in P, then all the assumptions above hold.

Proof: Let β be as in the theorem. For each $n \in N$,

$$\begin{aligned} & E_{\pi,s}\{e^{-\alpha t_{n+1}}\} \\ & \quad = E_{\pi,s}\{e^{-\alpha t_n} \cdot E_{\pi,s}\{e^{-\alpha(t_{n+1}-t_n)} | s_n, a_n\}\} \\ & \quad \leq \beta \cdot E_{\pi,s}\{e^{-\alpha t_n}\} \end{aligned}$$

since

$$E_{\pi,s}\{e^{-\alpha(t_{n+1}-t_n)} | s_n, a_n\} \leq \beta .$$

This implies that

$$E_{\pi,s}\{e^{-\alpha t_n}\} \leq \beta^{n-1} ,$$

for $n \in N$. For each m in N,

$$\begin{aligned} E_{\pi,s}\{\sum_{n \geq m} e^{-\alpha t_n}\} &= \sum_{n \geq m} E_{\pi,s}\{e^{-\alpha t_n}\} \\ &\leq \sum_{n \geq m} \beta^{n-1} = \beta^{m-1} \cdot (1-\beta)^{-1} . \end{aligned}$$

This implies that

$$(1-\beta)^{-1} \geq E_{\pi,s}\{\sum_{n\in N} e^{-\alpha t_n}\} \geq E_{\pi,s}\{\sum_{n \leq m} e^{-\alpha t_n}\}$$

$$\geq me^{-\alpha t} \cdot P_{\pi,s}\{t_n \leq t \text{ for } n \leq m\}$$

for $t \in R_+$ and $m \in N$. Thus

$$P_{\pi,s}\{t_n \leq t \text{ for } n \leq m\} \leq e^{\alpha t} \cdot (1-\beta)^{-1} \cdot m^{-1}$$

for $t \in R_+$ and $m \in N$. Assumption 3 follows by taking the limits as m goes to infinity.

Let M be an upper bound on $-c_\alpha(s,a)$. Then

$$E_{\pi,s}\{\sum_{n \geq m} e^{-\alpha t_n} \cdot c_\alpha(s_n,a_n)^-\}$$

$$\leq M \cdot E_{\pi,s}\{\sum_{n \geq m} e^{-\alpha t_n}\}$$

$$\leq M \cdot \beta^{m-1} \cdot (1-\beta)^{-1}$$

for $m \in N$. This shows that the rest of the assumptions hold.

2. The Operators Q_π and T_π.

Let B be the set of mappings from S into $R \cup \{\infty\}$. For each $\pi \in P$, define the operators Q_π and T_π from B into B by

$$(Q_\pi v)(s) = E_{\pi,s}\{e^{-\alpha t_2} \cdot v(s_2)\}, \text{ for } s \in S,$$

$$(T_\pi v)(s) = E_{\pi,s}\{c_\alpha(s_1,a_1) + e^{-\alpha t_2} \cdot v(s_2)\}, \text{ for } s \in S,$$

for v in B. For some v in B, the above expressions may not be well-defined. Those functions will, however, not be used.

Some compact notation will be used. If u and v are functions in B, then $u \leq v$ means that $u(s) \leq v(s)$ for $s \in S$, $u + v$ is the function such that $(u+v)(s) = u(s) + v(s)$ for $s \in S$, and if c is a constant, then cv is the function such that $(cv)(s) = c \cdot v(s)$ for $s \in S$, etc.

Lemma 2: If u and v in B are such that $u \leq v$, then $Q_\pi u \leq Q_\pi v$ and $T_\pi u \leq T_\pi v$, provided the expressions are well-defined.

For each $n \in N$ and $\pi \in P$, define the operators Q_π^n and T_π^n by

$$(Q_\pi^n v)(s) = E_{\pi,s}\{e^{-\alpha t_{n+1}} \cdot v(s_{n+1})\}, \quad \text{for } s \in S,$$

$$(T_\pi^n v)(s) = E_{\pi,s}\{\sum_{i \leq n} e^{-\alpha t_i} \cdot c_\alpha(s_i,a_i) + e^{-\alpha t_{n+1}} \cdot v(s_{n+1})\}, \text{ for } s \in S,$$

for v in B. Again, these expressions need not always be well-defined for each v in B. If, however, v is the value function of a policy, then the expressions are clearly well-defined.

Let θ be the function (from S into R) which is zero everywhere. Then

$$(T_\pi^n \theta)(s) = E_{\pi,s}\{\sum_{i \leq n} e^{-\alpha t_i} \cdot c_\alpha(s_i,a_i)\}, \quad s \in S,$$

and

$$v_\pi = \lim_{n \to \infty} T_\pi^n \theta,$$

for any π in P.

Lemma 3: For each $n \in N$ and $\pi \in T$, $Q_\pi^n v_\alpha$ and $T_\pi^n v_\alpha$ are well-defined.

Proof: Let $\epsilon \geq 0$ be given, and let π' be an ϵ-optimal policy. This means that $v_\alpha \geq v_{\pi'} + \epsilon \cdot \mathbb{1}$ where $\mathbb{1}$ is the function from S into {1}. This implies that

$$v_\alpha^- \leq (v_{\pi'} + \epsilon \cdot \mathbb{1})^- \leq v_{\pi'}^- + \epsilon \cdot \mathbb{1} .$$

Therefore

$$Q_\pi^n v_\alpha^- \leq Q_\pi^n (v_{\pi'}^- + \epsilon \cdot \mathbb{1}) \leq Q_\pi^n v_{\pi'}^- + \epsilon \cdot \mathbb{1} ,$$

since

$$Q_\pi^n \mathbb{1} \leq \mathbb{1} .$$

This implies that $Q_\pi^n v_\alpha^-$ is finite-valued, and thus $Q_\pi^n v_\alpha$ is well-defined. Also

$$T_\pi^n v_\alpha^- \leq T_\pi^n v_{\pi'}^- + \epsilon \cdot \mathbb{1} ,$$

since

$$Q_\pi^n \mathbb{1} \leq \mathbb{1} .$$

This implies that $T_\pi^n v_\alpha^-$ is finite-valued, and thus $T_\pi^n v_\alpha$ is well-defined.

Lemma 4: For each $\pi \in P$,

$$\lim_{n \to \infty} \inf Q_\pi^n v_\alpha \geq 0 .$$

Proof: Let $\epsilon > 0$ be given. From the proof of Lemma 3, there is a policy π' such that

$$Q_\pi^n \bar{v}_\alpha \leq Q_{\pi'}^n \bar{v}_{\pi'} + \epsilon \cdot \mathbb{1} ,$$

for all π in P. This implies that

$$\lim_{n \to \infty} Q_\pi^n \bar{v}_\alpha \leq \lim_{n \to \infty} Q_{\pi'}^n \bar{v}_{\pi'} + \epsilon \cdot \mathbb{1}$$

$$= \epsilon \cdot \mathbb{1} ,$$

by Assumption 4. The lemma follows, since ϵ is arbitrary.

3. The Optimality Equation.

Bellman (1957) introduced the principle of optimality for dynamic programming. He says (p. 83), "An optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision." Since an optimal policy need not always exist, the principle has a limited potential use. More useful is the optimality equation, given in the theorem below. For a discussion of the principle of optimality and the optimality equation, see Porteus (1975a).

Let q_α be the mapping from $S \times A \times S$ into R such that

$$q_\alpha(s,a,s') = \int_0^\infty e^{-\alpha t} dq(s,a,s',t) ,$$

for $a \in A_s$ for $s', s \in S$.

