1970

Markovian decision processes with uncertain rewards

Franklin Kreamer Wolf
Iowa State University

MARKOVIAN DECISION PROCESSES WITH UNCERTAIN REWARDS

by

Franklin Kreamer Wolf

A Dissertation Submitted to the

Graduate Faculty in Partial Fulfillment of

The Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subjects: Engineering Valuation
Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

Heads of Major Departments

Signature was redacted for privacy.

Dean of Graduate College

Iowa State University
Of Science and Technology
Ames, Iowa

1970

TABLE OF CONTENTS

I. INTRODUCTION AND REVIEW OF LITERATURE

Consider a system which may be defined as a stochastic process whose realization is a series of transitions between a finite number of states. If the probability of transition to another state depends only on the current state of the system and not on the series of transitions leading to the current state, the process may be called a Markov chain. Suppose that a reward is generated immediately after each transition and that the value of the reward depends on the state of the system prior to and immediately after the transition. For a given number of transitions, the expected value of the sum of future rewards will be called the value of the system. Also suppose that one or more alternatives are associated with each state. Prior to the next transition one of the alternatives must be selected; the alternative selected will determine the probability of transition to other states and the value of the reward received due to the transition. The duty of a decision maker is to choose alternatives in a manner which will maximize the value of the system. Bellman (2) has called this model a Markovian decision process.

A Markovian decision process can be classified into discounted and non-discounted models. Call β, $0 < \beta < 1$, the discount factor and discount the reward received due to the $h + 1^{th}$ transition by β^h. The value of the discounted model is the expected value of the sum of discounted rewards and, under weak restrictions, will be bounded in the infinite transition horizon situation. Two arguments for concentrating on the discounted model follow. First is the psychological consideration that a reward received immediately has a greater intrinsic value than the

same reward received at a future date, and that the same reward received in the infinite future would have no value. Another view might be taken by the engineering economist who would equate the discount factor β^h to the present worth factor $(1 + i)^{-h}$, where i is the effective rate of interest for the period of time between transitions. In this context the value of the discounted model could be labled the present worth of future rewards.

The stochastic process describing the state transitions could be other than a Markov process. However, the Markov process is mathematically tractable and is often a satisfactory assumption when modeling a physical system. Examples are found in many areas including inventory control (8), production planning (14), equipment replacement (13) and Marketing (10). Numerous examples are also found in the natural and physical sciences.

Howard (12) draws on the extensive accumulated knowledge of the properties of a Markov process and on work by Bellman (2) to define, with admirable simplicity, the Markovian decision process as a dynamic programming problem. A major contribution by Howard is the development of a procedure to determine the maximum value of a system when the state transition horizon is infinite. Manne (15), Wagner (20) and Derman (7) formulated Howard's model as a linear programming problem, thus establishing an interesting link between dynamic and linear programming.

Howard assumed the transition probabilities and the rewards to be constants. An extension of the Markovian decision process is obtained by presuming that the decision maker is uncertain of either the transition probabilities or the reward structure. The decision maker faces the dual problem of choosing alternatives to maximize the value of the system and

using the information gained from observation of past transitions or rewards to improve the quality of future decisions, thus suggesting the addition of a Bayesian component to the Markovian decision process. Robbins (18), in a paper written before Howard's work, raised a question related to this problem. Given two statistical distributions and knowing only the class of these distributions, what sequential sampling strategy will maximize $E(S_n)$, $S_n = x_1 + x_2 + \ldots + x_n$, when x_i may be drawn from either distribution? Modification of this problem by the assignment of some prior knowledge of the two distributions leads to the "two-armed bandit" problem discussed in papers by Bradt, Johnson and Karlin (6) and Feldman (9), and a variation of the problem by Box and Hill (5).

Martin (16) first considered the Markovian decision process described by Howard, and then assumed the transition probabilities to be random variables with uncertain parameters. The parameters are described by prior distributions. A strategy will depend on the past history of transitions, and the expected value of the sum of future rewards is conditioned on the history of transitions. To obtain a computable model Martin required the prior distributions to be natural conjugates of the densities of the transition probabilities. Raiffa and Schlaifer (17) analyse this topic in some detail. Silver (19) considered convenient prior distributions to use with reward distributions. Others, including Billingsley (4) and Anderson and Goodman (1), have considered statistics associated with Markov chains.

This thesis is primarily concerned with a Markovian decision process with transitions occurring at fixed intervals of time, state stationary probabilities and discounted rewards. It differs from previous models by assuming uncertain rewards. The rewards are considered to be random

variables from a known class of distributions. The parameters of these distributions are described by a set of prior distributions. Chapter II considers a general decision model with uncertain rewards which places few restrictions on the stochastic process involved. A less general model is developed by restricting the manner in which transitions and rewards are generated. A recursive equation of the value of the system is developed. Chapter III applies the results of the preceding chapter to a Markovian decision process with uncertain rewards. This allows the modification of previous notation to a more economical form. Chapter IV examines the Markovian decision process when the rewards are generated by a Bernoulli process with a beta prior density function. A strategy is defined and several theorems by Martin (16) concerning the existence and uniqueness of the value of the system are given. Chapter V is concerned with a method of calculating the value of a system when the state transition horizon is infinite. Upper and lower bounds for the value are developed as well as a method of selecting the alternative which should be chosen to govern the next transition.

II. A DECISION PROCESS WITH UNCERTAIN REWARDS

Before proceeding to a discussion of a Markovian decision process with uncertain rewards, a more general decision process with uncertain rewards will be examined. The reader may find the general model to be an interesting topic in its own right, and the results of this chapter are of direct use in Chapter III.

A. A General Decision Model

Consider a system which must be in one of a finite number of states. At discrete intervals of time the system undergoes transitions which allow it to change state. Immediately after each transition, the decision maker receives a reward; the value of the reward received due to the h^{th} transition is discounted by β^{h-1}, $0 \leq \beta < 1$. The value of the system is defined to be the expected value of the sum of the discounted rewards received over a specified number of transitions. The transition horizon is the number of transitions remaining before termination and may be infinite. Prior to each transition the decision maker selects a single course of action from among the alternatives available; the set of available alternatives is a function of the current state of the system. It is assumed that the decision maker has available to him the record of transitions, rewards received and alternatives used. The alternative chosen will govern both the transition and the reward received. Future transitions and rewards may be dependent on the previous transitions and rewards. Uncertainty concerning the reward enters the model by assuming the reward to be a sample from a distribution with an unknown parameter, and a prior distribution of the parameter is specified.

It is now necessary to briefly describe a strategy for a decision process with uncertain rewards. A more thorough description for a Markovian decision process with Bernoulli rewards is found in section C of Chapter IV. As mentioned, the decision maker is assumed to have perfect knowledge of past transitions and rewards. When the system is in an initial state i_0 and n transitions remain before termination, the decision maker can specify the alternative to be chosen to govern the first transition; call this specification $k^1(i_0,n)$. Denote the states and rewards by

$$\begin{aligned} i_h &= \text{state of the system after the } h^{th} \text{ transition,} \\ R_h &= \text{reward received due to the } h^{th} \text{ transition.} \end{aligned} \tag{2.1}$$

A bar superscript signifies a $1 \times h$ vector, e.g. $\bar{i}_h = (i_1, i_2, \ldots i_h)$. The alternative chosen to govern the second transition will depend on the result of the first transition, i_1 and R_1, which was governed by the specification $k^1(i_0,n)$. Denote the specifications used for the second transition by $k^2_{i_1,R_1}(i_0,n)$. The alternative chosen to govern the h^{th} $(h \leq n)$ transition will be a function of all previous transitions and rewards. Denote the specifications for the h^{th} transition by

$$k^h_{\bar{i}_{h-1},\bar{R}_{h-1}}(i_0,n). \tag{2.2}$$

The number of specifications with the superscript h is equal to the number of possible histories of transitions and rewards leading to the h^{th} transition. For a particular history leading through the first

h-1 transitions, specification (2.2) dictates the alternative to be chosen to govern the h^{th} transition. The collection of specifications $k^h_{\bar{i}_{h-1},\bar{R}_{h-1}}(i_0,n)$, h = 1, 2, ... n, is a strategy $D(i_0,n)$. This strategy specifies the alternative to be chosen prior to any transition in the horizon for all possible histories leading to that transition. Note that the strategy $D(i_0,n)$ can be partitioned into those specifications pertaining to the first h transitions and those specifications pertaining to the last n-h transitions; in addition, the specifications which pertain only to the h^{th} transition may be considered. This allows the following definitions.

D^h = those specifications of $D(i_0,n)$ pertaining to the first h transitions.

$D^{n-h}_{\bar{i}_h,\bar{R}_h}$ = those specifications of $D(i_0,n)$ pertaining to the final n-h transitions given the history $\bar{i}_h$, $\bar{R}_h$.

$k^h_{\bar{i}_{h-1},\bar{R}_{h-1}}$ = those specifications of $D(i_0,n)$ pertaining to the h^{th} transition. (2.3)

The most generalized decision model considered in this thesis is developed using the following additional symbols.

β = a discount factor, $0 \leq \beta < 1$

m_h = a random variable representing the parameter of the distribution from which R_h is sampled. (2.4)

Let the following stochastic triple represent the realization of the h^{th} transition.

$$\{i_h, m_h, R_h\} = a_h \tag{2.5}$$

Since only i_h and R_h are observable, it is convenient to define

$$\{i_h, R_h\} = b_h \ . \tag{2.6}$$

When the system is initially in state i_0 and strategy $D^n = D(i_0,n)$ is used, denote the joint likelihood of the sequence $\{i_1, m_1, R_1\}$, $\{i_2, m_2, R_2\}$, ... $\{i_n, m_n, R_n\}$ by

$$\pi_{a_1,a_2,\ldots a_n}(\ldots; i_0, D^n) \ . \tag{2.7}$$

Since the sequence $a_1, a_2, \ldots a_h$ cannot depend on any member of D^n other than the members of D^h, the marginal likelihood of $a_1, a_2, \ldots a_h$ may be written

$$\int_{a_{h+1}} \int_{a_{h+2}} \cdots \int_{a_n} \pi_{a_1,a_2,\ldots a_n}(\ldots; i_0, D^h)\, da_n \ldots da_{h+2}\, da_{h+1}$$

$$= \pi^*_{a_1,a_2,\ldots a_h}(\ldots; i_0, D^h) \ . \tag{2.8}$$

The conditional likelihood of a_h given $a_1, a_2, \ldots a_{h-1}$ depends only on those members of D^n which pertain to the choice of alternatives to

govern the h^{th} transition, and is written

$$\pi_{a_h}(./\bar{a}_{h-1}; i_0, k^h_{\underline{b}_{h-1}}) = \frac{\pi^*_{\underline{a}_h}(\ldots; i_0, D^h)}{\pi^*_{\underline{a}_{h-1}}(\ldots; i_0, D^{h-1})} \quad . \tag{2.9}$$

Analogously, the conditional likelihood of the sequence $a_h, a_{h+1}, \ldots a_n$ given $a_1, a_2, \ldots a_{h-1}$ depends on the members of D^n pertaining to the final $n-h$ transitions and is written

$$\pi_{a_h, a_{h+1}, \ldots a_n}(\ldots/a_1, a_2, \ldots a_{h-1}; i_0, D^{n-h}_{\underline{b}_{h-1}})$$

$$= \frac{\pi_{a_1, a_2, \ldots a_n}(\ldots; i_0, D^n)}{\pi^*_{a_1, a_2, \ldots a_{h-1}}(\ldots; i_0, D^{h-1})} \quad . \tag{2.10}$$

When a system which started in state i_0 has $n-h$ transitions remaining until termination, the sequence $a_1, a_2, \ldots a_h$ has occurred and strategy $D(i_0, n) = D^n$ is being used, denote the expected value of the sum of the remaining $n-h$ discounted rewards by

$$w(a_1, a_2, \ldots a_h; i_0, D^{n-h}_{\underline{b}_h}) \quad . \tag{2.11}$$

Use the dummy variable a_0 to denote the lack of history when writing the value of the system before the first transition.

The expected value of the sum of the discounted rewards when the

system is in state i_0 and strategy $D(i_0,n)$ is used is

$$w(a_0; i_0, D^n) = \int_{a_1}\int_{a_2}\cdots\int_{a_n} \pi_{a_1,a_2,\ldots a_n}(a_1, a_2, \ldots a_n; i_0, D^n)$$

$$\times (R_1 + \beta R_2 + \ldots + \beta^{n-1}R_n)da_n, \ldots da_2, da_1. \quad (2.12)$$

In the following chapters, the desirability of writing equation (2.12) in a recursive form will become evident. As the first step in this direction, write equation (2.12) in the following manner:

$$w(a_0; i_0, D^n) = \int_{a_1} R_1 \cdot \pi^*_{a_1}(a_1; i_0, D^1)da_1$$

$$+ \beta \int_{a_1}\int_{a_2}\cdots\int_{a_n} \pi_{a_1,a_2,\ldots a_n}(a_1, a_2, \ldots a_n; i_0, D^n)$$

$$\times (R_2 + \beta R_3 + \ldots + \beta^{n-2}R_n)da_n \ldots da_2, da_1$$

$$= \int_{a_1} R_1 \cdot \pi^*_{a_1}(a_1; i_0, D^1)da_1$$

$$+ \beta \int_{a_1} \pi^*_{a_1}(a_1; i_0, D^1)$$

$$\times \left[\int_{a_2}\int_{a_3}\cdots\int_{a_n} \frac{\pi_{a_1,a_2,\ldots a_n}(a_1,a_2,\ldots a_n;i_0,D^n)}{\pi^*_{a_1}(a_1;i_0,D^1)}\right.$$

$$\left.\times (R_2 + \beta R_3 + \ldots + \beta^{n-2}R_n)da_n \ldots da_3, da_2\right] da_1$$

$$= \int_{a_1} R_1 \cdot \pi^*_{a_1}(a_1;\, i_0,\, D^1)\,da_1$$

$$+ \beta \int_{a_1} R_1 \cdot \pi^*_{a_1}(a_1;\, i_0,\, D^1)$$

$$\times \left[\int_{a_2} \int_{a_3} \cdots \int_{a_n} \pi^*_{a_2,a_3,\ldots a_n}(a_2,a_3,\ldots a_n/a_1;i_0,\underline{D}^{n-1}_{b_1}) \right.$$

$$\left. \times\, (R_2 + \beta R_3 + \ldots \beta^{n-2} R_n)\,da_n \ldots da_3 da_2 \right] da_1 \quad .$$

(2.13)

The first addend of equation (2.13) is the expected value of the immediate reward (the reward received due to the next transition), and the second the expected value of the sum of the remaining discounted rewards. The value of the last n-1 transitions given a_i is

$$w(a_1;\, i_0,\, \underline{D}^{n-1}_{b_1}) = \int_{a_2} \int_{a_3} \cdots \int_{a_n} \pi_{a_2,a_3,\ldots a_n}(a_2,a_3,\ldots a_n/a_1;\, i_0,\, \underline{D}^{n-1}_{b_1})$$

$$\times\, (R_2 + \beta R_3 + \ldots + \beta^{n-2} R_n)\,da_n \ldots da_3 da_2 \quad .$$

(2.14)

Equations (2.12) can now be written in the following recursive form.

$$w(a_0;\, i_0,\, D^n) = \int_{a_1} R_1 \quad \pi^*_{a_1}(a_1;\, i_0,\, D^1)\,da_1$$

$$+ \beta \int_{a_1} \pi^*_{a_1}(a_1;\, i_0,\, D^1)\, w(a_1;\, i_0,\, \underset{\bar{b}_1}{D^{n-1}})\,da_1 \qquad (2.15)$$

The maximum value of the system and a strategy which will achieve that value are of major interest to the decision maker. Now introduced are some problems that will be approached in greater detail in the following chapters, particularly Chapter IV.

