Continuous Time Markovian Sequential
Control Processes

by

S. S. Chitgopekar

F.S.U. Statistics Report M127

December, 1967
Department of Statistics
Florida State University
Tallahassee, Florida

Work supported by the Office of Naval Research
Contract No. NONR 988(13)

Reproduction in whole or in part is permitted for
any purpose of the United States Government.
Section I. Introduction and Summary.

Consider a stochastic system which at any time can be in any one of several possible states. Having observed the system in some state, one could take one of several possible actions. As a result, the system moves to some other state according to some law of motion, depending on the state of the system and the action taken. Associated with the system is a reward function, depending on the state of the system, the action taken and also possibly on the state to which the system moves next.

Interest lies in the problems of existence and nature of "optimal" policies—a policy being a rule of action—according to different chosen criteria. These problems, variously referred to in the literature as Dynamic Programming problems, Surveillance problems, Markovian Sequential Control Processes etc., have received considerable attention during the last few years. See, for example, Blackwell [1962, 1965], Derman [1962, 1966], Maitra [1965], Strauch [1966], Howard [1960, 1963] and Jewell [1963a, 1963b].

The work of Blackwell, Derman, Maitra and Strauch referred to above assumes that the system makes transitions at discrete, uniformly spaced intervals of time and the actions are taken at the instants of transitions. In many applications, however, the system may make transitions at random time intervals i.e., the "waiting time" of the system in any state is a non-negative random variable. Howard [1960, 1963] and Jewell [1963a, 1963b] have considered this problem.

Howard [1963] considers a system with a finite state space and finite action space. In state i, under action k, the system earns at the rate $a_{ik}$ per unit of time and by a transition to state j, it earns a reward $r_{ij}$. Under
action $k$, $h_{ij}^k(\cdot)$ is the probability density function of the waiting time in state $i$ given that the next transition is to state $j$. The probability of transition from state $i$ to state $j$ under action $k$ is $p_{ij}^k$.

Howard restricts attention to a finite class of stationary policies where the actions are taken at the instants of transitions only. He develops a simple, finite, iterative technique to find optimal stationary policies. Jewell [1963a, 1963b] considers essentially the same problem as Howard with similar results. Fox [1967] has generalised Howard's results to the case where the action space is not necessarily finite.

In this study, in Sections 2 - 7, we consider a system similar to Howard [1963]. However, unlike Howard, Jewell and Fox, we allow actions to be taken between transitions and we are interested in the existence and nature of optimal policies in a very wide class of policies.

A policy $S$ is a rule of action that specifies the action to be taken at any instant of time, possibly randomized and possibly depending on the entire past history of the system. If a policy $S$ is such that there is a positive probability of an action being taken between transitions, following Antelman, Russell and Savage [1967], we shall say that the policy $S$ involves "hesitation". Let $G(\cdot)$ be a distribution function (on $[0,\infty]$) such that $G(t)$ denotes the probability that an action would be taken by time $t$ since the previous action, given that there is no transition before then. We shall refer to such a distribution $G(\cdot)$ as a "hesitation distribution".

In Sections 2 - 6, we consider the case when the cost of taking any action is positive; in Section 7, the case where all these costs are zero; and in Section 8, we consider the case when there is a positive cost of taking an observation. These cases will be referred to as the costly actions case, the costless actions case and the costly observations case respectively.
In the costly actions case, we shall restrict ourselves to the class of policies $S$ where for any policy $S$ in $S$, the actions to be taken form a sequence with probability one. In the costless actions case, we restrict ourselves to the subclass $S^\alpha \subset S$, where for any policy in $S^\alpha$, the hesitation distributions involved are distributions on $[\alpha, \infty]$ with $\alpha > 0$. All the policies in $S$ necessarily take actions at the instants of transitions. In the costly observations case, we restrict ourselves to the class of policies $R^\Delta$ where for any policy in $R^\Delta$, the observations taken form a sequence and the distributions of the times between successive observations are distributions on $[0, \Delta]$ with $\Delta < \infty$.

Remark. Our first restriction on the class of policies to be considered is that the actions taken form a sequence with probability one. An example of a policy that is eliminated from consideration because of this is the following: Two actions are available to us in the initial state of the system and the policy takes the first action at every rational time point and the second action at every irrational time point.

For the costly actions case and the costly observations case, such policies would result in an expected income of $-\infty$ in a finite length of time and hence may be deemed undesirable.