Theorem 5: For each s in S,

$$v_\alpha(s) = \inf_{a \in A_s} \{c_\alpha(s,a) + \sum_{s' \in S} q_\alpha(s,a,s') \cdot v_\alpha(s')\} .$$

Proof: The proof is similar to the one given in Ross (1970, p. 121) for the case when the action spaces are finite. Let π' be an ϵ-optimal policy. This exists for each $\epsilon > 0$, since $v_\alpha(s) > -\infty$ for each $s \in S$ by Assumption 2. Then

$$v_\alpha \leq T_{\pi'} v_{\pi'} \leq T_{\pi'}(v_\alpha + \epsilon \cdot \mathbb{1}) = T_{\pi'} v_\alpha + Q_{\pi'}(\epsilon \cdot \mathbb{1})$$

$$\leq T_{\pi'} v_2 + \epsilon \cdot \mathbb{1}$$

for all $\pi \in \mathcal{P}$. Since ϵ is arbitrary, $v_\alpha \leq T_\pi v_\alpha$ for $\pi \in \mathcal{P}$. This is equivalent to

$$v_\alpha(s) \leq \inf_{a \in A_s} \{c_\alpha(s,a) + \sum_{s' \in S} q_\alpha(s,a,s') \cdot v_\alpha(s')\} ,$$

for $s \in S$. We now show that this inequality also holds in the opposite direction.

For each $s \in S$,

$$v_\pi(s) = E_{\pi,s}\{\sum_{n \in N} e^{-\alpha t_n} \cdot c_\alpha(s_n,a_n)\}$$

$$= E_{\pi,s}\{c_\alpha(s_1,a_1) + e^{-\alpha t_2} \cdot E_{\pi,s}\{\sum_{n > 1} e^{-\alpha(t_n - t_1)} \cdot c_\alpha(s_n,a_n) | a_1, s_2, t_2\}\} .$$

Now

$$E_{\pi,s}\{\sum_{n > 1} e^{-\alpha(t_n - t_1)} \cdot c_\alpha(s_n,a_n) | a_1, s_2, t_2\} \geq v_\alpha(s_2) .$$

To see this, suppose the opposite. Then there must be a', s' and t' such that

$$E_{\pi,s}\{\sum_{n>1} e^{-\alpha(t_n-t_1)} \cdot c_\alpha(s_n,a_n)|a_1 = a',\ s_2 = s',\ t_2 = t'\} < v_\alpha(s') .$$

For each $n \in N$, let h_n denote the history of the process up to the n^{th} decision epoch (including the state at that time). Let π' be a policy such that for each history h,

$$P_{\pi',s'}\{a_n = a|h_n = h\} = P_{\pi,s}\{a_{n+1} = a|h_{n+1} = (a',s',t',h)\} .$$

Then

$$v_{\pi'}(s') = E_{\pi,s}\{\sum_{n>1} e^{-\alpha(t_n-t_1)} \cdot c_\alpha(s_n,a_n)|a_1 = a',\ s_2 = s',\ t_2 = t'\}$$

$$< v_\alpha(s') ,$$

which is a contradiction. Therefore

$$v_\pi(s) \geq E_{\pi,s}\{c_\alpha(s_1,a_1) + e^{-\alpha t_2} \cdot v_\alpha(s_2)\}$$

$$= T_\pi v_\alpha(s) .$$

This implies that

$$v_\pi(s) \geq \inf_{a\in A_s} \{c_\alpha(s,a) + \sum_{s'\in S} q_\alpha(s,a,s')\cdot v_\alpha(s')\} ,$$

for $s \in S$. But this holds for each π in $\mathcal{P}$, so

$$v_\alpha(s) \geq \inf_{a\in A_s} \{c_\alpha(s,a) + \sum_{s'\in S} q_\alpha(s,a,s')\cdot v_\alpha(s')\} ,$$

for $s \in S$. Combining this with the result above, the theorem follows.

4. <u>On the Existence of Stationary Optimal and Stationary ϵ-Optimal Policies</u>.

In this section the existence of stationary optimal and stationary

ϵ-optimal policies is investigated. It is important to distinguish between stationary optimal policies and optimal stationary policies. While the former policies are truly optimal, the latter ones are only optimal in the class of stationary policies. Conditions are given for optimal stationary policies to be stationary optimal policies.

Theorem 6: If π is a stationary policy such that $v_\alpha = T_\pi v_\alpha$, then π is optimal.

Proof: Since π is stationary, we obtain

$$v_\alpha = T_\pi^n v_\alpha ,$$

by applying T_π on both sides of $v_\alpha = T_\pi v_\alpha$ repeatedly. This implies that

$$v_\alpha = \lim_{n \to \infty} T_\pi^n v_\alpha = \lim_{n \to \infty} \{T_\pi^n \theta + Q_\pi^n v_\alpha\}$$

$$\geq \lim_{n \to \infty} T_\pi^n \theta + \lim_{n \to \infty} \inf Q_\pi^n v_\alpha$$

$$= v_\pi ,$$

by Lemma 4. Thus, π is optimal.

Corollary 7: If each A_s is finite, then there is a stationary optimal policy.

Proof: The existence of a policy π as in the theorem is in this case guaranteed by the optimality equation.

Corollary 8: If there is an optimal policy, then there is one which is stationary.

Proof: Let π be an optimal policy. From the proof of the optimality equation, $v_\pi \geq T_\pi v_\alpha$. Since π is optimal, we obtain $T_\pi v_\alpha \leq v_\alpha$. But $v_\alpha \leq T_{\pi'} v_\alpha$ for all $\pi' \in P$, so $v_\alpha = T_\pi v_\alpha$. Let π'' be the stationary policy such that $T_{\pi''} \equiv T_\pi$. By the theorem, π'' is optimal. Thus, there is a stationary optimal policy.

Theorem 9: If for each $s,s' \in S$,

$$E_{\pi,s}\{\sum_{n\in N} e^{-\alpha t_n} \cdot \delta(s_n,s')\}$$

is uniformly bounded, then an optimal stationary policy is a stationary optimal policy.

Proof: For each $s,s' \in S$, let $M(s,s')$ be an upper bound on

$$E_{\pi,s}\{\sum_{n\in N} e^{-\alpha t_n} \cdot \delta(s_n,s')\} .$$

Let $\epsilon > 0$ be given. Let v be a mapping from S into R_+ such that $v(s') > 0$ for $s' \in S$ and

$$\sum_{s'\in S} M(s,s')\cdot v(s') < \infty ,$$

where s is an element of S. Let π be a stationary policy such that

$$T_\pi v_\alpha \leq v_\alpha + \epsilon \cdot v .$$

Such a policy exists by the optimality equation. Applying T_π on both

sides of this inequality repeatedly, we obtain

$$T_\pi^n v_\alpha \leq v_\alpha + \epsilon \sum_{i < n} Q_\pi^i v ,$$

for $n \in N$. Letting n go to infinity in the expression above, we obtain

$$\lim_{n \to \infty} \inf T_\pi^n v_\alpha \leq v_\alpha + \epsilon \sum_{n \in N} Q_\pi^n v .$$

Now

$$\lim_{n \to \infty} \inf T_\pi^n v_\alpha = \lim_{n \to \infty} \inf\{T_\pi^n \theta + Q_\pi^n v_\alpha\}$$

$$= \lim_{n \to \infty} T_\pi^n \theta + \lim_{n \to \infty} \inf Q_\pi^n v_\alpha$$

$$\geq v_\pi ,$$

by Lemma 4. Thus

$$v_\pi \leq v_\alpha + \epsilon \sum_{n \in N} Q_\pi^n v ,$$

and in particular,

$$v_\pi(s) \leq v_\alpha(s) + \epsilon \sum_{n \in N} (Q_\pi^n v)(s)$$

$$\leq v_\alpha(s) + \epsilon \sum_{s' \in S} M(s,s') \cdot v(s') .$$

Let π' be an optimal stationary policy. From above,

$$v_{\pi'}(s) \leq v_\alpha(s) + \epsilon \sum_{s' \in S} M(s,s') \cdot v(s') .$$

But $\epsilon > 0$ is arbitrary and $\sum_{s' \in S} M(s,s')v(s)$ is finite, so

$v_{\pi'}(s) \leq v_{\alpha}(s)$. The argument can be repeated for each $s \in S$, so π' must be optimal.

Theorem 10: If for each $s' \in S$,

$$E_{\pi,s}\{\sum_{n\in N} e^{-\alpha t_n} \cdot \delta(s_n,s')\}$$

is uniformly bounded, then there are stationary ϵ-optimal policies for all $\epsilon > 0$.