If $\Delta(i_0,\, n)$ is the set of all strategies $D(i_0,\, n)$ then define

$$v(a_0;\, i_0,\, n) = \sup_{D(i_0,n)\varepsilon\Delta(i_0,n)} \{w(a_0;\, i_0,\, D(i_0,\, n))\} \qquad (2.16)$$

In the same sense, $v(\bar{a}_h;\, i_0,\, n\text{-}h)$ will denote the supremum of the expected value of the sum of the discounted rewards due to the remaining $n\text{-}h$ transitions, given that the sequence $a_1, a_2, \ldots a_h$ has occurred. Equation (2.15) suggests a dynamic programming problem and application of Bellman's "Principle of Optimality" (3). Represent the alternative chosen by the decision maker to govern the h^{th} transition by k_h, ($k_h = 1, 2, \ldots K_i$), where K_i is the number of alternatives available when the system is in state i. In equation (2.15), replace D^1, the alternative chosen to govern the first transition under strategy $D(i_0,\, n)$, by k_1 and write

$$v(a_0;\, i_0,\, n) = \max_{1 \le k_1 \le K_{i_0}} \Big\{ \int_{a_1} R_1 \pi^*_{a_1}(a_1;\, i_0,\, k_1)\, da_1 + \beta \int_{a_1} \pi^*_{a_1}(a_1;\, i_0,\, k_1)\, v(a_1;\, i_0,\, n-1)\, da_1 \Big\} \quad . \tag{2.17}$$

When an infinite transition horizon is considered, the parameter n will be dropped from the symbols denoting a strategy and the value of the system. In this case equation (2.16) will be written

$$v(a_0;\, i_0) = \sup_{D(i_0) \in \Delta(i_0)} \{\, w(a_0;\, i_0,\, D(i_0))\} \quad . \tag{2.18}$$

Of interest are the conditions under which

$$\lim_{n \to \infty} v(a_0;\, i_0,\, n) = v(a_0;\, i_0); \tag{2.19}$$

this will be discussed in Chapter IV.

B. A Less General Decision Process

Some restrictions will now be placed on the model of section A of this chapter. A particular method of generating the conditional likelihood $\pi^*_{a_h}(a_h / \bar{a}_{h-1};\, i_0,\, \underline{k}^h_{\underline{b}_{h-1}})$ under strategy $D(i_0, n)$ will be described and denoted. From this conditional likelihood, recursive equations similar to (2.15) and (2.17) but for the less general decision process with

uncertain rewards will be obtained.

Consider a process which generates the stochastic triple (i_h, m_h, R_h) in the following manner. The probability of transition from state i_{h-1} to state i_h depends on the past history of states, $i_1, i_2, \ldots i_{h-1}$, and the alternative chosen to govern the h^{th} transition. Denote this by

$$P_{i_h}(\cdot / i_1, i_2, \ldots i_{h-1}; i_0, k_h) \quad . \tag{2.20}$$

There is a reward distribution associated with each transition from state i_{h-1} to state i_h and each alternative available when in state i_{h-1}. If N denotes the number of states in the system, there are $L = \sum_{i=1}^{N} NK_i = N \sum_{i=1}^{N} K_i$ reward distributions from which R_h can be sampled. The particular distribution sampled is indexed by i_{h-1}, i_h and k_h. Let m_h represent the parameter of the sampled distribution and denote the likelihood of R_h by

$$\ell_{R_h}(\cdot / i_{h-1}, i_h; m_h, k_h) \quad . \tag{2.21}$$

The decision maker is uncertain of the value of the parameter m_h and views it as a random variable. Because there are L distributions from which R_h can be sampled, there are also L distributions from which m_h can be sampled. These distributions are also indexed by i_{h-1}, i_h and k_h, so that with each reward distribution there is associated a distribution of m_h with identical indices. As part of the initial conditions the decision maker must specify L independent prior

distributions. The distribution from which m_h is sampled is one of the set of L prior distributions, the class of which is denoted by

$$\Phi_m(\cdot/\psi) = \{\phi_\mu(\cdot/\psi_\mu); \mu = 1, 2, \ldots L\} \quad . \tag{2.22}$$

The specifications of the strategy discussed are dependent on the transitions and rewards observed, and the expected value of the system will be taken with respect to those observations. It will be necessary to compute the conditional distribution of m_h given the observations $\bar{i}_{h-1}$ and $\bar{R}_{h-1}$; in Bayesian terminology this is the posterior distribution of m_h and is denoted by

$$\phi_{m_h}(\cdot/R_1, R_2, \ldots R_{h-1}, i_1, i_2, \ldots i_h; i_0, k_h, \psi) \quad . \tag{2.23}$$

Although it is necessary to state the alternatives chosen to govern the first h-1 transitions to completely index the reward distributions sampled, these alternatives are known when the strategy is specified and the history of transitions and rewards is given. The posterior distribution (2.23) is obtained by application of Bayes theorem.

$$\phi_{m_h}(\cdot/ \bar{R}_{h-1}, \bar{i}_h; i_0, k_h, \psi) = \frac{\phi_{m_h}(\cdot/\psi_{m_h}) \ell_{R_1}(R_1/i_0, i_1; m_1, k_1)}{\int_{m_h} \phi_{m_h}(m_h/\psi_{m_h}) \ell_{R_1}(R_1/i_0, i_1; m_1, k_1)}$$

$$\times \frac{\ell_{R_2}(R_2/i_1, i_2; m_2, k_2) \ldots \ell_{R_{h-1}}(R_{h-1}/i_{h-2}, i_{h-1}; m_{h-1}, k_{h-1})}{\ell_{R_2}(R_2/i_1, i_2; m_2, k_2) \ldots \ell_{R_{h-1}}(R_{h-1}/i_{h-2}, i_{h-1}; m_{h-1}, k_{h-1}) dm_h} \quad . \tag{2.24}$$

Because it is difficult to devise an economical notation which records both the decision state and the distribution sampled, equation (2.24) is somewhat cumbersome. When the indices of the reward distribution sampled are not equal to i_{h-1}, i_h and k_h, the indices determining m_h, the likelihood of that reward is functionally independent of m_h. Those likelihoods in the denominator which are functionally independent of m_h can be placed outside the integral and will cancel with the corresponding likelihoods in the numerator. The fact that the posterior distribution of m_h is altered only by rewards sampled from the reward distribution associated with m_h is clear but somewhat obscured by the notation.

One interpretation of the reward structure just described is that the decision maker knows the family of reward distributions but is uncertain of at least one of the distribution parameters. Specification of the unknown parameter as a random variable reflects the decision makers uncertainty. In spite of this uncertainty the decision maker must initially estimate ψ, a term containing the parameters of the prior distributions and indirectly representing a partial knowledge of the parameters of the reward distributions. The decision maker systematically updates his initial estimate of ψ by conditioning the distribution of m_h on past observations. In this manner the decision maker bases future decisions on both ψ and the observed rewards. But he must not lose sight of the fact that the expected value of the system is functionally dependent on the initial estimate of ψ.

The joint likelihood $\pi_{\underset{a_n}{}}(\cdot; i_0, D^n)$ of (2.7) can be written to include the additional parameter ψ.

$$\pi_{\underline{a}_n}(.;\, i_0,\, D^n,\, \psi) \;=\; P_{i_1}(.;\, i_0,\, k^1_{\underline{b}_0},\, \psi)\phi_{m_1}(./\,\bar{i}_1;\, i_0,\, k^1_{\underline{b}_0},\, \psi)$$

$$\times\, \ell_{R_1}(./\,\bar{i}_1,\, \bar{m}_1;\, i_0,\, k^1_{\underline{b}_0},\, \psi)P_{i_2}(./\,\bar{i}_1,\, \bar{m}_1,\, \bar{R}_1;\, i_0,\, k^2_{\underline{b}_1},\, \psi)$$

$$\times\, \phi_{m_2}(./\,\bar{i}_2,\, \bar{m}_1,\, \bar{R}_1;\, i_0,\, k^2_{\underline{b}_1},\, \psi)\ell_{R_2}(./\,\bar{i}_2,\, \bar{m}_2,\, \bar{R}_1;\, i_0,\, k^2_{\underline{b}_1},\, \psi)\;\ldots$$

$$P_{i_n}(./\,\bar{i}_{n-1},\, \bar{m}_{n-1},\, \bar{R}_{n-1};\, i_0,\, k^n_{\underline{b}_{n-1}},\, \psi)$$

$$\times\, \phi_{m_n}(./\,\bar{i}_n,\, \bar{m}_{n-1},\, \bar{R}_{n-1};\, i_0,\, k^n_{\underline{b}_{n-1}},\, \psi)$$

$$\times\, \ell_{R_n}(./\,\bar{i}_n,\, \bar{m}_n,\, \bar{R}_{n-1};\, i_0,\, k^n_{\underline{b}_{n-1}},\, \psi)\;. \qquad (2.25)$$

From equation (2.9), the conditional distribution of a_h given $\bar{a}_{h-1}$, is

$$\pi_{a_h}(./\,\bar{a}_{h-1};\, i_0,\, k^h_{\underline{b}_{h-1}},\, \psi) \;=\; P_{i_h}(./\,\bar{i}_{h-1},\, \bar{m}_{h-1},\, \bar{R}_{h-1};\, i_0,\, k^h_{\underline{b}_{h-1}},\, \psi)$$

$$\times\, \phi_{m_h}(./\,\bar{i}_h,\, \bar{m}_{h-1},\, \bar{R}_{h-1};\, i_0,\, k^h_{\underline{b}_{h-1}},\, \psi)$$

$$\times \ell_{R_h}(./\bar{i}_h, \bar{m}_h, \bar{R}_{h-1}; i_0, k^h_{\bar{b}_{h-1}}, \psi) \quad . \tag{2.26}$$

The assumptions made when describing the less general decision process allow equation (2.26) to be written

$$\pi_{a_h}(./\bar{a}_{h-1}; i_0, k^h_{\bar{b}_{h-1}}, \psi) = P_{i_h}(./\bar{i}_{h-1}; i_0, k^h_{\bar{b}_{h-1}})$$

$$\times \phi_{m_h}(./\bar{R}_{h-1}, \bar{i}_h; i_0, k^h_{\bar{b}_{h-1}}, \psi)\ell_{R_h}(./i_{h-1}, i_h, m_h; k^h_{\bar{b}_{h-1}}) \quad . \tag{2.27}$$

An equation analogous to equation (2.15) but expressing the expected value of the sum of discounted rewards for the less general decision process when strategy $D(i_0,n)$ is used can now be written

$$w(a_0; i_0, D^n, \psi) = \int_{i_1}\int_{m_1}\int_{R_1} R_1 \cdot P_{i_1}(i_1/\bar{i}_0; i_0, k^1_{\bar{b}_0})$$

$$\times \phi_{m_1}(m_1/\bar{R}_0, \bar{i}_0; i_0, k^1_{\bar{b}_0}, \psi)\ell_{R_1}(R_1/i_0, i_1, m_1; k^1_{\bar{b}_0})dR_1dm_1di_1$$

$$+ \beta\int_{i_1}\int_{m_1}\int_{R_1} P_{i_1}(i_1/\bar{i}_0; i_0; k^1_{\bar{b}_0})\phi_{m_1}(m_1/\bar{R}_0, \bar{i}_0; i_0, k^1_{\bar{b}_0}, \psi)$$

$$\times \ell_{R_1}(R_1/i_0, i_1, m; k^1_{\overline{b}_0}) w(a_1; i_0, D^{n-1}_{\overline{b}_1}, \psi)\, dR_1\, dm_1\, di_1, \qquad (2.28)$$

where $\overline{i}_0$, $\overline{R}_0$ and $\overline{b}_0$ are dummy variables representing the lack of history on the first transition. The equation similar to equation (2.17) but for the less general decision process is

$$v(a_0; i_0, n, \psi) = \max_{1 \le k \le K_{i_0}} \Big\{ \int_{i_1} \int_{m_1} \int_{R_1} R_1 \cdot P_{i_1}(i_1/\overline{i}_0; i_0, k)$$

$$\times \phi_{m_1}(m_1/\overline{R}_0, \overline{i}_0; i_0, k, \psi) \ell_{R_1}(R_1/i_0, i_1, m_1; k)\, dR_1\, dm_1\, di_1$$

$$+\beta \int_{i_1} \int_{m_1} \int_{R_1} P_{i_1}(i_1/\overline{i}_0; i_0, k) \phi_{m_1}(m_1/\overline{R}_0, \overline{i}_0; i_0, k, \psi)$$

$$\times \ell_{R_1}(R_1/i_0, i_1, m_1; k) v(a_1; i_0, n-1, \psi)\, dR_1\, dm_1\, di_1 \Big\} . \qquad (2.29)$$

III. A MARKOVIAN DECISION PROCESS WITH UNCERTAIN REWARDS

The models of the preceding chapter did not restrict the probability structure underlying the state transitions. By assuming that the state transitions can be described by a stationary Markov chain, the results of Chapter II can be modified to specify the characteristics of a Markovian decision process with uncertain rewards. The singular properties of the stationary Markov chain, as applied to the decision process being considered, are that the conditional probability of transition to state i_h given the transition history $i_0, i_1, \ldots i_{h-1}$, is dependent only on state i_{h-1} and the alternative chosen to govern the h^{th} decision; the probability is functionally independent of the state history leading to i_{h-1}. Stationarity refers to the functional independence of the transition probability and the number of previous transitions. Assuming that the state transition probabilities are represented by a stationary Markov chain, the probability (2.20) may be written

$$P_{i_h}(./\bar{i}_{h-1};\, i_0,\, k_h) = P_{i_h}(./i_{h-1};\, k_h) \quad . \tag{3.1}$$

Equation (2.28) stated the expected value of the sum of future discounted rewards under strategy $D(i_0;n)$ for the less general model. For a Markovian decision process with uncertain rewards, the value of the system is

$$w(a_0;\, i_0,\, D^n,\, \psi) = \sum_{i_1=1}^{N} \int_{m_1} \int_{R_1} R_1 \cdot P_{i_1}(i_1/i_0,\, k_{b_0}^{1})$$

$$\times \phi_{m_1}(m_1/\bar{R}_0, \bar{i}_0; i_0, k^1_{\bar{b}_0}, \psi)\ell_{R_1}(R_1/i_0, i_1, m_1; k^1_{\bar{b}_0})dR_1 dm_1$$

$$+ \beta \sum_{i_1=1}^{N} \int_{m_1}\int_{R_1} P_{i_1}(i_1/i_0; k^1_{\bar{b}_0})\phi_{m_1}(m_1/\bar{R}_0, \bar{i}_0; i_0, k^1_{\bar{b}_0}, \psi)$$

$$\times \ell_{R_1}(R_1/i_0, i_1, m_1; k^1_{\bar{b}_0})w(a_1; i_0, D^{n-1}_{\bar{b}_1}, \psi)dR_1 dm_1 \qquad (3.2)$$

It is possible to write equation (3.2) in a more compact form by modifying some of the notation of Chapter II. Rather than specifying the state of the system before and after the h^{th} transition by i_{h-1} and i_h, use, when possible, the indices i and j to refer directly to the state of the system. Use the following notation for a Markovian decision process.