For the costly as well as costless actions cases, for any policy $S$, the criterion of interest is $I(S)$ given by

\[(1.1) \quad I(S) = \liminf_{N \to \infty} I_N(S) = \liminf_{N \to \infty} \frac{\sum_{n=1}^{N} E_i(n)}{\sum_{n=1}^{N} E_T(n)}\]

where $i_n(S)$ is the income earned under the $n^{th}$ action and $T_n(S)$ is the time spent by the system under the $n^{th}$ action of the policy $S$.

For the costly observations case, the criterion of interest for any
policy S is I(S), given by

\[(1.2) \quad I(S) = \lim \inf_{N \to \infty} I_N(S) = \lim \inf_{N \to \infty} \frac{\sum_{n=1}^{N} E_i(S)}{\sum_{n=1}^{N} EY(n)}\]

where \(Y_n(S)\) is the time from the \(n^{th}\) to the \(n+1^{th}\) observation and \(i_n(S)\) is the income earned during this period. Let

\[(1.3) \quad I^* = \begin{cases} 
\sup_{S \in S} I(S) & \text{for the costly actions case,} \\
\sup_{S \in S^c} I(S) & \text{for the costless actions case,} \\
\sup_{S \in S^d} I(S) & \text{for the costly observations case.}
\end{cases}\]

We are interested in the existence and nature of policies \(S^*\) such that

\[(1.4) \quad I^* = I(S^*).\]

In Section 2, we introduce the model for the costly actions case and state the main result—Theorem 2.1—that there exists a non-randomized, stationary policy, involving only one-point hesitation distributions, that is optimal in \(S\).

Section 3 gives some formulae that are used repeatedly in the following sections and some preliminary theorems that are used later.

Section 4 gives an approximation theorem which is crucial in the proof of Theorem 2.1.

In Section 5 we consider the subclass of stationary policies and show that, (Theorem 5.2) in this class, we need only consider one-point hesitation distributions and further (Theorem 5.3) that there exists an optimal policy in this class. The section ends with an example showing the necessity of the assumption of continuity of the waiting time distribution functions.
In Section 6 we give the proof of Theorem 2.1 and as a Corollary (Corollary 2.2) we show that when the waiting time distributions are exponential, hesitation can be eliminated. We give an example to show that hesitation cannot always be eliminated.

In Section 7 we consider the costless actions case and show (Theorem 7.1) that there exists a non-randomized stationary policy that is optimal in $S^\alpha$. We investigate situations when this optimal policy would be independent of $\alpha$, for $\alpha$ sufficiently close to zero.

In Section 8, we consider the costly observations case and show (Theorem 8.5) that there exists a non-randomized stationary policy $S^*_\Delta$ that is optimal in $R^\Delta$. This optimal policy involves only one-point observation distributions. We show that for sufficiently large $\Delta$, $S^*_\Delta$ is independent of $\Delta$ and as such, the restriction that the observation distributions are distributions on $[0,\Delta]$ is of no serious consequence.

In Section 9, we show that for the three cases considered, $I^*$ has certain monotonic properties (Theorem 9.1). In Theorem 9.2, we show the equality of various criteria of interest for stationary policies in the three cases considered.

In the Appendix, we quote a result in Markov Chains theory from Chung [1966] and deduce some results that are relevant to our problem and are used in Sections 1 - 9.

Section 2. Costly actions; the model.

We are interested in a stochastic system which at any time can be in one of a finite number of $L$ states. In each of these $L$ states, we have $K (\leq \infty)$ alternative actions available to us. In state $i$, it costs $c^k_i (>0)$ to take action $k$ and then the system earns at the rate $a^k_i$ per unit of time; $i =
1, 2, ..., L, \( k = 1, 2, \ldots, K \). The probability of transition from state \( i \) to state \( j \), under action \( k \), in the absence of any further intervening actions, is denoted by \( p_{ij}^k \). We have

\[
(2.1) \quad p_{ij}^k \geq 0 \quad \text{for all } i, j, k,
\]

\[
\sum_{j=1}^{L} p_{ij}^k = 1 \quad \text{for all } i, k.
\]

We assume that between actions, the waiting time to transition is a random variable with a continuous distribution function with a finite mean and depending only on the current state of the system and the action taken. We denote the waiting time in state \( i \) under action \( k \) by \( T_i^k \), its distribution function by \( F_i^k(\cdot) \) and its mean by \( \theta(i,k) \); \( k = 1, 2, \ldots, K \); \( i = 1, 2, \ldots, L \). We further assume that

\[
F_i^k(0) = 0; \quad F_i^k(x) > 0 \quad \text{for all } x > 0
\]

and

\[
\frac{d}{dx} F_i^k(x) \mid_{x=0} \text{ exists and is finite; } k = 1, 2, \ldots, K,
\]

\[
i = 1, 2, \ldots, L.
\]

We also assume that in each state \( i, i = 1, 2, \ldots, L \), there is at least one action \( k_i \) such that the resulting transition probability matrix \( P = (p_{ij}^k) \) is the transition probability matrix of an irreducible Markov Chain. This assumption will be discussed in the Appendix.