Proof: For each $s' \in S$, let $M(s')$ be a bound on

$$E_{\pi,s}\{\sum_{n\in N} e^{-\alpha t_n} \cdot \delta(s_n,s')\} .$$

Following the proof of the previous theorem, we obtain

$$v_{\pi}(s) \leq v_{\alpha}(s) + \epsilon \sum_{s'\in S} M(s')\cdot v(s') ,$$

for some stationary policy π. Since $\epsilon > 0$ is arbitrary and

$$\sum_{s'\in S} M(s')\cdot v(s') < \infty ,$$

the theorem follows directly.

Corollary 11: If there is a $\beta < 1$ such that

$$E_{\pi,s}\{e^{-\alpha t_2}\} \leq \beta ,$$

for $s \in S$ and $\pi \in \Pi$, then there are stationary ϵ-optimal policies for arbitrarily small ϵ and every optimal stationary policy is a stationary optimal policy.

Proof: We only need to show that the conditions of the two previous theorems are satisfied. It is enough to show that

$$E_{\pi,s}\{\sum_{n\in N} e^{-\alpha t_n}\} \leq (1-\beta)^{-1} ,$$

for $s \in S$, $\pi \in \Pi$. But this follows from the proof of Theorem 1.

5. Policy Improvement

By the optimality equation, an optimal policy, π^*, must satisfy the equation

$$v_{\pi^*} = \inf_{\pi\in\Pi} T_\pi v_{\pi^*} .$$

A policy π^* which satisfies this equation is called unimprovable. An unimprovable policy need not be optimal, as the following example shows.

Example: Consider a discrete time Markov decision process with state space N_0 and action space $\{0,1\}$. If the system enters state 0, the process stops. In every other state there are two permissible actions, 0 and 1. If action 0 is taken in state $i (i > 0)$, an immediate cost a^i is incurred and the state 0 is entered. If action 1 is taken in state $i (i > 0)$, no immediate costs are incurred and the state $i + 1$ is entered. Let β be a given discount factor. If $a\beta > 1$, then the policy which always chooses action 0 is unimprovable, but it is not optimal.

If there is an $s \in S$ such that

$$v_{\pi'}(s) > \inf_{\pi\in\Pi} (T_\pi v_{\pi'})(s) ,$$

then π' is called improvable.

Theorem 12: If $\pi' \in \mathcal{P}$ and $\pi \in \mathcal{S}$ are such that $T_{\pi}v_{\pi'} \leq v_{\pi'}$, then $v_{\pi} \leq T_{\pi}v_{\pi'}$.

Proof: Applying T_{π} on both sides of $T_{\pi}v_{\pi'} \leq v_{\pi'}$ repeatedly, we obtain

$$T_{\pi}v_{\pi'} \geq T_{\pi}^{n}v_{\pi'} ,$$

for $n \in N$. Letting n go to infinity yields

$$T_{\pi}v_{\pi'} \geq \lim_{n \to \infty} \inf T_{\pi}^{n}v_{\pi'} = \lim_{n \to \infty} \inf\{T_{\pi}^{n}\theta + Q_{\pi}^{n}v_{\pi'}\}$$

$$= \lim_{n \to \infty} T_{\pi}^{n}\theta + \lim_{n \to \infty} \inf Q_{\pi}^{n}v_{\pi'}$$

$$\geq v_{\pi} ,$$

by Lemma 4. Thus, the theorem is proved.

This theorem may be useful for the development of a policy improvement procedure like that of Howard (1960). The problem is that one has to avoid convergence to a suboptimal solution.

6. Necessary and Sufficient Conditions for Optimality.

In Section 5, it was shown that an unimprovable policy need not always be optimal. Here, necessary and sufficient conditions for a policy to be optimal are presented. If v_{α} is known, then the optimality equation can be used to find out whether a given policy is optimal or not. If v_{α} is not known in advance, the following theorems may be more useful for proving that a given policy is optimal.

Theorem 13: Let S' be the set of s in S for which $v_{\alpha}(s)$ is

finite. Let $\mathcal{P}'$ be any subset of $\mathcal{P}$ such that for each π in $\mathcal{P}$ there is a π' in $\mathcal{P}$ such that $v_{\pi'} \leq v_{\pi}$.

If π^* is an unimprovable policy such that

$$\lim_{n \to \infty} (Q_{\pi}^{n} v_{\pi^*})(s) = 0 \quad \text{for} \quad s \in S', \ \pi \in \mathcal{P}' ,$$

then π^* is optimal.

Proof: We first prove that $v_{\pi^*} \leq T_{\pi}^{n} v_{\pi^*}$ for $n \in N$ and $\pi \in \mathcal{P}$. This clearly holds for $n = 1$. Now

$$(T_{\pi}^{n} v_{\pi^*})(s) = E_{\pi,s}\{ \sum_{i \leq n} e^{-\alpha t_i} \cdot c_{\alpha}(s_i,a_i) + e^{-\alpha t_{n+1}} \cdot v_{\pi^*}(s_{n+1})\}$$

$$= E_{\pi,s}\{ \sum_{i \leq n} e^{-\alpha t_i} \cdot c_{\alpha}(s_i,a_i)$$

$$+ e^{-\alpha t_n} \cdot E_{\pi,s}\{c_{\alpha}(s_n,a_n) + e^{-\alpha(t_{n+1}-t_n)} \cdot v_{\pi^*}(s_{n+1}) | t_n, s_n\}\} .$$

Furthermore,

$$E_{\pi,s}\{c_{\alpha}(s_n,a_n) + e^{-\alpha(t_{n+1}-t_n)} \cdot v_{\pi^*}(s_{n+1}) | t_n, s_n\} \geq v_{\pi^*}(s_n) ,$$

since π^* is unimprovable. Therefore

$$(T^{n} v_{\pi^*})(s) \geq E_{\pi,s}\{ \sum_{i \leq n} e^{-\alpha t_i} \cdot c_{\alpha}(s_i,a_i) + e^{-\alpha t_n} \cdot v_{\pi^*}(s_n)\}$$

$$\geq E_{\pi,s}\{ \sum_{i \leq n-1} e^{-\alpha t_i} \cdot c_{\alpha}(s_i,a_i) + e^{-\alpha t_{n-1}} \cdot v_{\pi^*}(s_{n-1})\}$$

$$\vdots$$

$$\geq v_{\pi^*}(s) ,$$

for $s \in S$. This implies that

$$v_{\pi^*} \leq \lim_{n \to \infty} T_{\pi}^{n} v_{\pi^*} = \lim_{n \to \infty} \{T_{\pi}^{n}\theta + Q_{\pi}^{n} v_{\pi^*}\}$$

$$= \lim_{n \to \infty} T_{\pi}^{n}\theta + \lim_{n \to \infty} Q_{\pi}^{n} v_{\pi^*}$$

$$= v_{\pi} ,$$

by the last condition of the theorem. Thus, π^* is optimal.

Corollary 14: If π^* is an unimprovable policy such that

$$\sum_{s' \in S} E_{\pi,s}\{e^{-\alpha t_n} \cdot \delta(s', s_n)\} \cdot |v_{\pi^*}(s')| < \infty, \quad \text{for } s \in S', \pi \in \mathcal{P}' ,$$

then π^* is optimal.

Proof: It is sufficient to recognize that

$$\sum_{s' \in S} E_{\pi,s}\{e^{-\alpha t_n} \cdot \delta(s', s_n)\} \cdot |v_{\pi^*}(s')|$$

$$= \sum_{n \in N} (Q_{\pi}^{n} |v_{\pi^*}|)(s)$$

$$\geq \sum_{n \in N} |(Q_{\pi}^{n} v_{\pi^*})(s)| ,$$

for $s \in S'$ and $\pi \in \mathcal{P}'$. Thus the conditions of the theorem hold, and the corollary follows.

Corollary 15: If π^* is an unimprovable policy such that its value-function is bounded, then π^* is optimal.