$p^k_{i,j}$ = the probability of transition from state i to state j when alternative k governs the transition.

$\ell^k_{i,j}(./m)$ = the likelihood function of the reward received due to transition from state i to state j when alternative k governs the transition; m is a random variable representing the parameter of the function and the density of m is indexed by i, j and k.

$\phi^k_{i,j}(.;\psi)$ = the prior density function of m indexed by i, j and k.

$$\phi^{k}_{i,j}(./R_1, R_2, \ldots R_{h-1}, i_1, i_2 \ldots i_{h-2}; i_0, \psi) =$$

the posterior density function of m indexed by i, j and k and conditional on the observations $R_1, R_2, \ldots R_{h-1}$; let $i = i_{h-1}$ and $j = i_h$. (3.3)

Reduce the notation required when denoting the posterior distribution of m by letting

$$\phi^{k_h}_{i,j}(./ \bar{R}_{h-1}, \bar{i}_{h-2}; i_0, \psi) = \phi^{k_h}_{i,j}(.; \psi^h) \quad . \tag{3.4}$$

This notation may be taken to imply that the posterior distribution of m is of the same family as the prior distribution, with ψ^h denoting the value of the parameters of the prior distributions updated through the $(h-1)^{th}$ transition. Chapter IV considers a reward structure of this type.

The author is indebted to J. J. Martin whose book (16) suggested the following notation. Let $T^k_{i,j}(R,\psi)$ denote the parameters of the posterior distributions given one additional reward observation R sampled from the distribution indexed by i, j and k, so that

$$\phi^{k}_{i,j}(./R; \psi) = \frac{\phi^{k}_{i,j}(.;\psi)\ell^{k}_{i,j}(R/m)}{\int_m \phi^{k}_{i,j}(m;\psi)\ell^{k}_{i,j}(R/m)\,dm} = \phi^{k}_{i,j}(.; T^k_{i,j}(R,\psi)) \quad . \tag{3.5}$$

The conditional likelihood of a_h given $\bar{a}_{h-1}$ for the less general decision model was specified in equation (2.27). For the Markovian

decision process, the analogous equation is

$$\pi_{a_h}(./\bar{a}_{h-1};\, i_0,\, k^h_{\bar{b}_{h-1}},\, \psi) = p_{i,j}^{k^h_{\bar{b}_{h-1}}}\, \ell_{i,j}^{k^h_{\bar{b}_{h-1}}}(./m)\, \phi_{i,j}^{k^h_{\bar{b}_{h-1}}}(.;\, \psi^h) \quad . \tag{3.6}$$

Before rewriting equation (3.2), consider the first addend on the right hand side of that equation, which represents the expected value of the reward received due to the next transition. When the system is in state i prior to the h^{th} transition, denote the expected value of R_h by

$$q_i^{k_h}(\psi^h) = \sum_{j=1}^{N} \int_m \int_R R \cdot p_{i,j}^{k_h}\, \phi_{i,j}^{k_h}(m;\, \psi^h)\, \ell_{i,j}^{k_h}(R/m)\, dR\, dm \quad . \tag{3.7}$$

The value of the sum of the remaining $n-h$ discounted rewards is now a function of the current state of the system and the updated parameters of the prior distributions. When $i_h = i$, let

$$w(\bar{a}_h;\, i_0,\, D^{n-h}_{\bar{b}_h},\, \psi) = w_i(D^{n-h}_{\bar{b}_h},\, \psi^{h+1}) \quad . \tag{3.8}$$

Rewrite equation (3.2) using the notation developed in this chapter.

$$w_i(D^n,\, \psi) = q_i^{k^1_{\bar{b}_0}}(\psi) + \beta \sum_{j=1}^{N} \int_m \int_R p_{i,j}^{k^1_{\bar{b}_0}}\, \phi_{i,j}^{k^1_{\bar{b}_0}}(m;\, \psi)\, \ell_{i,j}^{k^1_{\bar{b}_0}}(R/m)$$

$$\times w_j(D^{n-1}, T_{i,j}^{\underline{k}^1_{b_0}}(R, \psi))dRdm \tag{3.9}$$

The dynamic programming formulation of the Markovian decision process with uncertain rewards is

$$v_i(n, \psi) = \max_{1 \le k \le K_i} \{ q_i^k(\psi) + \beta \sum_{j=1}^{N} \int_m \int_R p_{i,j}^k \phi_{i,j}^k(m; \psi)$$

$$\times \ell_{i,j}^k(R/m) v_j(n-1, T_{i,j}^k(R, \psi))dRdm \} , \tag{3.10}$$

where $v_i(n, \psi)$ is the supremum of the value of the remaining n transitions when the system is state i.

IV. A MARKOVIAN DECISION PROCESS WITH UNCERTAIN BERNOULLI REWARDS

This chapter will analyse a Markovian decision process with uncertain rewards and with a specific reward structure. First a particular reward likelihood and prior distribution will be specified, and then the results of Chapter III will be used to obtain equations describing a Markovian decision process with Bernoulli rewards. A strategy for the case when the number of possible rewards is finite will then be described in detail. Finally the existence and uniqueness of $v_i(\psi)$, an optimal strategy and bounds of the function $|v_i(n,\psi) - v_i(\psi)|$ will be considered.

A. A Discrete Reward Structure

The reward structure to be considered assumes that rewards are generated by a Bernoulli process. The parameter of this process is uncertain; this uncertainty will be described by assuming the parameter to be a random variable defined by a prior distribution. Describe the likelihood of the reward sampled from the distribution indexed by i, j and k by

$$\ell^k_{i,j}(R/m) = m^x(1-m)^{1-x},$$

$$0 \leq m \leq 1, \qquad R = R(1), R(2), \qquad x = 1 \text{ when } R = R(1)$$
$$x = 0 \text{ when } R = R(2) \ . \qquad (4.1)$$

The parameter m in (4.1) is a random variable to which a prior distribution also indexed by i, j and k must be assigned. Two criteria

should be considered when selecting a prior distribution. First, the distribution must appeal to the model builder's intuition and seem a reasonable way in which to describe m. The second criterion is more objective. For the model to be useful, the posterior distribution of m must be calculable. Raiffa and Schlaifer (17) have extensively examined the class of prior distributions which are natural conjugates of the process (i.e. reward) distribution. The natural conjugate prior density function has the characteristic that the posterior density is of the same family as the prior density, and that the parameters of the posterior density are often simple functions of the parameters of the prior density. If the model builder's intuition allows him to reduce the set of candidates for the prior distribution to the natural conjugate prior, he will achieve a large return in terms of computability.

When rewards are generated by the Bernoulli process described in (4.1), the likelihood function of the sample $R_1, R_2, \ldots R_t$ from the distribution indexed by i, j and k is

$$\prod_{i=1}^{t} m^{x_i}(1-m)^{1-x_i} = m^s(1-m)^{t-s};$$

$$s = \sum_{i=1}^{t} x_i, \qquad x_i = 1 \text{ when } R_i = R(1)$$
$$x_i = 0 \text{ when } R_i = R(2) \quad . \tag{4.2}$$

The natural conjugate prior density of the likelihood (4.1) is the beta distribution. This is verified by observing that the density function of the prior distribution varies as the likelihood function of the rewards.

$$f_{beta}(m;\ s,\ t) \propto m^{s-1}(1-m)^{t-s-1} \quad . \tag{4.3}$$

When the beta distribution is chosen to describe m,

$$\phi^k_{i,j}(m;\ \psi) = \phi^k_{i,j}(m;\ s,\ t) = \frac{1}{\beta(s,\ t-s)}\ m^{s-1}(1-m)^{t-s-1} \quad . \tag{4.4}$$

Because the natural conjugate prior distribution has been selected, the posterior distribution is also a beta distribution with parameters as shown below.

$$\phi^k_{i,j}(m/R_1,\ R_2,\ \ldots\ R_{t'};\ s,\ t) = \frac{m^{s-1}(1-m)^{t-s-1}\ \prod_{h=1}^{t'} m^{x_h}(1-m)^{1-x_h}}{\int_0^1 m^{s-1}(1-m)^{t-s-1}\ \prod_{h=1}^{t'} m^{x_h}(1-m)^{1-x_h}\ dm}$$

$$= \frac{1}{\beta(s+s',\ t+t'-(s+s'))}\ m^{s+s'-1}\ (1-m)^{t+t'-(s+s')-1}$$

$$= \phi^k_{i,j}(m;\ s+s',\ t+t') \quad ,$$

$$s' = \sum_{h=1}^{t'} x_h, \qquad x_h = 1 \text{ when } R = R(1)$$

$$x_h = 0 \text{ when } R = R(2) \quad . \tag{4.5}$$

As indicated, the parameters of the posterior are a simple function of the parameters of the prior, the total number of observations and the number of those observations equal R(1).

The marginal distribution of R is

$$\ell^k_{i,j}(R;\psi) = \ell^k_{i,j}(R;\, s,t) = \int_0^1 m^x(1-m)^{1-x}\,\frac{1}{\beta(s,t-s)}\,m^{s-1}(1-m)^{t-s-1}\,dm$$

$$= \left(\frac{s}{t}\right)^x \left(\frac{t-s}{t}\right)^{1-x} \quad ;$$

$$x = 1 \text{ when } R = R(1)$$

$$x = 0 \text{ when } R = R(2) \quad , \qquad (4.6)$$

where s and t are the parameters of the prior distribution indexed by i, j and k. The expected value of R is

$$E^k_{i,j}(R;\, s,t) = \sum_R R\,\ell^k_{i,j}(R;\, s,t) = R(1)\,\frac{s}{t} + R(2)\,\frac{t-s}{t} \quad . \qquad (4.7)$$

There are $L = N \sum_{i=1}^{N} K_i$ prior distributions, and the parameters s and t must be specified for each distribution. The symbol ψ denotes a $1 \times L$ vector containing the parameters of all prior distributions. The decision maker must estimate s and t for each of these distributions in the manner which best reflects any prior intelligence about the corresponding reward distribution. One point of view is that the ratio s/t should be selected to correspond with the decision maker's estimate of the expected value of R, and that the magnitude of t will reflect his certainty of the estimate of the expected value. A large value of t indicates a great deal of confidence in that estimate, while a small value

of t would allow a more rapid relative change in the posterior parameters and indicate less confidence in the initial estimate of the expected value of R.

B. The Expected Value of the Sum of Discounted Rewards

Equation (3.9) defined, in recursive form, the expected value of the sum of discounted rewards for the Markovian decision process with uncertain rewards. Using the reward likelihood (4.1), the prior distribution (4.4) and the notation of (4.6), the value of the Markovian decision process with uncertain Bernoulli rewards under the strategy D(i, n) can now be written

$$w_i(D^n, \psi) = q_i^{k_{\bar{b}_0}^1}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k_{\bar{b}_0}^1} \sum_{R} \ell_{i,j}^{k_{\bar{b}_0}^1}(R; \psi) w_j(D^{n-1}, T_{i,j}^{k_{\bar{b}_0}^1}(R, \psi)),$$

$$i = 1, 2, \ldots N, \qquad R = R(1), R(2) \quad . \tag{4.8}$$

Analogously, equation (3.10) is now

$$v_i(n, \psi) = \max_{1 \le k \le K_i} \{ q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_{R} \ell_{i,j}^k(R; \psi) v_j(n-1, T_{i,j}^k(R, \psi)) \} \quad ,$$

$$i = 1, 2, \ldots N, \qquad R = R(1), R(2) \quad . \tag{4.9}$$

For an infinite transition horizon, equation (4.9) will be written

$$v_i(\psi) = \max_{1 \le k \le K_i} \{q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_{R} \ell_{i,j}^k(R; \psi) v_j(T_{i,j}^k(R, \psi))\},$$

$$i = 1, 2, \ldots N, \qquad R = R(1), R(2) \quad . \tag{4.10}$$

The remainder of this thesis will focus on the properties and solution of equation (4.10).

C. A Strategy for a Markovian Decision Process with Uncertain Bernoulli Rewards

A strategy for the decision process described in the previous section will now be described. For an n transition horizon, a sampling strategy will be constructed by first specifying a strategy for a transition horizon of one, amending that to obtain a strategy for a horizon of two and proceeding sequentially to the n transition horizon strategy. Since there are a finite number of states and rewards, the number of strategies will be finite if the transition horizon is finite.

If the system is in state i and the alternative to govern the next transition has been chosen, the decision maker can select a policy vector which will specify the alternative to be chosen after the next transition, give the outcome of the next transition and the reward received due to that transition. Because there are $2N$ possible outcomes of a transition and a reward, the policy vector is a $1 \times 2N$ vector denoted by

$$\sigma = (k_{1,1}, k_{1,2}, k_{2,1}, k_{2,2}, \ldots k_{N,1}, k_{N,2}) \quad . \tag{4.11}$$

Element $k_{j,\ell} = 1, 2, 3, \ldots K_j$ denotes the alternative to be chosen if the next transition is to state j and the reward received due to the transition is $R = R(\ell)$, $\ell = 1, 2$. Let $\sum$ be the finite set of the $J = \prod_{i=1}^{N} K_i^2$ policy vectors σ. Index the policy vectors by the integers 0 through $J - 1$.

$$\sum = (\sigma_0, \sigma_1, \sigma_2, \ldots \sigma_{J-1}) \tag{4.12}$$

Assume that the system is initially in state i_0 and that alternative k, $k = (1, 2, 3, \ldots K_{i_0})$, has been chosen to govern the first transition. Before the first transition the decision maker can specify a policy vector $d(1) = \sigma_{\alpha_{1,1}}$, $\alpha_{1,1} = 0, 1, 2, \ldots J - 1$, which specifies the alternative to be chosen to govern the second transition. Let $D^2 = D(2, i_0, k, d(1))$, be called a strategy for a horizon of two transitions.