Throughout, we have made the restrictive assumption that the distribution of the waiting time to transition in any state under any action is independent of the state to which the system makes a transition next.
A consequence of this assumption is that the probabilities of transition from one given state to some other state are independent of any conditioning event about the transition times.

Any policy $S$ in $S$ is of the following type:

1) After any transition, $S$ specifies

(a) the action to be taken, possibly randomized and depending on the past history of the system; and

(b) a hesitation distribution, possibly depending on the past history of the system, and if there is a positive probability of the next action being taken before the next transition, $S$ specifies the next action to be taken, possibly randomized.

2) If an action has been taken before a transition, then the policy $S$ specifies only (b) above.

**Definition.** A policy $S$ is said to be stationary if for each $i$, $i = 1, 2, \ldots, L$, $S$ specifies a pair $(k_i, G_i)$ such that whenever an action is to be taken in state $i$, $S$ prescribes the action $k_i$ and the hesitation distribution $G_i(\cdot)$.

**Definition.** A hesitation distribution $G(\cdot)$ is said to be a one-point distribution if there exists an $x$, $0 \leq x \leq \infty$, such that

$$G(t) = \begin{cases} 0 & \text{for } t < x \\ 1 & \text{for } t \geq x. \end{cases}$$

We shall denote such a distribution by $G_x(\cdot)$.

**Definition.** If, on some occasion, a policy $S$ involves the hesitation distribution $G_0(\cdot)$, then we say that the policy $S$ involves instantaneous hesitation on that particular occasion.
Let $S_0 \subseteq S$ be the subclass of stationary policies. Natural questions that arise are: Does there exist an $S^*$ such that

$$I(S^*) = \sup_{S \in S} I(S)$$

where $I(S)$ is given by (1.1)? If such a $S^*$ exists, does $S^*$ belong to $S_0$?

The following theorem answers the above questions.

**Theorem 2.1.** There exists a non-randomized policy $S^*$ in $S_0$ that is optimal in $S$. Further, $S^*$ involves only one-point hesitation distributions.

**Corollary 2.2.** If all the waiting time distributions are exponential, then hesitation can be eliminated.

We give the proofs of Theorem 2.1 and Corollary 2.2 in Section 6.

Let

$$\theta^* = \max_{i,k} \theta(i,k)_{i,k}$$

$$\theta_* = \min_{i,k} \theta(i,k)$$

$$a^* = \max_{i,k} a_i$$

$$a_* = \min_{i,k} a_i$$

$$c^* = \max_{i,k} c_i$$

$$c_* = \min_{i,k} c_i$$

and

$$I^* = I(S^*)$$

**Lemma 2.3.** $a_* - \frac{c^*}{\theta^*} < I^* < a^* - \frac{c_*}{\theta^*}$. 
Proof. For any policy $S$, we have

$$I_N(S) = \frac{\sum_{n=1}^{N} E_i n(S)}{\sum_{n=1}^{N} E T n(S)}$$

$$= \left( \frac{\sum_{n=1}^{N} a_n E T n(S) - \sum_{n=1}^{N} c_n}{\sum_{n=1}^{N} E T n(S)} \right)$$

where $a_n$ is the earning rate of the $n^{th}$ action and $c_n$ is the cost of the $n^{th}$ action. Therefore,

$$I_N(S) \leq \frac{a - N c}{\sum_{n=1}^{N} E T n(S)}$$

$$\leq a - c/\theta^*.$$

Hence

$$I(S) = \lim_{N \to \infty} \inf I_N(S)$$

$$\leq a - c/\theta^*.$$

Now suppose $S$ does not involve any hesitation. Then

$$I_N(S) \geq a^* - c/\theta^*.$$ 

Hence

$$I(S) \geq a^* - c/\theta^*.$$ 

Hence the Lemma.

Let

$$\mu_0 = c^*/(a^* - a^* + c^* / \theta^*).$$

We then have

Lemma 2.4. Any policy $S$ for which $\lim \inf_{N \to \infty} \left\{ \frac{\sum_{n=1}^{N} E T n(S)}{N} \right\} < \mu_0$ is not optimal.