Proof: Let M be an upper bound on $|v_{\pi^*}(s)|$. Then

$$\lim_{n \to \infty} \left| E_{\pi,s}\{e^{-\alpha t_n} \cdot v_{\pi^*}(s_n)\} \right|$$

$$= \lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n} \cdot |v_{\pi^*}(s)|\}$$

$$\leq M \cdot \lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n}\} = 0 ,$$

by Assumption 3. The corollary now follows from the theorem.

Theorem 16: Suppose that there is an optimal policy, π. Then a policy π^* is optimal if and only if

$$\lim_{n \to \infty} (Q_{\pi}^{n} v_{\pi^*})(s) = 0, \quad \text{for} \quad s \in S' .$$

Proof: The if part of the theorem follows from Theorem 13 by letting $\mathcal{P}' = \{\pi\}$. The only if part is proven as follows. Suppose that π^* is optimal. Then

$$(Q_{\pi}^{n} v_{\pi^*})(s) = (Q_{\pi}^{n} v_{\pi})(s) = v_{\pi}(s) - (T_{\pi}^{n}\theta)(s) ,$$

for $s \in S'$, $n \in N$. This implies that

$$\lim_{n \to \infty} (Q_{\pi}^{n} v_{\pi^*})(s) = \lim_{n \to \infty} \{v_{\pi}(s) - (T_{\pi}^{n}\theta)(s)\} = 0 ,$$

for $s \in S'$. This completes the proof of the theorem.

7. Norms and Contraction Mappings.

It may sometimes be more convenient to work with norms and contraction mappings. Denardo (1967) did this, and developed an elegant analysis. Recently, Lippman (1975) used these concepts.

As before, let $\mathbb{1}$ be the function from S into R with value 1 everywhere. Let $\|\cdot\|$ be a norm on B such that

(a) $\| \underline{1} \| = 1$,

(b) $\|u\| \leq \|v\|$ if $0 \leq u \leq v$.

The sup norm, given by

$$\|v\| = \sup_{s \in S} |v(s)| ,$$

is such a norm. Lippman (1975) has considered other norms.

A mapping T from B into B is called a contraction mapping if there is a $\beta < 1$ such that

$$\|Tv\| \leq \beta \|v\| ,$$

for $v \in B$. Denardo's n-stage contraction condition is as follows. There is an $n \in N$ and a $\beta < 1$ such that

$$\|Q_\pi^n v\| \leq \beta \|v\| ,$$

for $v \in B$ and $\pi \in \mathcal{P}$. We weaken the n-stage contraction condition so that it reads as follows. For each $v \geq 0$, there is an n N and a $\beta < 1$ such that

$$\|Q_\pi^n v\| \leq \beta \|v\| ,$$

for all π in $\mathcal{P}$.

Lemma 17: If there is an $n \in N$ and a $\beta < 1$ such that

$$E_{\pi,s}\{e^{-\alpha t_n}\} \leq \beta ,$$

for $s \in S$ and $\pi \in \mathcal{P}$, then the sup norm satisfies the n-stage contraction condition.

Proof: We have

$$\|Q_\pi^n v\| = \sup_{s\in S}|(Q_\pi^n v)(s)| = \sup_{s\in S}|E_{\pi,s}\{e^{-\alpha t_n} v(s_n)\}|$$

$$\leq \sup_{s\in S} E_{\pi,s}\{e^{-\alpha t_n} \cdot \sup_{s\in S}|v(s')|\}$$

$$= \|v\| \cdot \sup_{s\in S} E_{\pi,s}\{e^{-\alpha t_n}\} ,$$

and the lemma follows.

Let $\rho(\cdot,\cdot)$ be a metric on $B \times B$ such that for u,v in B, $\rho(u,v) = \|w\|$, where

$$w(s) = \begin{cases} u(s) - v(s), & \text{if } u(s) < \infty \text{ or } v(s) < \infty , \\ 0, & \text{if } u(s) = v(s) = \infty . \end{cases}$$

Theorem 18: If $\|\cdot\|$ satisfies the n-stage contraction condition, then a policy π^* is optimal if and only if π^* is unimprovable and $\rho(v_{\pi^*}, v_\alpha) < \infty$.

Proof: The only if part of the theorem is trivial. We now prove the if part. Let w be such that

$$w(s) = \begin{cases} v_{\pi^*}(s) - v_\alpha(s), & \text{if } v_\alpha(s) < \infty \text{ or } v_{\pi^*}(s) < \infty , \\ 0, & \text{if } v_{\pi^*}(s) = v_\alpha(s) = \infty . \end{cases}$$

Let $n \in N$ and $\beta < 1$ be as in the contraction condition. Let $\epsilon > 0$ be given, and let π be a stationary policy such that

$$T_\pi v_\alpha \leq v_\alpha + (\epsilon/n)\cdot \mathbb{1} .$$

π^* is unimprovable, so $v_{\pi^*} \leq T_\pi v_{\pi^*}$. This implies that

$$w \leq Q_\pi w + (\epsilon/n)\cdot \mathbb{1} ,$$

since $w \geq \theta$. Applying Q_π to both sides of this inequality repeatedly yields

$$w \leq Q_\pi^n w + (\epsilon/n) \sum_{i < n} Q_\pi^i \mathbb{1}$$

$$\leq Q_\pi^n w + \epsilon \cdot \mathbb{1} ,$$

since $Q_\pi \mathbb{1} \leq \mathbb{1}$. Taking the norm of the functions on both sides of the inequality yields

$$\|w\| \leq \|Q_\pi^n w + \epsilon \cdot \mathbb{1}\|$$

$$\leq \|Q_\pi^n w\| + \epsilon \|\mathbb{1}\|$$

$$\leq \beta\|w\| + \epsilon ,$$

by the contraction condition. But $\epsilon > 0$ is arbitrary, $\beta < 1$ and $\|w\| < \infty$. This implies that $\|w\| = 0$, and thus $v_{\pi^*} = v_\alpha$.

Corollary 19: If π^* is an unimprovable policy such that $v_{\pi^*}(s) - v_\alpha(s)$ is bounded, and if the condition of Lemma 17 is satisfied, then π^* is optimal.

Notice the similarity of this corollary and Corollary 15.

Corollary 20: If $\|\cdot\|$ satisfies the n-stage contraction condition, if v_α is bounded below, and if π^* is an unimprovable policy such that $\|v_{\pi^*}\| < \infty$, then π^* is optimal.

Proof: Let $-M(M \geq 0)$ be a lower bound on $v_\alpha(s)$. Let w be as in the proof of the theorem. Then

$$0 \leq w \leq v_{\pi^*} + M \cdot \mathbb{1} ,$$

so

$$\|w\| \leq \|v_{\pi^*} + M \cdot \mathbb{1}\| \leq \|v_{\pi^*}\| + M < \infty .$$

CHAPTER 4

OPTIMAL CONTROL OF QUEUEING SYSTEMS

There has been a considerable interest in the control of queueing systems in the last decade. Often the control problems have been formulated in the framework of semi-Markov decision processes. The existence of certain simple and intuitive optimal policies have been proven for many different queueing systems. For a brief (but excellent) survey of the literature in this area, see Gross and Harris (1974, pp. 364-380).

In this chapter, three aspects of the control of queueing systems are considered. In Section 1, the formulation of queueing control problems is discussed. Section 2 elaborates upon two general approaches to the solution of queueing control problems. In Section 3, four different methods for proving the optimality of an unimprovable policy are developed.

1. Formulation of Queueing Control Problems.

The formulation of queueing control problems plays an important role in the solution of these problems. Sometimes, a queueing control problem may be formulated in two different but equivalent ways, where only one is amenable to analysis. Special queueing control problems may have special desirable formulations. But since a general formulation of queueing control problems may yield a better perspective, we shall now briefly describe the various components of a controllable queueing system.

A queueing system consists of an input source, a queue and a service mechanism. The input source generates customers which need certain services

provided by the service mechanism. A customer generated by the input source is said to arrive at the queueing system. The times between two consecutive arrivals are the interarrival times. On arrival, a customer either is given service immediately or is placed in the queue of customers waiting to be served. There may be several customer classes, reflecting the special needs of the customers. The service mechanism may consist of one or several service facilities, each of which has a certain number of servers. When the customers have received their service(s), they leave the system.