A strategy for a horizon of three transitions can be defined by stating D^2 and specifying the $2N$ policy vectors which will dictate the alternative chosen to govern the third transition. There are $2N$ possible state-reward histories leading from i_0 to the outcome of the first transition. These may be denoted by

$$\begin{aligned} &\mathbf{x}_2(i_1, \ell_1), \qquad i_1 = 1, 2, \ldots N \\ &\qquad\qquad\qquad\quad \ell_1 = 1, 2, \end{aligned} \tag{4.13}$$

where i_1 is the state of the system after the first transition and ℓ_1

completes the description of the observed reward. Because i_0 and k are known, the reward received due to the first transition is known to have been a sample from the reward distribution indexed by i_0, i_1 and k. There must be a policy vector specified for each possible state-reward history. The following function of $x_2(i_1, \ell_1)$ is a 2N-ary number which will be used to order the state-reward histories.

$$z(x_2(i_1, \ell_1)) = y(i_1, \ell_1)(2N)^{-1} ,$$

$$y(i_1, \ell_1) = 2(i_1 - 1) + \ell_1 - 1 , \qquad i = 1, 2, 3, \ldots N$$
$$\ell = 1, 2 . \qquad (4.14)$$

Associated with each state-reward history is the unique number $z(x_2(i_1,\ell_1))$. Order the state-reward histories so that $x_2(i_1', \ell_1') < x_2(i_1'', \ell_1'') < x_2(\ell_1''', \ell_1''') < \ldots$ when $z(x_2(i_1', \ell_1')) < z(x_2(i_1'', \ell_1'')) < z(x_2(i_1''', \ell_1''')) \ldots$ and index the histories with the digits 1 through 2N, assigning the history associated with the smallest value of $z(x_2(i_1, \ell_1))$ the integer 1, the next smallest the integer 2 etc. . The state-reward histories from i_0 through the first transition may now be denoted $x_{2,1}, x_{2,2}, x_{2,3}, \ldots x_{2,2N}$ where the first subscript indicates a history through the first transition and the second subscript refers to the ordering index just described. Denote the policy vector selected by the decision maker when strategy D^2 is used and state reward history $x_{2,g}$ is observed by $\gamma(2, g, D^2) = \sigma_{\alpha_{2,g}}$, $\alpha_{2,g} = 0, 1, 2, \ldots J - 1$. Define the $1 \times 2N$ vector $\gamma(2) = (\sigma_{\alpha_{2,1}}, \sigma_{\alpha_{2,2}}, \sigma_{\alpha_{2,3}}, \ldots \sigma_{\alpha_{2,2N}})$.

Lengthen this vector by attaching the element $d(1)$ to the front of $Y(2)$ and denote the result by

$$d(2) = (\sigma_{\alpha_{1,1}}, \sigma_{\alpha_{2,1}}, \sigma_{\alpha_{2,2}}, \ldots \sigma_{\alpha_{2,2N}}) \quad . \tag{4.15}$$

A strategy D^3 for a horizon of three transitions is

$$D^3 = D(3, i_0, k, d(2)) \quad . \tag{4.16}$$

The strategy D^3 explicitly states the alternative to be chosen to govern the first, second and third transitions as a function of the observed transitions and rewards.

In general, to construct a strategy D^h given D^{h-1} it is necessary to specify $(2N)^{h-2}$ additional policy vectors since that is the number of possible state-reward histories $x_{h-1}(i_1, \ell_1, i_2, \ell_2, \ldots i_{h-2}, \ell_{h-2})$ leading from i_0 through the $(h-2)^{th}$ transition. As before, the state-reward histories can be assigned a $2N$-ary number.

$$z(x_{h-1}(i_1, \ell_1, i_2, \ell_2, \ldots i_{h-2}, \ell_{h-2})) = \sum_{m=1}^{h-2} y(i_m, \ell_m)(2N)^{-m} \quad ;$$

$$y(i_m, \ell_m) = 2(i_m - 1) + \ell_m - 1, \qquad i = 1, 2, \ldots N$$
$$\ell = 1, 2 \quad . \tag{4.17}$$

Order the histories as before, so that

$$x_{h-1}(i_1', \ell_1', i_2', \ell_2', \ldots i_{h-2}', \ell_{h-2}')$$

$$< x_{h-1}(i_1'', \ell_1'', i_2'', \ell_2'', \cdots i_{h-2}'', \ell_{h-2}'')$$

when $z(x_{h-1}\ (i_1', \ell_1', i_2', \ell_2', \cdots i_{h-2}', \ell_{h-2}'))$

$$< z(x_{h-1}(i_1'', \ell_1'', i_2'', \ell_2'', \cdots i_{h-2}'', \ell_{h-2}''))\ ,$$

and index the histories with the integers 1 through $(2N)^{h-2}$. The policy vector to be selected prior to the $(h-1)^{th}$ transition, when the history through the $(h-2)^{th}$ transition identified by the ordering index g has been observed, is $\gamma(h-1, g, D^{h-1}) = \sigma_{\alpha_{h-1,g}}$. Let the $1 \times (2N)^{h-2}$ vector $\gamma(h-1) = (\sigma_{\alpha_{h-1,1}}, \sigma_{\alpha_{h-1,2}}, \cdots \sigma_{\alpha_{h-1,(2N)^{h-2}}})$ specify the policy vectors selected prior to the $(h-1)^{th}$ transition and which in turn will dictate the alternative to be chosen to govern the h^{th} transition. Combine $\gamma(h-1)$ and $d(h-2)$ to obtain $d(h-1)$.

$$d(h-1) = (\sigma_{\alpha_{1,1}}, \sigma_{\alpha_{2,1}}, \sigma_{\alpha_{2,2}}, \cdots \sigma_{\alpha_{2,2N}}, \sigma_{\alpha_{3,1}}, \cdots \sigma_{\alpha_{h-1,(2N)^{h-2}}})\ . \tag{4.18}$$

The symbol $d(h-1)$ denotes a $1 \times M(h-1)$ vector where

$$M(0) = 0$$

$$M(h) = \sum_{g=1}^{h} (2N)^{g-1};\ h = 1, 2, 3, \cdots\ . \tag{4.19}$$

The h transition horizon strategy is

$$D^h = D(h, i_0, k, d(h-1)) \quad . \tag{4.20}$$

Call $\Delta(i_0, n)$ the set of all n transition strategies when the system is initially in state i_0. When n is finite, the total number of unique strategies contained in the set $\Delta(i_0, n)$ is finite and equal $S = K_{i_0} J^{M(n)}$.

D. The Existence and Uniqueness of $v_i(\psi)$

Some results from Bayesian Decision Problems and Markov Chains by J. J. Martin (16) are very useful at this point. A model which Martin developed and the model of this chapter have certain similarities, and several of his theorems, with only slight modifications, apply to a Markovian decision process with uncertain Bernoulli rewards. In this spirit, four theorems based on Martin (16, p. 38-44) follow. The proofs given are from Martin but with the required changes.

Theorem 4.1. Let $w_i(D, \psi)$ be the expected value of the sum of discounted rewards when the system is in state i, strategy D is used and the transition horizon is infinite. Let

$$v_i(\psi) = \sup_{D \in \Delta(i)} \{w_i(D, \psi)\} \quad . \tag{4.21}$$

Then there is a strategy $D^* \in \Delta(i)$ such that

$$v_i(\psi) = w_i(D^*, \psi) \quad . \tag{4.22}$$

Proof. First it will be shown that $v_i(\psi)$ is bounded. Denote the possible rewards from the distribution indexed by i, j and k by $R^k_{i,j}(\ell)$, $\ell = 1, 2$. If $R^* = \max_{i,j,k,\ell} \{R^k_{i,j}(\ell)\}$, then the maximum value of the sum of discounted rewards which can be received is

$$\sum_{h=1}^{\infty} \beta^{h-1} R^* = \frac{R^*}{1-\beta} . \tag{4.23}$$

Letting δ denote the set of all d in the strategy $D = D(i, k, d)$, equation (4.21) can be written

$$v_i(\psi) = \max_{1 \le k \le K_i} \sup_{d \in \delta} \{w_i(k, d, \psi)\} \tag{4.24}$$

To each $d \in \delta$ let there correspond the J-ary number

$$a(d) = \sum_{h=1}^{\infty} \sum_{j=1}^{(2N)^{h-1}} \alpha_{h,j} J^{-(M(h-1)+j)} \tag{4.25}$$

where M(h) is defined in equation (4.19). For any $d \in \delta$, $0 \le a(d) \le 1$, and in addition equation (4.25) is a one-to-one mapping of the set δ onto the closed interval $[0,1]$. For fixed i and k let $g^k_i(a, \psi)$ be a function defined on $[0,1]$ by

$$g^k_i(a, \psi) = w_i(k, d, \psi) . \tag{4.26}$$

Then equation (4.24) can be written

$$v_i(\psi) = \max_{1 \le k \le K_1} \sup_{0 \le a \le 1} \{g_i^k(a, \psi)\} \tag{4.27}$$

To show that for fixed k, $g_i^k(a, \psi)$ is continuous in a let

$$R^{**} = \max_{i,j,k,\ell} \{|R_{i,j}^k(\ell)|\}, \quad r^{**} = \min_{i,j,k,\ell} \{|R_{i,j}^k(\ell)|\} . \tag{4.28}$$

For a given ξ choose a positive integer n such that

$$\beta^{n+1} \left(\frac{R^{**}-r^{**}}{1-\beta} \right) < \xi \tag{4.29}$$

For a fixed $a \in [0, 1]$ let a' be any number such that $0 \le a' \le 1$ and $|a - a'| < J^{-\nu}$. If $a = a(d)$ and $a' = a'(d')$ then

$$\sigma_{\alpha_{h,g}} = \sigma'_{\alpha_{h,g}}, \qquad \begin{matrix} h = 1, 2, \ldots (\nu - 1) \\ g = 1, 2, \ldots (2N)^{\nu-2} \end{matrix} . \tag{4.30}$$

Since both strategies are identical through the first ν transitions,

$$|g_i^k(a, \psi) - g_i^k(a', \psi)| \le \sum_{h=\nu+1}^{\infty} \beta^{h-1} (R^{**}-r^{**})$$

$$= \beta^{\nu} \left(\frac{R^{**}-r^{**}}{1-\beta} \right) < \xi . \tag{4.31}$$

So $g_i^k(a, \psi)$ is a continuous function of a on the compact set $[0, 1]$ and for each k there exists an $a_k^* \in [0, 1]$ such that

$$g_i^k(a_k^*, \psi) = \sup_{0 \le a \le 1} \{g_i^k(a, \psi)\} \quad . \tag{4.32}$$

Letting $d^*(k)$ denote the inverse image of $a_k^* = a_k^*(d^*)$,

$$v_i(\psi) = \max_{1 \le k \le K_1} \{v_i(k, d^*(k), \psi)\} \quad , \tag{4.33}$$

and there exists a strategy $D^* = D^*(i, k^*, d^*(k^*))$ such that

$$v_i(\psi) = v_i(D^*, \psi) \quad . \qquad \text{QED.} \tag{4.34}$$

Theorem 4.2. If the set of functions $\{v_i(n, \psi)\}$ is defined by equation (4.9) then the limits

$$\lim_{n \to \infty} v_i(n, \psi) = v_i(\psi), \qquad i = 1, 2, \ldots N \tag{4.35}$$

exist and $\{v_i(\psi)\}$ is a set of solutions to equation (4.10).

Proof. It will be established inductively that for arbitrary positive integers n and m,

$$|v_i(n, \psi) - v_i(m, \psi)| \le \frac{\beta^n - \beta^m}{1-\beta} R^{**} \quad ,$$

$$i = 1, 2, \ldots N, \qquad n, m = 0, 1, 2, \ldots \quad , \tag{4.36}$$

where $R^{**} = \max_{i,j,k,\ell} \{|R_{i,j}^k(\ell)|\}$. Because $0 \le \beta < 1$ it follows by the

Cauchy criterion that $\lim_{n \to \infty} v_i(n, \psi)$ exists for $i = 1, 2, \ldots N$. By allowing n to go to ∞ in equation (4.9), it follows that the limiting functions satisfy equation (4.10). Let

$$S_i^k(v, n, \psi) = q_i^k(\psi) + \sum_{j=1}^{N} p_{i,j}^k \sum_{R} \ell_{i,j}^k(R; \psi) v_j(n, T_{i,j}^k(R; \psi)) \tag{4.37}$$

To establish equation (4.32) let

$$v_i(n, \psi) = S_i^{\alpha}(v, n-1, \psi) = \max_{1 \le k \le K_i} \{S_i^k(v, n-1, \psi)\} ,$$

$$v_i(m, \psi) = S_i^{\delta}(v, m-1, \psi) = \max_{1 \le k \le K_i} \{S_i^k(v, m-1, \psi)\} ,$$

then

$$v_i(n, \psi) - v_i(m, \psi) \le S_i^{\alpha}(v, n-1, \psi) - S_i^{\alpha}(v, m-1, \psi) ,$$

$$v_i(n, \psi) - v_i(m, \psi) \ge S_i^{\delta}(v, n-1, \psi) - S_i^{\delta}(v, n-1, \psi) . \tag{4.38}$$

Let k^* index the larger of $|S_i^{\alpha}(v, n-1, \psi) - S_i^{\alpha}(v, m-1, \psi)|$ and $|S_i^{\delta}(v, n-1, \psi) - S_i^{\delta}(v, m-1, \psi)|$. Then

$$|v_i(n, \psi) - v_i(m, \psi)| \le |S_i^{k^*}(v, n-1, \psi) - S_i^{k^*}(v, m-1, \psi)|$$

$$\le \beta \sum_{j=1}^{N} p_{i,j}^{k^*} \sum_{R} \ell_{i,j}^{k^*}(R; \psi) |v_j(n-1, T_{i,j}^{k^*}(R; \psi)$$

$$- v_j(m-1, T_{i,j}^{k*}(R, \psi))| \quad ,$$

$$i = 1, 2, \ldots N, \qquad n, m = 0, 1, 2, \ldots \tag{4.39}$$

Assuming that $v_i(0, \psi) = 0$, $i = 1, 2, \ldots N$, then

$$|v_i(n, \psi)| \leq \sum_{h=1}^{n} \beta^{h-1} R^{**} = \frac{1-\beta^n}{1-\beta} R^{**} \quad . \tag{4.40}$$

Therefore, assuming that $n \geq m$,

$$|v_i(n-m, \psi)| \leq \frac{1-\beta^{n-m}}{1-\beta} R^{**} \quad . \tag{4.41}$$

An inductive argument using equations (4.39) and (4.41) shows that

$$|v_i(n, \psi) - v_i(m, \psi)| \leq \frac{\beta^m-\beta^n}{1-\beta} R^{**} \quad , \tag{4.42}$$

and a similar argument for the case $m > n$ yields equation (4.36). Q.E.D. Proofs of the remaining theorems in this section will not be given. Modifications of the proofs to Theorems 4.1 and 4.2 are typical of those necessary for the remaining theorems, and the required proofs follow almost directly from Martin.