Proof. For any policy $S$ we have
\[ I_N(S) \leq a^* - c^* \sqrt{\sum_{n=1}^{N} E T_n(S)/N}. \]

Let \[ \mu(S) = \lim \inf_{N \to \infty} \left\{ \sum_{n=1}^{N} E T_n(S)/N \right\}. \]

Then

\[ (2.3) \quad I(S) \leq a^* - c^*/\mu(S). \]

If \( \mu(S) < \mu_0 \), we have

\[ I(S) < a^* - c^*/\mu_0 \quad \text{(from (2.3))} \]

\[ = a^* - (a^* - a^* + c^*/\theta^*), \quad \text{(from (2.2))} \]

\[ = a^* - c^*/\theta^* \]

\[ \leq I^*. \quad \text{(from Lemma 2.3)} \]

Hence the Lemma.

In view of the above Lemma, we shall only consider policies \( S \) for which \( \mu(S) \geq \mu_0 \).

Section 3. Some preliminary formulae and theorems.

In this Section we give some formulae and theorems that are used in the following sections. \( F(\cdot) \) is a waiting time distribution and \( G(\cdot) \) is a hesitation distribution. When \( F(\cdot) = F_i^k(\cdot) \), we shall replace \( F \) by \( i,k \) in all the expressions. Similarly, when \( G(\cdot) = G_t(\cdot) \), we shall replace \( G \) by \( t \) in all the expressions:

(3.1) \[ \Theta(F) = \text{the expected value of a random variable with distribution function } F(\cdot) \]

\[ = \int_0^\infty dF(t) = \int_0^\infty (1-F(t))dt. \]
\[(3.2) \eta(F;G) = \text{expected time to the next action when the waiting time}
\]
\[
distribution is } F(\cdot) \text{ and the hesitation distribution is } G(\cdot)
\]
\[
= \int_{0}^{\infty} t(1-G(t)) \, dF(t) + \int_{0}^{\infty} t(1-F(t)) \, dG(t)
\]
\[
= \int_{0}^{\infty} t \, dF(t) - \int_{0}^{\infty} (1-F(t)) \, G(t) \, dt
\]
\[
= \int_{0}^{\infty} (1-F(t))(1-G(t)) \, dt
\]
\[
= \int_{0}^{\infty} \{ \int_{0}^{x} (1-F(t)) \, dt \} \, dG(x).
\]
\[(3.3) \eta(F;\tau) = \int_{0}^{\tau} (1-F(x)) \, dx.
\]

From (3.3) and the last form of (3.2) we get
\[(3.4) \eta(F;G) = \int_{0}^{\infty} \eta(F;\tau) \, dG(\tau).
\]

\[(3.5) q(F;G) = \text{probability of a transition before the next action when the}
\]
\[
\text{waiting time distribution is } F(\cdot) \text{ and the hesitation distribution is } G(\cdot)
\]
\[
= \int_{0}^{\infty} F(t) \, dG(t).
\]

\[(3.6) q(F;\tau) = F(\tau).
\]

\[
N(F;G) = \text{expected number of actions before the next transition when}
\]
\[
\text{we hesitate repeatedly with the hesitation distribution}
\]
\[
G(\cdot) \text{ while the waiting time distribution is } F(\cdot).
\]

If \(q(F;G) > 0\), that is \(G(\cdot) \neq G_{0}(\cdot)\), we have
(3.7) \[ N(F;G) = 1 \cdot q(F;G) + 2 \cdot (1-q(F;G))q(F;G) + 3 \cdot (1-q(F;G))^2q(F;G) + \ldots \]
\[ = \frac{1}{q(F;G)}. \]

When \( q(F;G) = 0 \), we have

(3.8) \[ N(F;G) = \infty. \]

\[ \theta(F;G) = \text{expected time to transition when the waiting time distribution is } F(\cdot) \text{ and we hesitate repeatedly with the hesitation distribution } G(\cdot) \text{ until a transition occurs} \]
\[ = \int_0^\infty t(1-G(t)) dF(t) + \int_0^\infty (t+\theta(F;G))(1-F(t)) dG(t) \]
\[ = \eta(F;G) + \theta(F;G)(1-q(F;G)). \]

Thus, when \( q(F;G) > 0 \),

(3.9) \[ \theta(F;G) = \frac{\eta(F;G)}{q(F;G)}. \]

(3.10) \[ \theta(F;\dagger) = \int_0^\dagger (1-F(x)) dx / F(t). \]

Note that \( \theta(F) = \theta(F;\infty) = \eta(F;\infty) \). We shall define \( \theta(F;0) = \lim_{t \to 0} \theta(F;t) \).