The control of queueing systems can take various forms. Sometimes, the arrival rate may be adjusted dynamically. Other times, the service rate(s) or the number of active servers may be controlled. A third possibility is to control the order in which the customers are given service.

There are various costs that may need to be considered when analyzing queueing systems. For example, there may be a service cost which is incurred each time a customer is served. If the server(s) can be turned on and off, there may be start-up and shut-down costs when the server(s) are turned on and turned off, respectively. There may be an idling cost which is incurred at a positive and constant rate for each server when he is not giving service or performing other useful duties. There may be a customer holding cost which is incurred at a rate which is a function of the number of customers in the system.

There may, of course, be many other types of controls and costs than those which have been mentioned here. But surprisingly many of the queueing control problems which have been considered in the literature fit the above description.

By formulating a queueing control problem as a semi-Markov decision process, the theory for such processes may be used in developing a solution procedure or to prove that a given policy is optimal (or not). The formulation is usually quite straightforward. One only has to define the state of the system and the decision epochs. The state space, the set of action spaces, the law of motion and the cost function of the semi-Markov decision process are then determined by the specification of the queueing system.

The definition of the state of the system is crucial. The state must characterize the queueing system completely at each decision epoch. Since a queueing system consists of an input source, a queue and a service mechanism, one may define the state of the input source, the state of the queue and the state of the service mechanism. The state of the system is then given by these three states. The state space of the system may be defined as the Cartesian product of the state spaces of the input source, the queue and the service mechanism, respectively.

The state space of a queueing system is often countable. If the input source, the queue and the service mechanism all have countable state spaces, then the state of the system is countable.

Consider the state space of the queue. Suppose that there is a countable number of customer classes. If the state of the queue is defined as the vector whose i^{th} component indicates the number of customers in class i (for each $i \in N$), then the state space of the queue is countable. This follows from the fact that there are only a finite number of customers in the queue at any given time.

Consider the state space of the service mechanism. One case is the system which can be controlled by turning serving on or off. For

this case, if there is a countable number of servers, and if the state of the service mechanism is defined as the vector whose i^{th} component indicates whether the i^{th} server is on or off (for each $i \in N$), then the state space of the service mechanism is countable. For a more general case, suppose now that the service rate of each server may be adjusted to a countable number of levels. Also suppose that there are a countable number of servers and that the service rate is only non-zero for a finite number of servers at any given point in time. If the state of the service mechanism is defined as the vector whose i^{th} component indicates the level of the service rate of the i^{th} server, then the state space is still countable.

The definition of the decision epochs is also crucial. As mentioned before, the state of the system must characterize the queueing system completely at each decision epoch. The most natural way to define the decision epochs is by letting them be the epochs when the state of the system changes. If the state of the system (as it happens to be defined) does not characterize the queueing system completely at each of these decision epochs, one can try to eliminate some of the decision epochs.

Sometimes it may be desirable to have the decision epochs equally spaced in time. In this case, the decision epochs are determined by specifying the length of time between two consecutive decision epochs. Magazine (1971) used this approach. Other times, it may be desirable to define the decision epochs such that the times between two consecutive decision epochs are independent and identically distributed random variables. Lippman (1975) used this approach. Both of these ways of defining the decision epochs are motivated by a certain solution method which will be elaborated upon in the next section.

2. Analytical Solution Methods.

A large variety of queueing control problems have been successfully analyzed by a number of investigators. Their successes have to some extent depended on the special features of the problems they considered. But many of the queueing problems also have much in common. Therefore, there is some basis for developing general approaches for solving them. Prabhu and Stidham (1973) attempted to develop a unified view of the different approaches that have been used previously.

If the state and action spaces are finite, then there are well-known (policy improvement, policy iteration) algorithms for finding an optimal policy. But in the context of queueing systems, one is often more interested in showing that there is an optimal policy of a simple and intuitive form. As a by-product of this, one may perhaps develop especially efficient algorithms for finding an optimal policy. Two general approaches for analyzing queuing systems will now be presented.

The first approach consists of solving the problem for one period (stage) and then extending the results to arbitrarily many periods by an inductive argument. This approach was initially used for solving inventory problems (e.g. by Iglehart (1963)). Because of the similarity between queueing and inventory problems, the approach was later adopted by queueing theoreticians. McGill (1969) used the approach in his analysis of the M/M/c queueing system with controllable servers. A full development of this approach can be found in Porteus (1975b).

This approach has two advantages. First, the one-period problem is usually easier to analyze than the infinite period problem. A successful analysis solves both the finite and infinite horizon problems.

However, this approach of first solving the one-period problem can also have its disadvantages. In fact, for many queueing problems, the one-period problem is rather meaningless. One reason is that the length of the first period may not be nearly the same for different start-states and different actions. Furthermore, many important costs may be neglected in the one-period problem (e.g., switching costs). Nevertheless, the approach is still attractive for many problems.

The second approach consists of restricting one's search for an optimal policy to a small class of stationary policies (hopefully not excluding the optimal policy) and then proving that the policy which is optimal in this class is also optimal among all policies. To prove that a policy believed to be optimal is indeed optimal among all policies, one usually only has to prove that the policy is unimprovable. This approach has been used by, among others, Reed (1974a), (1974b).

This approach has the advantage that it usually only requires the analysis of relatively simple stationary policies. If one can obtain an explicit expression for the value functions of these policies, then it is usually a simple matter to prove when one of these policies is unimprovable (and thus probably optimal). Even if such explicit results cannot be obtained, the approach may still be used with success (e.g., see Orkenyi (1976)).

The disadvantage of the approach lies in the fact that an unimprovable policy need not necessarily be optimal. In the previous chapters, several conditions for an unimprovable policy to be optimal were given. For example, when discounting is used, it was shown that if the value function of the unimprovable policy is bounded, then the policy is optimal.

But queueing control problems are often characterized by giving rise to unbounded value functions. This is often due to the holding costs being unbounded. In the next section, it is shown how this problem can be solved.

3. Solutions to the Problem of Unbounded Costs.

We now consider the problem of unbounded costs with discounting, and develop four different methods for proving that an unimprovable policy is optimal. The assumptions of chapter 3 are retained here.

3.1 A Reformulation.

Perhaps the easiest way to solve the problem of unbounded costs is by reformulating the cost structure of the system under consideration in such a way that the costs become bounded. There is, however, no single receipe for doing this. Different problems may require different reformulations. Here, an idea of Bell (1971) is generalized.

For the sake of simplicity, suppose that the expected discounted cost excluding the cost due to holding customers in the system is bounded. Also suppose that there are m customer classes and that a holding cost is incurred at a rate which is a given function, h, of the number of customers present in each customer class. Define the state of the queue as indicated in Section 1.

For each $n \in N$, let t_n denote the time of the n^{th} change in the state of the queue and let y_n denote the state of the queue immediately after the change. Without loss of generality, assume that $t_1 = 0$. For each policy π and state s, let $v_\pi^h(s)$ denote the expected discounted holding cost, given that the policy π is used and that the

start state is s. Clearly

$$v_{\pi}^{h}(s) = E_{\pi,s}\{\sum_{n\in N}\int_{t_n}^{t_{n+1}} h(y_n)e^{-\alpha t}dt\}$$

$$= \frac{1}{\alpha}h(y_1) + E_{\pi,s}\{\sum_{n\in N}\frac{1}{\alpha}(h(y_{n+1}) - h(y_n))e^{-\alpha t_{n+1}}\},$$

for each $s \in S$ and $\pi \in \mathcal{P}$.

Now, reformulate the holding cost structure such that at each time $t_n(n > 1)$, the holding cost

$$x_n = \frac{1}{\alpha}(h(y_n) - h(y_{n-1}))$$

is incurred. Formally, we choose to include the cost x_n in the costs incurred in the period from t_{n-1} to $t_n(n > 1)$. For each start-state s and policy π used, the expected discounted holding cost becomes

$$v_{\pi}^{h}(s) - \frac{1}{\alpha}h(y_1) .$$

Thus, the problem before the reformulation is equivalent to the problem after the reformulation with regard to optimal policies.