Theorem 4.3. There exists a unique set of functions $v_i(\psi)$ which satisfies the set of equations

$$v_i(\psi) = \max_{1 \le k \le K_i} \{q_i^k(\psi) + \alpha \sum_{j=1}^{N} p_{i,j}^k \sum_{R} \ell_{i,j}^k(R;\psi) v_j(T_{i,j}^k(R;\psi))\}$$

$$i = 1, 2, \ldots N , \qquad R = R(1), R(2) . \tag{4.43}$$

Theorem 4.4. If $\{v_i(\psi)\}$ is the unique bounded set of functions which satisfy equation (4.10) and if $\ell_{i,j}^k(R; \psi)$ is a continuous function of ψ ($k = 1, 2, \ldots K_i$; $i, j = 1, 2, \ldots N$), then $v_i(\psi)$ is a continuous function of ψ ($i = 1, 2, \ldots N$).

It has now been shown that the set of solutions $\{v_i(\psi)\}$ to equation (4.10) exist and are unique, and that there is an optimal strategy D^* which will achieve $\{v_i(\psi)\}$. The decision maker would like a method of determining, or at least approximating, the set of solutions $\{v_i(\psi)\}$. A more immediate problem facing the decision maker is the choice of the alternative to govern the next transition. Before proceeding to Chapter V and a discussion of these problems, three additional theorems from Bayesian Decision Problems and Markov Chains (16, p. 44-50), but modified to apply to the model of this chapter, will be stated.

Martin has developed a bound for the error function $|e(n, \psi)| = |v_i(\psi) - v_i(n, \psi)|$. The bound converges monotonically to zero, and n can be chosen such that the resulting error bound is small enough to make $v_i(n, \psi)$ a satisfactory approximation to $v_i(\psi)$. The following theorems concern this bound.

Theorem 4.5. The value $v_i(\psi)$ has the bounds

$$\frac{r^*}{1-\beta} \le v_i(\psi) \le \frac{R^*}{1-\beta} , \tag{4.44}$$

where $r^* = \max_{i,j,k,\ell} \{R^k_{i,j}(\ell)\}$ and $r^* = \min_{i,j,k,\ell} \{R^k_{i,j}(\ell)\}$.

Theorem 4.6. Let $v_i(n, \psi)$, as defined in equation (4.9), be a sequence of successive approximations. Then the error term of the n^{th} approximation has the bound

$$|e(n, \psi)| \leq \beta^n(\max \{ \frac{R^*}{1-\beta} ; \frac{-r^*}{1-\beta} \}) \quad . \tag{4.45}$$

Theorem 4.7. Let the generalized state (i, ψ) be fixed and let $\lambda(i, \psi) \epsilon \Delta(i)$ denote the set of optimal strategies for the Markovian decision process of equation (4.10). If $D^*(i, n)$ is an optimal strategy for the problem defined by equation (4.9) then, as $n \to \infty$, $D^*(i, n)$ ultimately lies in $\lambda(i, \psi)$.

V. CALCULATION OF THE SET OF SOLUTIONS $\{v_i(\psi)\}$ TO THE MARKOVIAN DECISION PROCESS WITH UNCERTAIN BERNOULLI REWARDS

Theorem 4.5 provides a bound for the error term $|e(n, \psi)| = |v_i(\psi) - v_i(n, \psi)|$ which is a monotonically decreasing function of n. The value $v_i(\psi)$ of equation (4.10) can be approximated by calculating $v_i(n, \psi)$ of equation (4.9), with n chosen large enough to reduce the bound of the error term to a magnitude acceptable to the decision maker. The practicality of this method of solution is seriously limited because of the excessive time required to calculate $v_i(n, \psi)$. Consider the simple "2×2" problem in which the system consists of two states (N=2), and there are two alternatives available in each state ($K_i = 2$, $i = 1, 2$). For fixed k the equation describing $v_i(n, \psi)$ contains four different values of $v_j(n-1, T^k_{i,j}(R, \psi))$ and since $k = (1,2)$, there are a total of eight values of $v_j(n-1, T^k_{i,j}(k, \psi))$ which must be calculated. Each of these in turn generates eight additional values until $v_j(1, T^k_{i,j}(R, \psi))$, which requires only two calculations, is reached. Solution of $v_i(n, \psi)$ for the "2×2" case therefore requires $8^{n-1} \cdot 2$ separate calculations.

The bound of $|e(n, \psi)|$ is a function of the discount factor, β. If β is the present worth factor for a compounding period of one, then $\beta = 1/(1 + i)$ where i is the effective rate of interest. A typical rate of interest might be 10% per year, so that $\beta \approx 0.9$. Should the time interval between transitions be less than one year, β will be greater than 0.9 if the annual rate of interest of 10% is to be maintained. Assuming $\beta = 0.9$ to be typical, then β^n converges rather slowly. For example, ten iterations would reduce the initial error bound by a factor

of approximately 0.35, while the number of separate calculations for the "2×2" case necessary to calculate $v_i(10, \psi)$ would be approximately 2.7×10^8. These considerations may create some concern about the practicability of applying the Markovian decision process with uncertain Bernoulli rewards to a real world problem unless a better method of solution can be developed. The purpose of this chapter is to develop bounds for $v_i(\psi)$ which are relatively quick to calculate, and to develop a method of determining the alternative which, for a fixed ψ, should be chosen to govern the next transition.

The Markovian decision process with uncertain Bernoulli rewards is conceptually similar to the discounted model discussed by Howard (12, p. 76-91), except that Howard assumed the rewards to be known constants. Let

$A^k_{i,j}$ = the reward received due to transition from state i to state j when alternative k is chosen to govern the transition.

A = a $1 \times L$ vector containing the rewards $A^k_{i,j}$ for all i, j and k. (5.1)

When the state transition horizon is infinite and the system is in state i, Howard has defined $u_i(\psi)$ as the expected value of the sum of future discounted rewards under an optimal strategy, where

$$u_i(A) = \max_{1 \le k \le K_i} \left\{ \sum_{j=1}^{N} p^k_{i,j} A^k_{i,j} + \beta \sum_{j=1}^{N} p^k_{i,j} u_j(A) \right\}$$

$$i = 1, 2, \ldots N, \qquad (5.2)$$

and has shown that the set of solutions $\{u_i(A)\}$ exist. An important contribution by Howard (12, p. 76-87) was the development of the value determination operation and the policy improvement routine, which together form an iterative method of calculating the set of solutions $\{u_i(A)\}$ to equation (5.2). From equation (5.2) it is clear that the alternative to be chosen to govern the next transition depends only on i, the current state of the system. Denote the state stationary strategy which yields $\{u_i(A)\}$ by $\Omega(A) = (\omega_1, \omega_2, \ldots \omega_N)$, where ω_i is the alternative to be chosen when the system is in state i. With reference to section C of Chapter IV, the decision maker always chooses the same policy vector and ignores past history.

Consider another similar Markovian decision process; suppose that the rewards are random variables whose distributions are known. Let

$f^k_{i,j}(.; \lambda)$ = the density function of the reward received due to transition from state i to state j when alternative k is chosen to govern the transition.

Λ = a $1 \times L$ vector containing the parameters of $f^k_{i,j}(.; \lambda)$ for all i, j and k. (5.3)

Then the following equation, which is analogous to equations (4.10) and (5.2), can be written

$$x_i(\Lambda) = \max_{1 \leq k \leq K_i} \{ \sum_{j=1}^{N} p^k_{i,j} \int_R R \, f^k_{i,j}(R; \lambda) dR$$

$$+ \beta \sum_{j=1}^{N} p_{i,j}^{k} \int_{R} f_{i,j}^{k}(R; \lambda) x_j(\Lambda) dR \Big\} ,$$

$$i = 1, 2, \ldots N . \qquad (5.4)$$

Denote the expected value of a reward by

$$E_{i,j}^{k}(R) = \int_{R} R \; f_{i,j}^{k}(R; \lambda) dR$$

and let

$$E(R_\Lambda) = \text{a } 1 \times L \text{ vector containing the expected values } E_{i,j}^{k}(R) \text{ for all i, j and k.} \qquad (5.5)$$

It will be shown that the set of solutions $\{x_i(\Lambda)\} = \{u_i(E(R_\Lambda))\}$.

Since $x_i(\Lambda)$ is a constant, equation (5.4) can be written

$$x_i(\Lambda) = \max_{1 \le k \le K_i} \Big\{ \sum_{j=1}^{N} p_{i,j}^{k} \int_{R} R \, f_{i,j}^{k}(R; \lambda) dR$$

$$+ \beta \sum_{j=1}^{N} p_{i,j}^{k} \, x_j(\Lambda) \int_{R} f_{i,j}^{k}(R; \lambda) dR \Big\}$$

$$= \max_{1 \le k \le K_i} \Big\{ \sum_{j=1}^{N} p_{i,j}^{k} \, E_{i,j}^{k}(R) + \beta \sum_{j=1}^{N} p_{i,j}^{k} \, x_j(\Lambda) \Big\}$$

$$i = 1, 2, \ldots N . \qquad (5.6)$$

Substitute $E^k_{i,j}(R)$ for $A^k_{i,j}$ in equation (5.2). Then $u_i(E(R_\Lambda))$ of equation (5.2) is of the same functional form as $x_i(\Lambda)$ of equation (5.6) so that $u_i(E(R_\Lambda)) = x_i(\Lambda)$, $i = 1, 2, \ldots N$. Therefore the Markovian decision model described by (5.3) and equation (5.4) is equivalent to the discounted model discussed by Howard.

The preceding discussion leads to a method of obtaining a lower bound for $v_i(\psi)$ which, in most situations, will be much larger than the lower bound given in Theorem 4.5. The expected value $E^k_{i,j}(R; \psi)$ was defined in equation (4.7). Let

$$E(R; \psi) = \text{a } 1\times L \text{ vector containing the values } E^k_{i,j}(R; \psi) \text{ for all } i, j \text{ and } k. \tag{5.7}$$

The state stationary strategy $\Omega(E(R; \psi))$ is the optimal strategy associated with the set of solutions $\{u_i(E(R; \psi))\}$ to equation (5.2).

Theorem 5.1. The value of a Markovian decision process with uncertain Bernoulli rewards, given that the state transition horizon is infinite and that the state stationary strategy $\Omega(E(R; \psi))$ is used, is, from equation (4.8)

$$w_i(\Omega(E(R;\psi)), \psi) = q_i^{\omega_i} + \beta \sum_{j=1}^{N} p_{i,j}^{\omega_i} \sum_{R} \ell_{i,j}^{\omega_i}(R; \psi) \times w_j(\Omega(E(R; \psi)), T_{i,j}^{\omega_i}(R, \psi)) ,$$

$$i = 1, 2, \ldots N, \quad R = R(1), R(2) . \tag{5.8}$$

If $u_i(E(R;\psi))$ is the solution to equation (5.2), then $w_i(\Omega(E(R;\psi)))$ $= u_i(E(R;\psi))$, $i = 1, 2, \ldots N$.

Proof. To simplify notation let $\Omega(E(R;\psi)) = (\omega_1, \omega_2, \ldots \omega_N) = (\omega)$. It will be established inductively that

$$\omega_{i_0}(\Omega(R;\psi)),\psi) = \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1} E^{\omega}_{i_0,i_1}(R_1;\psi)$$

$$+ \beta \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1} \left[\sum_{i_2=1}^{N} p^{\omega}_{i_1,i_2} E^{\omega}_{i_1,i_2}(R_2;\psi) \right.$$

$$+ \beta \sum_{i_2=1}^{N} p^{\omega}_{i_1,i_2} \left[\sum_{i_3=1}^{N} p^{\omega}_{i_2,i_3} E^{\omega}_{i_2,i_3}(R_3;\psi) \right.$$

$$+ \ldots$$

$$i_0 = 1, 2, \ldots N. \tag{5.9}$$

By letting $E^{\omega}_{i,j}(R;\psi) = A^{\omega}_{i,j}$ it is clear from equation (5.9) that $w_i(\Omega(E(R;\psi)),\psi) = u_i(E(R;\psi))$.

Let $\Omega^n(E(R;\psi))$ denote a n state transition horizon even though the state stationary strategy is independent of the number of transitions. Then equation (4.8) can be used to establish equation (5.9).

$$w_{i_0}(\Omega^1(E(R;\psi)),\psi) = \sum_{i_1=1}^{N} p^k_{i_0,i_1} \sum_{R_1} R_1 \, \ell^\omega_{i_0,i_1}(R_1;\psi)$$

$$= \sum_{i_1=1}^{N} p^\omega_{i_0,i_1} \, E^\omega_{i_0,i_1}(R;\psi);$$

$$w_{i_0}(\Omega^2(E(R;\psi)),\psi) = \sum_{i_1=1}^{N} p^\omega_{i_0,i_1} \sum_{R_1} R_1 \, \ell^\omega_{i_0,i_1}(R_1;\psi)$$

$$+ \beta \sum_{i_1=1}^{N} p^\omega_{i_0,i_1} \sum_{R_1} \ell^\omega_{i_0,i_1}(R_1;\psi) \left[\sum_{i_2=1}^{N} p^\omega_{i_1,i_2} \right.$$

$$\left. \times \, E^\omega_{i_1,i_2}(R_2; T^\omega_{i_0,i_1}(R_1,\psi)) \right] . \qquad (5.10)$$

It is necessary to show that

$$\sum_{R_1} \ell^\omega_{i,j}(R_1;\psi) \left[\sum_{h=1}^{N} p^\omega_{j,h} \, E^\omega_{j,h}(R_2; T^\omega_{i,j}(R_1,\psi)) \right]$$

$$= \sum_{h=1}^{N} p^\omega_{j,h} \, E^\omega_{j,h}(R_2;\psi) . \qquad (5.11)$$

Equation (5.11) can be written

$$\sum_{R_1} \ell^\omega_{i,j}(R_1;\psi) \left[\sum_{h=1}^{N} p^\omega_{j,h} \sum_{R_2} \int_m R_2 \, \ell^\omega_{j,h}(R_2/m) \phi^\omega_{j,h}(m/R_1;\psi) \, dm \right], \qquad (5.12)$$

where $\phi^{\omega}_{j,h}(m/R_1;\ \psi)$ is the posterior distribution of m given R_1 and is determined by the application of Bayes theorem as shown in equation (4.5). As mentioned in the discussion following equation (2.24), the posterior distribution of m differs from the prior distribution only when the reward observed is sampled from the distribution with indices identical to those of the prior distribution of m. If $i \neq j$ then $\phi^{\omega}_{j,h}(m/R_1;\ \psi) = \phi^{\omega}_{j,h}(m;\ \psi)$ since the likelihood of R_1 is $\ell^{\omega}_{i,j}(R;\ \psi)$. When $i \neq j$ equation (5.12) can be written

$$\sum_{R_1} \ell^{\omega}_{i,j}(R_1;\ \psi)\left[\sum_{h=1}^{N} p^{\omega}_{j,h} \sum_{R_2} \int_{m} R\ \ell^{\omega}_{j,h}(R_2/m)\phi^{\omega}_{j,h}(m;\ \psi)\,dm\right]$$

$$= \sum_{R_1} \ell^{\omega}_{i,j}(R_1;\ \psi)\left[\sum_{h=1}^{N} p^{\omega}_{j,h}\ E^{\omega}_{j,h}(R_2;\ \psi)\right]$$

$$= \sum_{h=1}^{N} p^{\omega}_{j,h}\ E^{\omega}_{j,h}(R;\ \psi) \quad . \tag{5.13}$$