Assume that the number of customers in each customer class only can change by one at a time and that changes in different customer classes cannot occur simultaneously. Let Y denote the state space of the queue, and for each $i(\leq m)$, let ω_i denote the m-vector whose components are all zero except for the i^{th} one which is equal to one. We can now state the following theorem.

Theorem 1: If for each policy π,

$$E_{\pi}\{\sum_{n\in N} e^{-\alpha t_n}\}$$

is uniformly bounded, and if there is an $M < \infty$ such that

$$|h(y + u_i) - h(y)| \leq M$$

for $1 \leq i \leq m$ and $y \in Y$, then every unimprovable policy is optimal.

Proof: Under the conditions of the theorem, the expected discounted holding cost after the reformulation is bounded. Therefore, any policy which is unimprovable for the problem after the reformulation is optimal for that problem. But the optimal policies are the same for both problems. The unimprovable policies are also the same for both problems. Therefore, we conclude that a policy which is unimprovable for the original problem is also optimal.

Example (The M/G/1 queueing system with removable server):

Excluding the policies which turn the server on and off repeatedly at a decision epoch, the expected discounted cost excluding those due to holding customers in the system is bounded. Let λ be the arrival rate of the customers, and let $\omega(< 1)$ be the Laplace transform of the service times (with its parameter being equal to the interest rate α).

Let $\{t'_n\}_{n\in N}$ be the sequence of times when customers arrive, and let $\{t''_n\}_{n\in N}$ be the sequence of times when customers depart. It can easily be shown that for each policy π used and each start-state s,

$$E_{\pi,s}\{\sum_{n\in N} e^{-\alpha t'_n}\} = 1 + \frac{\lambda}{\alpha},$$

and

$$E_{\pi,s}\{\sum_{n\in N} e^{-\alpha t''_n}\} \leq \frac{1}{1-\omega} < \infty .$$

Since $\{t_n\}_{n\in N}$ is a subsequence of $\{t'_n\}_{n\in N} \cup \{t''_n\}_{n\in N}$, the first condition of the theorem holds.

If the slope of h is bounded (in this case h is a function of one variable), then the second condition of the theorem holds. Thus, if the slope of the holding cost function is bounded, then every unimprovable policy is optimal. This is just the assumption made by Blackburn (1971) when he considered the convex holding cost model.

3.2 Comparison with the Policy which Shuts Down the System.

Assume as before that the customer holding cost is incurred at a rate $h(y_n)$ in each interval (t_n, t_{n+1}). Also assume that h is such that

$$0 \leq h(x) \leq h(y)$$

for $x \leq y$ and $x \in Y$, $y \in Y$.

Assume that the system can be shut down at any decision epoch and that the shut-down cost is bounded uniformly from above. Let π_0 denote the policy which always shuts the system down (or leaves it off). Assume that when the policy π_0 is used the total number of customers present in each customer class is at a maximum at all times for any given start-state.

Theorem 2: If π^* is an unimprovable policy such that, for each $s \in S$,

$$v_{\pi^*}(s) \leq v_{\pi_0}(s) < \infty ,$$

then π^* is optimal.

Proof: By Theorem 13 in Chapter 3, we only need to show that

$$\lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n} \cdot v_{\pi^*}(s_n)\} = 0$$

for each $s \in S$ and $\pi \in \mathcal{P}$. Here $\{t_n\}_{n \in N}$ is the sequence of the times of the decision epochs.

For each $s \in S$, let $R(s)$ denote the expected discounted shut-down cost when the system is in state s and the policy π_0 is used. For each $\pi \in \mathcal{P}$, $s \in S$ and $t \in R$, let $x_\pi(s,t)$ denote the discounted holding cost incurred from time t onward (the discounting starting at time 0), given that the start-state is s and that the policy π_0 is used. It follows from our assumptions that

$$x_\pi(s,t) \leq x_{\pi_0}(s,t), \quad \text{for } t \in R,\ s \in S,\ \pi \in \mathcal{P}.$$

Now

$$v_{\pi_0}(s) = R(s) + E\{x_{\pi_0}(s,0)\}, \quad \text{for } s \in S,$$

so

$$E\{x_{\pi_0}(s,0)\} < \infty, \quad \text{for } s \in S.$$

For each $\pi \in \mathcal{P}$ and $s \in S$, let $\{t_n(\pi,s)\}_{n \in N}$ be the sequence of the times of the decision epochs, given that the start-state is s and that the policy π is used.

Choose a $\pi \in \mathcal{P}$, and for each $n \in N$, let π_n be the policy which follows π until the n^{th} decision epoch and then shuts down the system. Then

$$E_{\pi,s}\{e^{-\alpha t_n} v_{\pi_0}^h(s_n)\} = E\{x_{\pi_n}(s,t_n(\pi,s))\}$$

$$= E\{1_{\{t_n(\pi,s) \leq t\}} \cdot x_{\pi_n}(s,t_n(\pi,s))\}$$

$$+ E\{1_{\{t_n(\pi,s) > t\}} \cdot x_{\pi_n}(s,t_n(\pi,s))\}$$

$$\leq E\{1_{\{t_n(\pi,s) \leq t\}} \cdot x_{\pi_0}(s,0)\}$$

$$+ E\{1_{\{t_n(\pi,s) > t\}} \cdot x_{\pi_0}(s,t_n(\pi,s))\}$$

$$\leq E\{1_{\{t_n(\pi,s) \leq t\}} \cdot x_{\pi_0}(s,0)\} + E\{x_0(s,t)\}$$

for $n \in \mathbb{N}$, $t \in R$ and $s \in S$. Here, we have used the fact that

$$x_{\pi}(s,t) \leq x_{\pi_0}(s,t), \quad \text{for} \quad t \in R, \ s \in S, \ \pi \in \mathcal{P}$$

and

$$x_{\pi}(s,t) \leq x_{\pi}(s,t'), \quad \text{for} \quad t \leq t', \ t,t' \in R, \ s \in S, \ \pi \in \mathcal{P},$$

and

$$x_{\pi}(s,t) \geq 0, \quad \text{for} \quad t \in R, \ s \in S, \ \pi \in \mathcal{P}.$$

By Lebesgue's bounded convergence theorem,

$$\lim_{n \to \infty} E\{1_{\{t_n(\pi,s) \leq t\}} \cdot x_{\pi_0}(s,0)\} = 0, \quad \text{for} \quad s \in S,$$

since

$$E\{x_{\pi_0}(s,0)\} < \infty,$$

and

$$\lim_{n \to \infty} P_{\pi,s}\{t_n \leq t\} = 0.$$

Thus

$$\lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n} \cdot v^h_{\pi_0}(s_n)\} \leq E\{x_0(s,t)\}, \quad \text{for} \quad t \in R, \ s \in S.$$

But

$$\lim_{t \to \infty} E\{x_0(s,t)\} = 0, \quad \text{for} \quad s \in S ,$$

since

$$E\{x_0(s,0)\} = \lim_{t \to \infty} E_{\pi_0,s}\{\int_0^t e^{-\alpha t} \cdot h(y_t)dt\} < \infty, \quad \text{for} \quad s \in S ,$$

where y_t denotes the state of the queue at time t. Therefore

$$\lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n} \cdot v_{\pi_0}(s_n)\}$$

$$= \lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n} \cdot R(s_n)\} + \lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n} \cdot v^h_{\pi_0}(s_n)\}$$

$$\leq \lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n} \cdot R(s_n)\} ,$$

for $s \in S$. Let M be a finite upper bound on $R(s)$. Then

$$\lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n} \cdot v_{\pi_0}(s_n)\} \leq M \cdot \lim_{n \to \infty} E_{\pi,s}\{e^{-\alpha t_n}\}$$

$$= 0, \quad \text{for} \quad s \in S ,$$

since

$$\lim_{n \to \infty} P_{\pi,s}\{t_n \leq t\} = 0, \quad \text{for} \quad t \in R, \; s \in S .$$

This completes the proof.