For the case $i = j$ and $h \neq j$ the preceding argument is valid, and when $i = j = h$, that element of equation (5.12) is

$$\sum_{R_1} \ell^{\omega}_{i,i}(R_1;\ \psi)\ p^{\omega}_{i,i} \sum_{R_2} \int_{m} \ell^{\omega}_{i,i}(R_2/m)\phi^{\omega}_{i,i}(m/R_1)\,dm$$

$$= p^{\omega}_{i,i}\left[\frac{s}{t}\left[R_2(1)\ \frac{s+1}{t+1} + R_2(2)\ \frac{t-s}{t+1}\right]\right.$$

$$+ \frac{t-s}{t}\left[R_2(1)\frac{s}{t+1} + R_2(2)\frac{t+1-s}{t+1}\right]\Bigg]$$

$$= p^{\omega}_{i,i}\left[\frac{s}{t}R_2(1) + \frac{t-s}{t}R_2(2)\right] = p^{\omega}_{i,i}E^{\omega}_{i,i}(R;\psi) \quad . \tag{5.14}$$

Using the results of equations (5.13) and (5.14), the second equation of (5.10) can be written

$$w_{i_0}(\Omega^2(E(R;\psi)),\psi) = \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1}E^{\omega}_{i_0,i_1}(R_1;\psi)$$

$$+ \beta \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1}\left[\sum_{i_1=1}^{N} p^{\omega}_{i_1,i_2}E^{\omega}_{i_1,i_2}(R_2;\psi)\right] ;$$

$$w_{i_0}(\Omega^3(E(R;\psi)),\psi) = \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1}E^{\omega}_{i_0,i_1}(R_1;\psi)$$

$$+ \beta \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1}\sum_{R_1} \ell^{\omega}_{i_0,i_1}(R_1;\psi)\left[\sum_{i_1=1}^{N} p^{\omega}_{i_1,i_2}\right.$$

$$\times \left. E^{\omega}_{i_1,i_2}(R_2; T^{\omega}_{i_0,i_1}(R_1,\psi))\right]$$

$$+ \beta \sum_{i_2=1}^{N} p^{\omega}_{i_1,i_2}\sum_{R_2} \ell^{\omega}_{i_1,i_2}(R_2; T^{\omega}_{i_0,i_1}(R_1,\psi))$$

$$\times \left[\sum_{i_j=1}^{N} p^{\omega}_{i_2,i_3} E^{\omega}_{i_2,i_3}(R_3; T^{\omega}_{i_1,i_2}(R_2, T^{\omega}_{i_0,i_1}(R_1, \psi))) \right] \Bigg]$$

$$- \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1} E^{\omega}_{i_0,i_1}(R_1; \psi)$$

$$+ \beta \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1} \left[\sum_{i_2=1}^{N} p^{\omega}_{i_1,i_2} E^{\omega}_{i_1,i_2}(R_2; \psi) \right.$$

$$+ \beta \sum_{R_1} \ell_{i_0,i_1}(R_1; \psi) \sum_{i_2=1}^{N} p^{\omega}_{i_1,i_2} \left[\sum_{i_3=1}^{N} p^{\omega}_{i_2,i_3} \right.$$

$$\times E^{\omega}_{i_2,i_3}(R_3; T^{\omega}_{i_0,i_1}(R_1, \psi)) \Bigg] \Bigg]$$

$$- \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1} E^{\omega}_{i_0,i_1}(R_1; \psi)$$

$$+ \beta \sum_{i_1=1}^{N} p^{\omega}_{i_0,i_1} \left[\sum_{i_2=1}^{N} p^{\omega}_{i_1,i_2} E^{\omega}_{i_1,i_2}(R_2; \psi) \right.$$

$$+ \beta \sum_{i_2=1}^{N} p^{\omega}_{i_1,i_2} \left[\sum_{i_3=1}^{N} p^{\omega}_{i_2,i_3} E^{\omega}_{i_2,i_3}(R_3; \psi) \right] \Bigg] . \quad (5.15)$$

An inductive argument using equations (5.13), (5.14) and (5.15) establishes

equation (5.9) for the infinite state transition case. Q.E.D.

A theorem concerning the lower bound of $v_i(\psi)$ is now stated.

Theorem 5.2. Let $u_i(E(R; \psi))$ be the solution to equation (5.2). Then $v_i(\psi)$, the solution to equation (4.10), has the lower bound

$$u_i(E(R; \psi)) \leq v_i(\psi) \quad ,$$

$$i = 1, 2, \ldots N \quad . \tag{5.16}$$

Proof. It was shown in Theorem 5.1 that $u_i(E(R; \psi)) = w_i(\Omega(E(R; \psi), \psi)$; by equation (2.18) $v_i(\psi) \geq w_i(\Omega(E(R; \psi), \psi)$ since $\Omega(E(R; \psi)) \epsilon \Delta(i)$. Therefore $v_i(\psi) \geq w_i(\Omega(E(R; \psi), \psi) = u_i(E(R; \psi))$. Q.E.D.

Both the values $v_i(\psi)$ and $v_i(T^k_{i,j}(R, \psi))$ appear in the recursive equation (4.10). Before developing an upper bound of $v_i(\psi)$, it is necessary to obtain bounds for $v_i(\psi) - v_i(T^k_{i,j}(R, \psi))$.

Theorem 5.3. If $T^{k*}_{i*,j*}(R(\ell*), \psi) = \psi'$ is such that

$$E^{k*}_{i*,j*}(R; \psi) > E^{k*}_{i*,j*}(R; \psi') \quad , \tag{5.17}$$

and if $R(1)$ and $R(2)$ are the possible rewards from the distribution indexed by $i*$, $j*$ and $k*$ then

$$0 \leq v_{i*}(\psi) - v_{i*}(\psi') < p^{k*}_{i*,j*} \frac{|R(1)-R(2)|}{1-\beta} \quad ,$$

$$i* = 1, 2, \ldots N \quad , \tag{5.18}$$

where $v_i(\psi)$ is defined in equation (4.10).

Proof. Without loss of generality assume that $i^* = 1$, $j^* = 1$ and $\ell^* = 2$. The $1 \times L$ vectors ψ and ψ' differ only with respect to the parameters of the prior distribution indexed by 1,1 and k^* so that

$$E^k_{i,j}(R;\ \psi) = E^k_{i,j}(R;\ \psi'),\ \ell^k_{i,j}(R;\ \psi) = \ell^k_{i,j}(R;\ \psi'),$$

$$\text{all } i,\ j \text{ and } k \neq 1,1 \text{ and } k^* \ , \tag{5.19}$$

and

$$\ell^{k^*}_{1,1}(R(1);\ \psi) > \ell^{k^*}_{1,1}(R(1);\ \psi') \ . \tag{5.20}$$

Under the initial assumptions of the proof, the inequalities (5.17) and (5.20) imply that $R(1) > R(2)$ for the distribution indexed by 1,1 and k^*. From equations (3.7) and (4.7)

$$q^{k^*}_1(\psi) - q^{k^*}_1(\psi') = \sum_{j=1}^{N} p^{k^*}_{1,j}(E^{k^*}_{1,j}(R;\ \psi) - E^{k^*}_{1,j}(R;\ \psi'))$$

$$= p^{k^*}_{1,1}(E^{k^*}_{1,1}(R;\ \psi) - E^{k^*}_{1,1}(R;\ \psi')) > 0 \ ;$$

$$q^{k^*}_1(\psi) - q^{k^*}_1(\psi') < p^{k^*}_{1,1}(\max_{\psi\in\Psi}\{E^{k^*}_{1,j}(R;\ \psi)\} - \min_{\psi\in\Psi}\{E^{k^*}_{1,j}(R;\ \psi)\})$$

$$= p^{k^*}_{1,1}\ |R(1) - R(2)| = \Delta^{k^*}_{1,1}\ q \ ,$$

and

$$q_1^k(\psi) - q_1^k(\psi') = 0, \quad \text{all } k \neq k^* ,$$

$$q_i^k(\psi) - q_i^k(\psi') = 0, \quad \text{all } i \neq i^* . \tag{5.21}$$

Let

$$S_i^k(n, \psi_*) = q_i^k(\psi_*) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_{R} \ell_{i,j}^k(R; \psi_*)$$

$$\times v_j(n, T_{i,j}^k(R, \psi_*)) ,$$

$$i = 1, 2, \ldots N, \quad R = R(1), R(2), \quad \psi_* = \psi, \psi', \tag{5.22}$$

and use the following notation,

$$v_i(n, \psi) = \max_{1 \leq k \leq K_i} \{S_i^k(n-1, \psi)\} = S_i^{k_1}(n-1, \psi) ,$$

$$v_i(n, \psi') = \max_{1 \leq k \leq K_i} \{S_i^k(n-1, \psi')\} = S_i^{k_2}(n-1, \psi') , \tag{5.23}$$

where $v_i(n, \psi)$ is defined in equation (4.9). The following inequality is developed using the above notation.

$$v_i(n, \psi) - v_i(n, \psi') = S_i^{k_1}(n-1, \psi) - S_i^{k_2}(n-1, \psi')$$

$$\leq S_i^{k_1}(n-1, \psi) - S_i^{k_1}(n-1, \psi') . \tag{5.24}$$

The inequality (5.18) will be established inductively. When $n = 1$ and assuming $k_1 = k^*$,

$$s_1^{k^*}(0, \psi) - s_1^{k^*}(0, \psi') = q_1^{k^*}(\psi) - q_1^{k^*}(\psi') < \Delta_{1,1}^{k^*} q \quad . \quad (5.25)$$

Next assume $k_1 \neq k^*$,

$$s_1^{k_1}(0, \psi) - s_1^{k_1}(0, \psi) = q_1^{k_1}(\psi) - q_1^{k_1}(\psi)$$

$$= 0 \quad ; \quad (5.26)$$

from equations (5.24), (5.25) and (5.26)

$$v_1(1, \psi) - v_1(1, \psi') \leq q_1^{k^*}(\psi) - q_1^{k^*}(\psi') < \Delta_{1,1}^{k^*} q \quad . \quad (5.27)$$

When $i \neq 1$,

$$v_i(1, \psi) - v_i(1, \psi') \leq s_i^{k_1}(0, \psi) - s_i^{k_1}(0, \psi')$$

$$= q_i^{k_1}(\psi) - q_i^{k_1}(\psi') = 0 \quad . \quad (5.28)$$

Next let $n = 2$ and assume $k_1 = k^*$,

$$s_1^{k^*}(1, \psi) - s_1^{k^*}(1, \psi') = q_1^{k^*}(\psi) - q_1^{k^*}(\psi')$$

$$+ \beta \left[p_{1,1}^{k*} \left[\sum_{R} \ell_{1,1}^{k*}(R; \psi) v_1(1, T_{1,1}^{k*}(R, \psi)) \right. \right.$$

$$\left. - \sum_{R} \ell_{1,1}^{k*}(R; \psi') v_1(1, T_{1,1}^{k*}(R, \psi')) \right]$$

$$+ \sum_{j=2}^{N} p_{1,j}^{k*} \sum_{R} \ell_{1,j}^{k*}(R; \psi)(v_j(1, T_{1,j}^{k*}(R, \psi))$$

$$\left. - v_j(1, T_{1,j}^{k*}(R, \psi'))) \right] . \qquad (5.29)$$

If two vectors ψ_1 and ψ_2 differ only with respect to the parameters of the distribution indexed by $1,1$ and k^* and are such that $E_{1,1}^{k*}(R; \psi_1) > E_{1,1}^{k*}(R; \psi_2)$ then, in a manner identical to that used in equation (5.21), it can be shown that

$$0 < q_i^{k*}(\psi_1) - q_i^{k*}(\psi_2) < p_{1,1}^{k*}(E_{1,1}^{k*}(\psi_1) - E_{1,1}^{k*}(\psi_2)) < \Delta_{1,1}^{k*} q \quad . \qquad (5.30)$$

From condition (5.17) and equations (5.27) and (5.30),

$$v_1(1, T_{1,1}^{k*}(R(1), \psi_*)) \geq v_1(1, \psi_*) \geq v_1(1, T_{1,1}^{k*}(R(2), \psi_*)), \psi_* = \psi, \psi'$$

$$\Downarrow$$

$$\sum_{R} \ell_{1,1}^{k*}(R; \psi) v_i(1, T_{1,1}^{k*}(R, \psi)) \leq v_1(1, T_{1,1}^{k*}(R(1), \psi)) \quad ,$$

$$\sum_{R} \ell_{1,1}^{k*}(R; \psi') v_1(1, T_{1,1}^{k*}(R, \psi')) \geq v_i(1, T_{1,1}^{k*}(R(2), \psi)) \quad . \qquad (5.31)$$

Using the inequalities of (5.28) and (5.31), and since $\sum_{j=1}^{N} p_{i,j}^{k} = 1$,

$k = 1, 2, \ldots K_i$, $i = 1, 2, \ldots N$, equation (5.29) can be written

$$S_1^{k*}(1, \psi) - S_1^{k*}(1, \psi') = q_1^{k*}(\psi) - q_1^{k*}(\psi')$$

$$+ \beta \Bigg[p_{1,1}^{k*}(v_1(1, T_{1,1}^{k*}(R(1), \psi)) - v_1(1, T_{1,1}^{k*}(R(2), \psi')))$$

$$+ \sum_{j=2}^{N} p_{1,j}^{k*} \sum_{R} \ell_{1,j}^{k*}(R; \psi)(v_j(1, T_{i,j}^{k*}(R, \psi))$$

$$- v_j(1, T_{1,j}^{k*}(R, \psi'))) \Bigg]$$

$$\leq q_1^{k*}(\psi) - q_1^{k*}(\psi') + \beta(q_1^{k*}(T_{1,1}^{k*}(R(1), \psi))$$

$$- q_1^{k*}(T_{1,1}^{k*}(R(2), \psi')))$$

$$< \Delta_{1,1}^{k*} q + \beta \cdot \Delta_{1,1}^{k*} q \quad . \tag{5.32}$$

Assuming $k_1 \neq k^*$, and from equations (5.27) and (5.28),

$$S_1^{k_1}(1, \psi) - S_1^{k_1}(1, \psi') = q_1^{k_1}(\psi) - q_1^{k_1}(\psi')$$

$$+ \beta \sum_{j=1}^{N} p_{1,j}^{k_1} \sum_{R} \ell_{1,j}^{k_1}(R; \psi)(v_j(1, T_{1,j}^{k_1}(R, \psi))$$

$$- v_j(1, T_{1,j}^{k_1}(R, \psi)))$$

$$\leq \beta \, (q_1^{k*}(\psi) - q_1^{k*}(\psi'))$$

$$< \beta \cdot \Delta_{1,1}^{k*} \, q \quad . \tag{5.33}$$

From equations (5.22), (5.32) and (5.33),

$$v_1(2, \psi) - v_1(2, \psi') \leq q_1^{k*}(\psi) - q_1^{k*}(\psi')$$

$$+ \beta(q_1^{k*}(T_{1,1}^{k*}(R(1), \psi)) - q_1^{k*}(T_{1,1}^{k*}(R(2), \psi')))$$

$$< \Delta_{1,1}^{k*} \, q + \beta \cdot \Delta_{1,1}^{k*} \, q \quad , \tag{5.34}$$

and when $i \neq 1$, from equations (5.22) and (5.27),

$$v_i(2, \psi) - v_i(2, \psi') \leq s_i^{k_1}(1, \psi) - s_i^{k_1}(1, \psi')$$

$$= q_i^{k_1}(\psi) - q_i^{k_1}(\psi')$$

$$+ \beta \sum_{j=1}^{N} p_{i,j}^{k_1} \sum_{R} \ell_{i,j}^{k_1}(R; \psi)(v_j(1, T_{i,j}^{k_1}(R, \psi))$$

$$- v_j(1, T_{i,j}^{k_1}(R, \psi')))$$

$$\leq \beta(q_i^{k^*}(\psi) - q_i^{k^*}(\psi'))$$

$$< \Delta_{1,1}^{k^*} q \quad . \tag{5.35}$$

When ψ is the vector containing the parameters of the prior distributions, denote the vector of parameters of the posterior distributions given n observations of reward $R(\ell)$ from the distribution indexed by i, j and k by $T_{i,j}^{k}(R(\ell)^n, \psi)$. An inductive argument using equations (5.27), (5.34) and (5.35) establishes that

$$v_1(n, \psi) - v_1(n, \psi') \leq q_1^{k^*}(\psi) - q_1^{k^*}(\psi')$$

$$+ \beta(q_1^{k^*}(T_{1,1}^{k^*}(R(1), \psi)) - q_1^{k^*}(T_{1,1}^{k^*}(R(2), \psi')))$$

$$+ \beta^2(q_1^{k*}(T_{1,1}^{k*}(R(1)^2, \psi)) - q_1^{k*}(T_{1,1}^{k*}(R(2)^2, \psi')))$$

$$+ \ldots + \beta^{n-1}(q_1^{k*}(T_{1,1}^{k*}(R(1)^{n-1}, \psi))$$

$$- q_1^{k*}(T_{1,1}^{k*}(R(2)^{n-1}, \psi')))$$

$$< \Delta_{1,1}^{k*} q(1 + \beta + \beta^2 + \ldots + \beta^{n-1}) \quad . \tag{5.36}$$

To obtain the upper bound for $v_1(\psi) - v_1(\psi')$ let $n \to \infty$ in equation (5.36).

$$v_1(\psi) - v_1(\psi') < \sum_{h=0}^{\infty} \Delta_{1,1}^{k*} q \cdot \beta^h = \Delta_{1,1}^{k*} q \cdot \frac{1}{1-\beta}$$

$$= p_{1,1}^{k*} \frac{|R(1)-R(2)|}{1-\beta} \tag{5.37}$$

The lower bound of $v_i(\psi) - v_i(\psi')$ can be established by first writing the inequality

$$v_1(n, \psi) - v_1(n, \psi') = S_1^{k_1}(n-1, \psi) - S_1^{k_2}(n-1, \psi')$$

$$\geq S_1^{k_2}(n-1, \psi) - S_1^{k_2}(n-1, \psi') \quad . \tag{5.38}$$

Let $n = 1$ and assume first that $k_2 = k^*$, then that $k_2 \neq k^*$. Since $q_1^{k*}(\psi) - q_1^{k*}(\psi') > 0$, and from equations (5.25), (5.26) and (5.38),

$$v_1(1, \psi) - v_1(1, \psi') \geq 0 \quad . \tag{5.39}$$

When $n = 2$ and assuming that alternative k^* is not chosen on either the first or second transition, then from equations (5.28), (5.33) and (5.38)

$$v_1(2, \psi) - v_1(2, \psi') \geq 0 \quad . \tag{5.40}$$

An inductive argument can be used to show that

$$v_1(\psi) - v_1(\psi') \geq 0 \quad . \quad \text{Q.E.D.} \tag{5.41}$$

Corollary 5.4. If $T^{k^*}_{i^*,j^*}(R(\ell^*), \tilde{\psi}) = \tilde{\psi}'$ is such that

$$E^{k^*}_{i^*,j^*}(R; \tilde{\psi}) < E^{k^*}_{i^*,j^*}(R; \tilde{\psi}') \quad , \tag{5.42}$$

and if $R(1)$ and $R(2)$ are the possible rewards from the distribution indexed by i^*, j^* and k^*, then

$$-p^{k^*}_{i^*,j^*} \frac{|R(1)-R(2)|}{1-\lambda} < v_{i^*}(\tilde{\psi}) - v_{i^*}(\tilde{\psi}') \leq 0 \quad ,$$

$$i^* = 1, 2, \ldots N, \tag{5.43}$$

where $v_i(\psi)$ is defined in equation (4.10).

Proof. The direct substitution of $\tilde{\psi}'$ for ψ and $\tilde{\psi}$ for ψ' in the proof of Theorem 5.3 is possible. Then

$$0 \leq v_{i*}(\tilde{\psi}') - v_{i*}(\tilde{\psi}) < p^{k*}_{i*,j*} \frac{|R(1)-R(2)|}{1-\beta}$$

$$\Downarrow$$

$$-p^{k*}_{i*,j*} \frac{|R(1)-R(2)|}{1-\beta} < v_{i*}(\tilde{\psi}) - v_{i*}(\tilde{\psi}') \leq 0 \quad . \quad \text{Q.E.D.} \quad (5.44)$$

It is possible to modify Theorem 5.3 and Corollary 5.4 to obtain a bound which is less than $p^{k*}_{i*,j*} \frac{|R(1)-R(2)|}{1-\beta}$. Suppose that the conditions of Theorem 5.3 are met and assume, without loss of generality that $i* = j* = 1$ and $R(1) > R(2)$. Let

$$\begin{aligned}
\Delta^{k*}_{1,1} q(n) &= q^{k*}_1(\psi) - q^{k*}_1(\psi') \\
&\quad + \beta(q^{k*}_1(T^{k*}_{1,1}(R(1), \psi)) - q^{k*}_1(T^{k*}_{1,1}(R(2), \psi'))) \\
&\quad + \beta^2(q^{k*}_1(T^{k*}_{1,1}(R(1)^2, \psi)) - q^{k*}_1(T^{k*}_{1,1}(R(2)^2, \psi'))) \\
&\quad + \ldots + \beta^{n-1}(q^{k*}_1(T^{k*}_{1,1}(R(1)^{n-1}, \psi)) \\
&\quad - q^{k*}_1(T^{k*}_{1,1}(R(2)^{n-1}, \psi'))) \quad , \\
&\qquad n = 1, 2, \ldots
\end{aligned} \quad (5.45)$$

From equations (5.36) and (5.37)

$$v_{i*}(\psi) - v_{i*}(\psi') < \Delta^{k*}_{i*,j*} q(n) + \Delta^{k*}_{i*,j*} q \cdot \sum_{h=n}^{\infty} \beta^h$$

$$= \Delta^{k*}_{i*,j*}\, q(n) + \Delta^{k*}_{i*,j*}\, q \cdot \frac{\beta^n}{1-\beta}$$

$$< \Delta^{k*}_{i*,j*}\, q \cdot \frac{1}{1-\beta} = p^{k*}_{i*,j*}\, \frac{|R(1)-R(2)|}{1-\beta},$$

$$n = 1, 2, \ldots \qquad (5.46)$$

Therefore the quantity $p^{k*}_{i*,j*}\, \frac{|R(1)-R(2)|}{1-\beta}$ can be replaced by $\Delta^{k*}_{i*,j*}\, q(n) + p^{k*}_{i*,j*}\, |R(1) - R(2)|\, \frac{\beta^n}{1-\beta}$ in Theorem 5.3 and Corollary 5.4.

The bounds for $v_i(\psi) - v_i(T^k_{i,j}(R, \psi))$ will be used to obtain an upper bound for $v_i(\psi)$. Letting $T^k_{i,j}(R, \psi) = \psi'$, the results of Theorem 5.3 and Corollary 5.4 can be written

$$v_i(\psi') = v_i(\psi) + \theta^k_{i,j}(R, \psi)\, \Delta v_i(\psi'),$$

where $\Delta v_i(\psi') = p^k_{i,j}\, \frac{|R(1)-R(2)|}{1-\beta}$, $E^k_{i,j}(R; \psi) > E^k_{i,j}(R; \psi')$;

$$\Delta v_i(\psi') = -p^k_{i,j}\, \frac{|R(1)-R(2)|}{1-\beta}, \qquad E^k_{i,j}(R; \psi) < E^k_{i,j}(R; \psi');$$

$$0 \leq \theta^k_{i,j}(R; \psi) \leq 1,$$

$$i,j = 1, 2, \ldots N, \quad k = 1, 2, \ldots K_i, \quad R = R(1), R(2). \qquad (5.47)$$

Using the notation of (5.47), equation (4.10) can now be written

$$v_i(\psi) = \max_{0 \le k \le K_i} \{q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_{R} \ell_{i,j}^k(R; \psi)$$

$$\times (v_j(\psi) + \theta_{i,j}^k(R, \psi) \Delta v_j(T_{i,j}^k(R, \psi)))\} ,$$

$$i = 1, 2, \ldots N, \qquad R < R(1), R(2) . \tag{5.48}$$

Let $\Omega = (\omega_1, \omega_2, \ldots \omega_N) = (\omega)$ denote a state stationary strategy where ω_i is the alternative to be chosen when the system is in state i, and define $w_i(\Omega, \psi)$ as

$$w_i(\Omega, \psi) = q_i^{\omega}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{\omega} \sum_{R} \ell_{i,j}^{\omega}(R; \psi)$$

$$\times (w_j(\Omega, \psi) + \theta_{i,j}^{\omega}(R, \psi) \Delta v_j(T_{i,j}^{\omega}(R, \psi)))$$

$$= \sum_{j=1}^{N} p_{i,j}^{\omega} \sum_{R} \ell_{i,j}^{\omega}(R; \psi)(R + \beta \theta_{i,j}^{\omega}(R, \psi)$$

$$\times \Delta v_j(T_{i,j}^{\omega}(R, \psi))) + \beta \sum_{j=1}^{N} p_{i,j}^{\omega} w_j(\Omega, \psi) ,$$

$$i = 1, 2, \ldots N, \qquad R = R(1), R(2) . \tag{5.49}$$

To express the set of equations of (5.49) in matrix form let

$$W(\Omega, \psi) = \text{COL}\left[w_1(\Omega, \psi), w_2(\Omega, \psi), \ldots w_N(\Omega, \psi)\right] ,$$

$$P(\Omega) = \begin{bmatrix} p^{\omega}_{1,1} & p^{\omega}_{1,2} & \cdot & \cdot & \cdot & p^{\omega}_{1,N} \\ p^{\omega}_{2,1} & p^{\omega}_{2,2} & \cdot & \cdot & \cdot & p^{\omega}_{2,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p^{\omega}_{N,1} & p^{\omega}_{N,2} & \cdot & \cdot & \cdot & p^{\omega}_{N,N} \end{bmatrix}$$

$$D(\Omega, \psi) = \text{COL}\left[d_1^{\omega}(\psi), d_2^{\omega}(\psi), \ldots d_N^{\omega}(\psi)\right],$$

where

$$d_i^{\omega}(\psi) = \sum_{j=1}^{N} p^{\omega}_{i,j} \sum_{R} \ell^{\omega}_{i,j}(R; \psi)(R + \beta\, \theta^{\omega}_{i,j}(R, \psi) \times \Delta v_j(T^{\omega}_{i,j}(R, \psi))) \quad . \tag{5.50}$$

Express the set of equations (5.49) as

$$W(\Omega, \psi) = D(\Omega, \psi) + \beta\, P(\Omega)\, W(\Omega, \psi) \quad ,$$

$$W(\Omega, \psi) - \beta\, P(\Omega)\, W(\Omega, \psi) = D(\Omega, \psi)$$

$$\left[I - \beta\, P(\Omega)\right] W(\Omega, \psi) = D(\Omega, \psi) \quad ,$$

$$W(\Omega, \psi) = \left[I - \beta\, P(\Omega)\right]^{-1} D(\Omega, \psi) \quad . \tag{5.51}$$

Howard (12, p. 82) has shown that $\left[I - \beta\, P(\Omega)\right]^{-1}$ exists and has non-negative elements. Note that equation (5.49) is of the same functional form as equation (5.48), and that the set of all state stationary strategies, denoted by $\Delta(\Omega)$, is finite. Calculation of $W(\Omega, \psi)$ is equivalent to Howard's value determination operation, and the policy improvement routine will determine Ω^* such that

$$v_i(\psi) = \max_{\Omega \in \Delta(\Omega)} \{w_i(\Omega, \psi)\} = w_i(\Omega^*, \psi) \quad ,$$

$$i = 1, 2, \ldots N \quad . \tag{5.52}$$

Letting $\Omega^* = (\omega_1^*, \omega_2^*, \ldots \omega_N^*) = (\omega^*)$ and the $N \times N$ matrix $\left[I - \beta\, P(\Omega^*)\right]^{-1} = C = \left[c_{i,j}\right]$, $i, j = 1, 2, \ldots N$, then

$$v_i(\psi) = \sum_{j=1}^{N} c_{i,j}\, d_j^{\omega^*}$$

$$= \sum_{j=1}^{N} c_{i,j} \sum_{h=1}^{N} p_{j,h}^{\omega^*} \sum_{R} \ell_{j,h}^{\omega^*}(R;\, \psi)$$

$$\times \left(R + \beta\, \theta_{j,h}^{\omega^*}(R,\, \psi) \Delta v_h(T_{j,h}^{\omega^*}(R,\, \psi))\right)$$

$$< \sum_{j=1}^{N} c_{i,j} \sum_{h=1}^{N} p_{j,h}^{\omega^*} \sum_{R} \ell_{j,h}^{\omega^*}(R;\, \psi)$$

$$\times \left(R + \beta \max_{0 \le \theta^{\omega*}_{j,h}(R,\psi) \le 1} \{\theta^{\omega*}_{j,h}(R, \psi)\Delta v_h(T^{\omega*}_{j,h}(R, \psi))\}\right)$$

$$= \sum_{j=1}^{N} c_{i,h} \sum_{h=1}^{N} p^{\omega*}_{j,h} \sum_{R} \ell^{\omega*}_{j,h}(R; \psi)$$

$$\times (R + \beta\{\max 0; \Delta v_h(T^{\omega*}_{j,h}(R, \psi))\}) \quad ,$$

$$i = 1, 2, \ldots N, \qquad R = R(1), R(2) \quad . \tag{5.53}$$

Let

$$F^k_{i,j}(\psi) = \sum_{R} \ell^k_{i,j}(R; \psi)(R + \beta \max\{0; \Delta v_j(T^k_{i,j}(R, \psi))\})$$

$$= L^k_{i,j}(R; \psi) + \sum_{R} \ell^k_{i,j}(R; \psi)\ \beta \max\{0; \Delta v_j(T^k_{i,j}(R, \psi))\},$$

$$\hat{q}^k_i(\psi) = \sum_{j=1}^{N} p^k_{i,j}\ F^k_{i,j}(\psi) \quad , \tag{5.54}$$

and, for a state stationary strategy Ω, let

$$\hat{Q}(\Omega, \psi) = \text{COL}\left[\hat{q}^{\omega}_1(\psi), \hat{q}^{\omega}_2(\psi), \ldots \hat{q}^{\omega}_n(\psi)\right] \quad ,$$

$$U(\Omega, \psi) = \text{COL}\left[u_1(\Omega,\psi), u_2(\Omega, \psi), \ldots u_n(\Omega, \psi)\right] \quad . \tag{5.55}$$