Q.E.D.

Example (The M/G/1 queueing system with removable server):

The state of the system is defined as a pair of integers (i,j), where i denotes the number of customers in the system and

$$j = \begin{cases} 0, & \text{if the server is off} \\ 1, & \text{if the server is on .} \end{cases}$$

It is easy to find that

$$v_{\pi_0}(i,j) = \begin{cases} \sum_{k \in N_0} \left(\frac{\lambda}{\lambda+\alpha}\right)^k \cdot h(i+k), & \text{for } j = 0,\ i \in N_0 , \\ R_2 + \sum_{k+N_0} \left(\frac{\lambda}{\lambda+\alpha}\right)^k \cdot h(i+k), & \text{for } j = 1,\ i \in N_0 . \end{cases}$$

Therefore, if π is an unimprovable policy such that

$$v_{\pi}(i,j) \leq v_{\pi_0}(i,j), \quad \text{for } j \in \{0,1\},\ i \in N_0 ,$$

and if

$$\sum_{i \in N_0} \left(\frac{\lambda}{\lambda+\alpha}\right)^i h(i) < \infty,$$

then π is optimal.

3.3 Comparison with the Policy which Minimizes the Expected Discounted Holding Cost.

Suppose that there is a policy which minimizes the expected discounted holding cost, and let π_0 denote such a policy. For each $\pi \in \mathcal{P}$ and $s \in S$, let $v_{\pi}^{nh}(s)$ denote the expected discounted cost excluding the holding costs, given that the start-state is s and that the policy π is used. Then

$$v_{\pi}(s) = v_{\pi}^{h}(s) + v_{\pi}^{nh}(s), \quad \text{for } s \in S,\ \pi \in \mathcal{P} .$$

Let ρ be a metric defined as in Chapter 3. Let Λ be the binary operator such that

$$x \bigwedge y = \min(x,y), \quad \text{for} \quad x \in R,\ y \in R\ .$$

We are now ready to state the following theorem.

Theorem 3: If π^* is an unimprovable policy such that $v_{\pi^*} \leq v_{\pi_0}$ and, in addition,

$$\rho(v^{nh}_{\pi_0},\ v^{nh}_{\pi_0} \bigwedge v^{nh}_{\pi}) < \infty, \quad \text{for} \quad \pi \in \mathcal{P}\ ,$$

then π^* is optimal.

Proof: By Theorem 18 of Chapter 3, we only have to show that

$$\rho(v_{\pi^*},\ v_{\pi^*} \bigwedge v_{\pi}) < \infty, \quad \text{for} \quad \pi \in \mathcal{P}\ .$$

But

$$\begin{aligned}
\rho(v_{\pi^*},\ v_{\pi^*} \bigwedge v_{\pi}) &\leq \rho(v_{\pi_0},\ v_{\pi_0} \bigwedge v_{\pi}) \\
&= \rho((v^{h}_{\pi_0} + v^{nh}_{\pi_0}),\ (v^{h}_{\pi_0} + v^{nh}_{\pi_0}) \bigwedge (v^{h}_{\pi} + v^{nh}_{\pi})) \\
&\leq \rho((v^{h}_{\pi_0} + v^{nh}_{\pi}),\ (v^{h}_{\pi_0} + v^{nh}_{\pi}) \bigwedge (v^{h}_{\pi_0} + v^{nh}_{\pi})) \\
&\leq \rho(v^{nh}_{\pi_0},\ v^{nh}_{\pi_0} \bigwedge v^{nh}_{\pi}) \\
&< \infty, \quad \text{for} \quad \pi \in \mathcal{P}\ .
\end{aligned}$$

Q.E.D.

Example (The M/G/1 queueing system with removable server):

In this case, let $\|\cdot\|$ be the sup norm. Excluding those policies which turn the server on and off repeatedly at a decision epoch, then the contraction condition of Section 7 of Chapter 3 is satisfied. For

any $\pi \in \mathcal{P}$,

$$\rho(v^{nh}_{\pi_0},\ v^{nh}_{\pi_0} \bigwedge v^{nh}_{\pi}) = \|v^{nh}_{\pi_0} - v^{nh}_{\pi_0} \bigwedge v^{nh}_{\pi}\|$$

$$= \sup_{s \in S} |v^{nh}_{\pi_0}(s) - v^{nh}_{\pi_0}(s) \bigwedge v^{nh}_{\pi}(s)|$$

$$< \infty \ .$$

We conclude that if π^* is an unimprovable policy such that $v_{\pi^*} \leq v_{\pi_0}$, then π^* is optimal.

3.4 Comparison with a Policy which Minimizes the Expected Discounted Holding Cost until a Finite Set of States is Reached.

We now generalize the result of Section 3. This time, let π_0 denote a policy which minimizes the expected discounted holding cost incurred until a given, finite set of states is reached. Assume that $v^h_{\pi_0}$ is finite-valued. Let ρ be defined as before.

Theorem 4: If π^* is an unimprovable policy such that $v_{\pi^*} \leq v_{\pi_0}$ and, in addition,

$$\rho(v^{nh}_{\pi_0},\ v^{nh}_{\pi_0} \bigwedge v^{nh}_{\pi}) < \infty, \quad \text{for} \quad \pi \in \mathcal{P} \ ,$$

then π^* is optimal.

Proof: Since π_0 minimizes the expected discounted holding cost incurred until a given, finite set of states is reached, and since $v^h_{\pi_0}$ is finite-valued, there must exist an $M < \infty$ such that

$$v^h_{\pi_0}(s) \leq v^h_{\pi}(s) + M, \quad \text{for} \quad s \in S,\ \pi \in \mathcal{P} \ .$$

Now

$$\rho(v_{\pi^*}, v_{\pi^*} \bigwedge v_\pi) \leq \rho(v_{\pi_0}, v_{\pi_0} \bigwedge v_\pi)$$

$$= \rho(v_{\pi_0}, v_{\pi_0} \bigwedge (v_\pi^h + v_\pi^{nh}))$$

$$\leq \rho(v_{\pi_0}, v_{\pi_0} \bigwedge (v_{\pi_0}^h - M\theta + v_\pi^{nh}))$$

$$\leq \rho(v_{\pi_0}^{nh}, v_{\pi_0}^{nh} \bigwedge (v_\pi^{nh} - M\theta))$$

$$\leq \rho(v_{\pi_0}^{nh}, v_{\pi_0}^{nh} \bigwedge v_\pi^{nh}) + M$$

$$< \infty, \quad \text{for} \quad \pi \in \mathcal{P}.$$

Theorem 18 of Chapter 3 now implies that π^* is optimal.

Q.E.D.

Example (The M/G/1 queueing system with removable server):

Let $\mathcal{Y}$ be the set of policies such that each policy in $\mathcal{Y}$ always turns the server on (or keeps him on) when the number of customers in the system is sufficiently large. It is easy to show that, for each $\pi \in \mathcal{Y}$, either $v_\pi^h(s)$ is finite-valued for each s in S or it is infinite-valued for each s in S. In the latter case, all policies may be regarded as optimal. Therefore, we now focus on the former case where $v_\pi^h(s)$ is always finite-valued.

Clearly, π_0 in the theorem may be any policy in $\mathcal{Y}$. As before,

$$\rho(v_{\pi_0}^{nh}, v_{\pi_0}^{nh} \bigwedge v_\pi^{nh}) < \infty, \quad \text{for} \quad \pi \in \mathcal{P}.$$

Therefore, we conclude that if π^* is an unimprovable policy in $\mathcal{Y}$, then π^* is also optimal.

REFERENCES

Bell, C. (1971), "Characterization and Computation of Optimal Policies for Operating an M/G/1 Queueing System with Removable Server," Oper. Res. 19, 208-218.

Bellman, R. (1957), Dynamic Programming, Princeton University Press.

Blackburn, J. (1971), "Optimal Control of Queueing Systems with Intermittent Service," Tech. Rep. No. 8, Department of Operations Research, Stanford University.