Then, for some $\Omega\varepsilon\Delta(\Omega)$, the last line of equation (5.53) can be set equal to $u_i(\Omega^*, \psi)$ and written in matrix form as

$$U(\Omega^*, \psi) = \left[I - \beta\, P(\Omega^*)\right]^{-1} \hat{Q}(\Omega^*, \psi) \quad ,$$

$$U(\Omega^*, \psi) = \hat{Q}(\Omega^*, \psi) + \beta\, P(\Omega^*)\, U(\Omega^*, \psi) \quad , \tag{5.56}$$

so that

$$u_i(\Omega^*, \psi) = \sum_{j=1}^{N} p_{i,j}^{\omega^*} F_{i,j}^{\omega^*}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{\omega^*} u_i(\Omega^*, \psi) \quad ,$$

$$i = 1, 2, \ldots N \quad . \tag{5.57}$$

Because the $\theta_{i,j}^{k}(R, \psi)$ of equation (5.49) are unknown, the strategy Ω^* which was defined in equation (5.52) is also unknown. There is, however, a strategy $\hat{\Omega}$ such that $u_i(\hat{\Omega}, \psi) = \max_{\Omega\varepsilon\Delta(\Omega)} \{u_i(\Omega, \psi)\} \geq u_i(\Omega^*, \psi)$. Let

$$F(\psi) = \text{a } 1\times L \text{ vector containing the values } F_{i,j}^{k}(\psi) \text{ for all i, j and k.} \tag{5.58}$$

Since, for fixed k, equation (5.2) is of the same functional form as equation (5.57), the following equation can be written.

$$u_i(\psi^*, \psi) \leq u_i(\hat{\Omega}, \psi) = u_i(F(\psi))$$

$$= \max_{1 \le k \le K_i} \left\{ \sum_{j=1}^{N} p_{i,j}^{k} \, F_{i,j}^{k}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k} \, u_j(F(\psi)) \right\}$$

$$i = 1, 2, \ldots N, \tag{5.59}$$

and the set of solutions $\{u_i(F(\psi))\}$ can be obtained by application of Howard's iterative procedure.

Theorem 5.5. If $u_i(F(\psi))$ is the solution to equation (5.59), where $F(\psi)$ is defined in (5.58) and equation (5.54), and $v_i(\psi)$ is the solution to equation (4.10), then

$$v_i(\psi) < u_i(F(\psi)), \qquad i = 1, 2, \ldots N \ . \tag{5.60}$$

Proof. From equations (5.53) and (5.59),

$$v_i(\psi) < \sum_{j=1}^{N} c_{i,j} \sum_{h=1}^{N} p_{j,h}^{\omega^*} \sum_{R} \ell_{j,h}^{\omega^*}(R; \psi)$$

$$\times \left(R + \beta \max\{0; \Delta v_h(T_{j,h}^{\omega^*}(R, \psi))\}\right)$$

$$\le u_i(F(\psi)) \ ,$$

$$i = 1, 2, \ldots N \ . \qquad \text{Q.E.D.} \tag{5.61}$$

Theorems 5.1 and 5.5 establish the following bounds for $v_i(\psi)$,

$$u_i(E(R;\psi)) \leq v_i(\psi) < u_i(F(\psi)) ,$$

$$i = 1, 2, \ldots N . \tag{5.62}$$

Theorem 5.6. If the state stationary strategy associated with $u_i(E(R;\psi))$ is

$$\Omega(E(R;\psi)) = (\omega_1, \omega_2, \ldots \omega_N) , \tag{5.63}$$

and if

$$u_i(E(R;\psi)) \geq q_i^{k'}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k'} \sum_{R} \ell_{i,j}^{k'}(R;\psi)$$

$$\times u_j(F(T_{i,j}^{k'}(R,\psi))) ,$$

$$\text{where } k' \text{ is the set of all } k \neq \omega_i,\ k = 1, 2, \ldots K_i , \tag{5.64}$$

then

$$v_i(\psi) = q_i^{\omega_i}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{\omega_i} \sum_{R} \ell_{i,j}^{\omega_i}(R;\psi)$$

$$\times v_j(T_{i,j}^{\omega_i}(R,\psi)) . \tag{5.65}$$

Proof. Since it was shown in Theorem 5.5 that $v_i(\psi) < u_i(F(\psi))$, then

$$q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k} \sum_{R} \ell_{i,j}^{k}(R;\psi) v_j(T_{i,j}^{k},\psi)$$

$$< q_i^k(\psi) + \beta \sum_{j=1} p_{i,j}^k \sum_R \ell_{i,j}^k(R;\psi) u_j(F(T_{i,j}^k(R,\psi)))$$

$$i = 1, 2, \ldots N, \qquad k = 1, 2, \ldots K_i \ . \tag{5.66}$$

Theorem 5.1 showed that $v_i(\psi) \geq u_i(E(R;\psi))$. If condition (5.64) is true for all k', then, since $k' \cup \omega_i = k$, $k = 1, 2, \ldots K_i$,

$$v_i(\psi) \geq u_i(E(R;\psi)) \geq q_i^{k'}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k'} \sum_R \ell_{i,j}^{k'}(R;\psi)$$

$$\times u_j(F(T_{i,j}^{k'}(R,\psi)))$$

$$> q_i^{k'}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k'} \sum_R \ell_{i,j}^{k'}(R;\psi) v_j(T_{i,j}^{k'}(R,\psi))$$

$$\Downarrow$$

$$v_i(\psi) = q_i^{\omega_i}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{\omega_i} \sum_R \ell_{i,j}^{\omega_i}(R;\psi) v_j(T_{i,j}^{\omega_i}(R,\psi)).$$

Q.E.D. (5.67)

If the conditions of Theorem 5.6 are met the decision maker can determine the optimal alternative to govern the next transition. If the conditions are not met it may be desirable to recalculate $\Delta v_j(\psi')$ using the modified bounds for $v_i(\psi) - v_i(T_{i,j}^k(R,\psi))$ suggested in equation (5.46). This will reduce the value of $F_{i,j}^k(\psi)$ which in turn will reduce

$u_i(F(\psi))$; the change may be enough so that the modified calculations will meet the conditions of Theorem 5.6. If the conditions of Theorem 5.6 can not be met, it is necessary to revert to a solution which is a variation of solution by successive approximations.

Theorem 5.7. If $v_i(n, \psi, L)$ and $v_i(n, \psi, U)$ are defined as

$$v_i(n, \psi, L) = \max_{1 \le k \le K_i} \{q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_{R} \ell_{i,j}^k(R; \psi) \times v_j(n-1, T_{i,j}^k(R, \psi), L)\} ,$$

$$n = 1, 2, \ldots ,$$

$$v_i(0, \psi, L) = u_i(E(R, \psi)) ,$$

and

$$v_i(n, \psi, U) = \max_{1 \le k \le K_i} \{q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_{R} \ell_{i,j}^k(R; \psi) \times v_j(n-1, T_{i,j}^k(R, \psi), U)\} ,$$

$$n = 1, 2, \ldots ,$$

$$v_i(0, \psi, U) = u_i(F(\psi)) ,$$

$$i = 1, 2, \ldots N, \quad R = R(1), R(2) , \qquad (5.68)$$

and if n is such that

$$v_i(n, \psi, L) = q_i^{k*}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k*} \sum_{R} \ell_{i,j}^{k*}(R; \psi)$$

$$\times v_j(n-1, T_{i,j}^{k*}(R, \psi), L) \quad ,$$

and

$$v_i(n, \psi, L) \geq q_i^{k'}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k'} \sum_{R} \ell_{i,j}^{k'}(R; \psi)$$

$$\times v_j(n-1, T_{i,j}^{k'}(R, \psi), U) \quad ,$$

where k' is the set of all $k \neq k^*$, $k = 1, 2, \ldots K_i$, (5.69)

then

$$v_i(\psi) = q_i^{k*}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k*} \sum_{R} \ell_{i,j}^{k*}(R; \psi) v_j(T_{i,j}^{k*}(R; \psi)) \quad . \tag{5.70}$$

Proof. An inductive argument will be used to establish that

$$v_i(n, \psi, U) > v_i(\psi) \geq v_i(n, \psi, L) \quad . \tag{5.71}$$

Theorem 5.1 shows that $v_i(\psi) \geq u_i(E(R; \psi))$, so that

$$q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_R \ell_{i,j}^k(R;\ \psi) u_j(E(R;\ T_{i,j}^k(R,\ \psi)))$$

$$\leq q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_R \ell_{i,j}^k(R;\ \psi) v_j(T_{i,j}^k(R;\ \psi)),$$

$$k = 1, 2, \ldots N$$

$$\Downarrow$$

$$v_i(1,\ \psi,\ L) \leq v_i(\psi)\ ;$$

$$q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_R \ell_{i,j}^k(R;\ \psi) v_j(1,\ T_{i,j}^k(R,\ \psi),\ L)$$

$$\leq q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_R \ell_{i,j}^k(R;\ \psi) v_j(T_{i,j}^k(R,\ \psi)),$$

$$k = 1, 2, \ldots N$$

$$\Downarrow$$

$$v_i(2,\ \psi,\ L) \leq v_i(\psi);$$

and by induction

$$q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_R \ell_{i,j}^k(R;\psi) v_j(n-1, T_{i,j}^k(R;\psi), L)$$

$$\leq q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_R \ell_{i,j}^k(R;\psi) v_j(T_{i,j}^k(R,\psi)),$$

$$k = 1, 2, \ldots N$$

$$\Downarrow$$

$$v_i(n, \psi, L) \leq v_i(\psi) \quad . \tag{5.72}$$

By a similar argument it is easily shown that

$$q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_R \ell_{i,j}^k(R;\psi) v_j(n-1, T_{i,j}^k(R;\psi), U)$$

$$> q_i^k(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^k \sum_R \ell_{i,j}^k(R;\psi) v_j(T_{i,j}^k(R;\psi)),$$

$$k = 1, 2, \ldots N$$

$$\Downarrow$$

$$v_i(n, \psi, U) > v_i(\psi) \quad . \tag{5.73}$$

Suppose that there is a n such that condition (5.69) is met. Since

$k^* \cup k' = k,\ k = 1, 2, \ldots K_i,$

$$v_i(\psi) > v_i(n, \psi, L) \geq q_i^{k'}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k'} \sum_{R} \ell_{i,j}^{k'}(R; \psi)$$

$$\times v_j(n-1, T_{i,j}^{k'}(R, \psi), U)$$

$$> q_i^{k'}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k'} \sum_{R} \ell_{i,j}^{k'}(R; \psi) v_j(T_{i,j}^{k'}(R, \psi))$$

$$\Downarrow$$

$$v_i(\psi) = q_i^{k^*}(\psi) + \beta \sum_{j=1}^{N} p_{i,j}^{k^*} \sum_{R} \ell_{i,j}^{k^*}(R; \psi) v_j(T_{i,j}^{k^*}(R, \psi)).$$

Q.E.D. (5.74)

VI. LITERATURE CITED

1. Anderson, T. W. and Goodman, Leo A. Statistical inference about Markov chains. Annals of Mathematical Statistics 28: 89-110. 1957.

2. Bellman, R. A Markovian decision process. Journal of Mathematics and Mechanics 6: 679-684. 1957.

3. Bellman, Richard. Dynamic programming. Princeton, New Jersey, Princeton University Press. 1957.

4. Billingsley, Patrick. Statistical inference for Markov processes. Chicago, Illinois, University of Chicago Press. c1961.

5. Box, G. E. P. and Hill, W. J. Discrimination among mechanistic models. Technometrics 9: 57-71. 1967.

6. Bradt, R. N., Johnson, S. M., and Karlin, S. On sequential designs for maximizing the sum of n observations. Annals of Mathematical Statistics 27: 1060-1074. 1956.

7. Derman, Cyrus. On sequential decisions and Markov chains. Management Science 9: 16-24. 1963.

8. Derman, C. and Lieberman, G. J. A Markovian decision process for a joint replacement and stocking problem. Management Science 13: 607-617. 1967.

9. Feldman, Dorian. Contributions to the "two-armed bandit" problem. Annals of Mathematical Statistics 33: 847-856. 1962.

10. Herniter, Jerome D. and Magee, John F. Customer behavior as a Markov process. Operations Research 9: 105-122. 1961.

11. Howard, Ronald A. Dynamic programming. Management Science 12: 317-348. 1966.

12. Howard, Ronald A. Dynamic programming and Markov processes. Cambridge, Massachusetts, The M.I.T. Press. c1960.

13. Klein, Morton. Inspection-maintenance-replacement schedules under Markovian deterioration. Management Science 9: 25-32. 1963.

14. Klein, Morton. Markovian decision models for reject allowance problems. Management Science 12: 349-358. 1966.

15. Manne, Alan S. Linear programming and sequential decisions. Management Science 6: 259-267. 1960.

16. Martin, J. J. Bayesian decision problems and Markov chains. New York, New York, John Wiley and Sons, Inc. c1967.

17. Raiffa, H. and R. Schlaifer. Applied statistical decision theory. Boston, Massachusetts, Graduate School of Business Administration, Harvard University. 1961.

18. Robbins, Herbert. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society 58: 527-535. 1952.

19. Silver, E. A. Markovian decision processes with uncertain transition probabilities or rewards. Technical Report No. 1, Research in the control of complex systems. Cambridge, Massachusetts, Operations Research Center, Massachusetts Institute of Technology. August 1963.

20. Wagner, Harvey M. On the optimality of pure strategies. Management Science 6: 268-269. 1960.

VII. ACKNOWLEDGMENTS

The author is indebted to many persons who have, both directly and indirectly, assisted and encouraged him during this period of graduate study. Special acknowledgment is given Professor H. T. David for his guidance, counsel and encouragement throughout the preparation of this dissertation. The author wishes to thank his wife, Priscilla, for her understanding and patience during his years of study.