Blackwell, D. (1962), "Discrete Dynamic Programming," Ann. Math. Stat. 33, 719-726.

Blackwell, D. (1965), "Discounted Dynamic Programming," Ann. Math. Stat. 36, 226-235.

Dantzig, G. B. and Wolfe, P. (1962), "Linear Programming in a Markov Chain," Oper. Res. 10, 707-710.

Denardo, E. (1967), "Contraction Mappings in the Theory Underlying Dynamic Programming," SIAM Rev. 9, 165-177.

Denardo, E. V. (1970a), "On Linear Programming in a Markov Decision Problem," Mgt. Sci. 16, 281-288.

Denardo, E. V. (1970b), "Computing Bias-Optimal Policies in Discrete and Continuous Markov Decision Problems," Oper. Res. 18, 279-289.

Denardo, E. V. (1971), "Markov Renewal Programs with Small Interest Rates," Ann. Math. Stat. 42, No. 2, 477-496.

Denardo, E. V. and Fox, B. L. (1968), "Multichain Markov Renewal Programs," SIAM J. Appl. Math. 16, 468-487.

Derman, C. (1966), "Denumerable State Markovian Decision Processes - Average Cost Criterion," Ann. Math. Stat. 37, 1545-1554.

Derman, C. (1970), Finite State Markovian Decision Processes, Academic Press.

D'Epenoux, F. (1960), "Sur un Probleme de Production et de Stockage dans L'a Leatoire," Rev. Francaise Informat. Recherche Operationelle 14, 3-16 [English Transl.: Mgt. Sci. 10, 98-108 (1963).]

Gross, D. and Harris, C. M. (1974), Fundamentals of Queueing Theory, John Wiley and Sons, Inc.

Harrison, M. (1972), "Discrete Dynamic Programming with Unbounded Rewards," Ann. Math. Stat. 43, 636-644.

Heyman, D. (1968), "Optimal Operating Policies for M/G/1 Queueing Systems," Oper. Res. 16, 362-382.

Hordijk, A. (1974a), Dynamic Programming and Markov Potential Theory, Matematisch Centrum.

Hordijk, A. (1974b), Convergent Dynamic Programming," Tech. Rep. No. 28, Department of Operations Research, Stanford University.

Howard, R. (1960), Dynamic Programming and Markov Processes, Technology Press of M.I.T., Cambridge.

Howard, R. A. (1964), "Research in Semi-Markovian Decision Structure," J. Oper. Res. Soc. Japan 6, No. 4.

Iglehart, D. L. (1963), "Optimality of (s,S) Policies in the Infinite Horizon Dynamic Inventory Problem," Mgt. Sci. 9, 259-267.

Jewell, W. S. (1963), "Markov Renewal Programming, I and II," Oper. Res. 11, 938-971.

Lippman, S. (1975a), "On Dynamic Programming with Unbounded Rewards," Mgt. Sci. 27, 1225-1233.

Lippman, S. A. (1976a), "Applying a New Device in the Optimization of Exponential Queueing Systems, Oper. Res. 23, 687-710.

Magazine, M. (1971), "Optimal Control of Multi-Channel Service Systems," Nav. Res. Log. Quart. 18, 177-183.

Manne, A. (1960), "Linear Programming and Sequential Decisions," Mgt. Sci. 6, No. 3, 259-267.

McGill, J. T. (1969), "Optimal Control of Queueing Systems with Variable Number of Exponential Servers," Tech. Rep. No. 123, Department of Operations Research, Stanford University.

Orkenyi, P. (1976), "Optimal Control of the M/G/1 Queueing System with Removable Server - Linear and Non-Linear Holding Cost Function," Tech. Rep. No. 65 , Department of Operations Research, Stanford University. Office of Naval Research Contract N00014-76-C-0418.

Porteus, E. L. (1975a), "An Informal Look at Principle of Optimality," Mgt. Sci. 21, 1346-1348.

Porteus, E. L. (1974b), "On the Optimality of Structured Policies in Countable Stage Decision Processes," Mgt. Sci. 22, 148-158.

Prabhu, N. U. and Stidham, Jr. S. (1973), "Optimal Control of Queueing Systems," in Mathematical Methods in Queueing Theory, Conference at Western Michigan University, May 10-12.

Pyke, R., "Markov Renewal Processes with Finitely Many States," Ann. Math. Stat. 32, 1243-1259.

Reed, C. (1973), "Denumerable State Decision Processes with Unbounded Costs," Tech. Rep. No. 22, Department of Operations Research, Stanford University.

Reed, C. (1974a), "Difference Equations and the Optimal Control of Single Server Queueing Systems," Tech. Rep. No. 23, Department of Operations Research, Stanford University.

Reed, F. C. (1974b), "The Effect of Stochastic Time Delays on Optimal Operating Policies for M/G/1 Queueing Systems with Intermittent Service," Tech. Rep. No. 45, Department of Operations Research, Stanford University.

Ross, S. (1968), "Arbitrary State Markovian Decision Processes," Ann. Math. Stat. 39, 2118-2122.

Ross, S. (1970), Applied Probability Models with Optimization Applications, Holden-Day.

Smith, W. L. (1955), "Regenerative Stochastic Processes," Proceedings Royal Society, Series A, 232, 6-31.

Strauch, R. E. (1966), "Negative Dynamic Programming," Ann. Math. Stat. 37, 871-890.

Veinott, A. F. Jr. (1966), "On Finding Optimal Policies in Discrete Dynamic Programming with No Discounting," Ann. Math. Stat. 37, 1284-1294.

Veinott, A. F. Jr. (1969), "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," Ann. Math. Stat. 40, 1635-1660.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

(14) 1. REPORT NUMBER: TR-464

2. GOVT ACCESSION NO.

(9) 3. RECIPIENT'S CATALOG NUMBER

(6) 4. TITLE (and Subtitle): A THEORY FOR SEMI-MARKOV DECISION PROCESSES WITH UNBOUNDED COSTS AND ITS APPLICATION TO THE OPTIMAL CONTROL OF QUEUEING SYSTEMS

5. TYPE OF REPORT & PERIOD COVERED: (2) Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): (10) Peter Orkenyi

8. CONTRACT OR GRANT NUMBER(s): (15) N00014-76-C-0418, NSF-Eng-75-14847

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Dept. of Operations Research, Stanford University, Stanford, California (12) 82p.

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (16) (NR-047-061)

11. CONTROLLING OFFICE NAME AND ADDRESS: Logistics & Mathematical Statistics Branch, Code 434, Office of Naval Research - Arlington, Va. 22217

12. REPORT DATE: (11) August 1976

13. NUMBER OF PAGES: 78

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, Distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This research was supported in part by National Science Foundation Grant ENG 75-14847 and The Norwegian Research Council for Science and the Humanities.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

DYNAMIC PROGRAMMING, SEMI-MARKOV DECISION PROCESSES, QUEUEING THEORY, QUEUEING SYSTEMS, UNIMPROVABLE POLICIES, OPTIMALITY CONDITIONS, POLICY IMPROVEMENT, STATIONARY OPTIMAL POLICIES

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(see reverse side)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

402 766

A THEORY FOR SEMI-MARKOV DECISION PROCESSES WITH UNBOUNDED COSTS AND ITS APPLICATION TO THE OPTIMAL CONTROL OF QUEUEING SYSTEMS

by

Peter Orkenyi

Abstract:

Semi-Markov decision processes with countable state and action spaces are investigated. The optimality criteria considered are the average cost criterion, the undiscounted cost criterion, and the discounted cost criterion. The common assumption of bounded costs has been replaced by some considerably weaker conditions. In particular, our assumptions are weaker than those made by Harrison, Hordijk, Lippman and Reed when they considered the same problem.

The existence of optimal, stationary optimal and stationary ε-optimal policies is investigated. Policy improvement is considered. Necessary and sufficient conditions for the optimality of a policy are given.

Then the optimal control of queueing systems is considered by formulating this general problem as a semi-Markov decision process. Finally, four different ways of proving the optimality of an unimprovable policy are developed in the context of queueing systems.