Note that \( q(F;G) = 0 \) if, and only if, \( G(\cdot) = G_0(\cdot) \). Hence, when \( q(F;G) = 0 \), we have \( \theta(F;G) = \theta(F;0) \). Let \( \underline{G}(\cdot) \) and \( \bar{G}(\cdot) \) be such that

(3.11) \[ \theta(F;\underline{G}) = \min_G \theta(F;G) \]

(3.12) \[ \theta(F;\bar{G}) = \max_G \theta(F;G). \]

Lemma 3.1. \( \underline{G}(\cdot) \) and \( \bar{G}(\cdot) \) can be taken to be one-point distributions.

Proof. From (3.9), (3.4) and (3.5), we have
\[ \theta(F; G) = \int_0^\infty \eta(F; t) dG(t) \bigg/ \int_0^\infty F(t) dG(t) \]
\[ = \int_0^\infty \theta(F; t) F(t) dG(t) \bigg/ \int_0^\infty F(t) dG(t) \quad \text{(from (3.3) and (3.10))}. \]

Let \( x_0 \) be such that

\[ \theta(F; x_0) = \min_{x \in [0, \infty]} \theta(F; x). \]  
(3.13)

Then

\[ \theta(F; x_0) - \theta(F; G) = \int_0^\infty \{ \theta(F; x_0) - \theta(F; t) \} dG(t) \bigg/ \int_0^\infty F(t) dG(t) \leq 0 \quad \text{for all } G, \]

in view of (3.13). Similarly, if \( y_0 \) is such that

\[ \theta(F; y_0) = \max_{x \in [0, \infty]} \theta(F; x) \]  
(3.14)

then \( \theta(F; y_0) \geq \theta(F; G) \) for all \( G \). Take \( \bar{G}(\cdot) = G_{x_0}(\cdot) \) and \( \tilde{G}(\cdot) = G_{y_0}(\cdot) \).

Hence the Lemma.

**Definition.** A non-discrete distribution \( F(\cdot) \) is an **increasing failure rate distribution (I.F.R.)** if

\[ \frac{F(t+x) - F(t)}{1-F(t)} \]

is increasing in \( t \) for \( x > 0 \), \( t \geq 0 \) such that \( F(t) < 1 \). If \( F(\cdot) \) has density \( f(\cdot) \), let \( r(\cdot) = f(\cdot)/1-F(\cdot) \). Then \( F(\cdot) \) is I.F.R. if, and only if, \( r(\cdot) \) is increasing in \( t \). A **decreasing failure rate distribution (D.F.R.)** is defined in an analogous manner.
Lemma 3.2. Let $F(\cdot)$ be a distribution with density $f(\cdot)$. Then, for all $G(\cdot)$,

a) $\theta(F) \leq \theta(F;G) \leq 1/f(0)$ if $F(\cdot)$ is L.F.R.

b) $1/f(0) \leq \theta(F;G) \leq \theta(F)$ if $F(\cdot)$ is D.F.R.

Proof. In view of Lemma 3.1, we can restrict ourselves to $\{G(x): 0 \leq x \leq \infty\}$. We have (from (3.10))

$$\theta(F;x) = \frac{\int_{0}^{x} (1-F(t))dt}{F(x)}.$$

Therefore,

$$F^2(x) \frac{d}{dx} \theta(F;x) = F(x)(1-F(x)) - f(x) \int_{0}^{x} (1-F(t))dt$$

$$= (1-F(x))f(x) - f(x) \int_{0}^{x} (1-F(t))dt.$$ 

Let $m(x) = F(x) - r(x) \int_{0}^{x} (1-F(t))dt$. If $F(\cdot)$ is L.F.R., then

$$m(x) \leq F(x) - \int_{0}^{x} r(t)(1-F(t))dt = 0.$$

Similarly, if $F(\cdot)$ is D.F.R., then $m(x) \geq 0$.

Hence $\frac{d}{dx} \theta(F;x) \geq (\leq 0)$ if $F(\cdot)$ is D.F.R. (L.F.R.). Since $\theta(F;\infty) = \theta(F)$ and $\theta(F;0) = 1/f(0)$, the Lemma follows.

Corollary 3.3. If $F(\cdot)$ is exponential, then $\theta(F;G) = \theta(F)$ for all $G(\cdot)$.

Theorem 3.4. Let $G(\cdot)$ be a distribution function on $[0,\Delta]$, $\Delta < \infty$, and let $h(\cdot)$ be some monotone continuous function on $[0,\Delta]$. For any given

$$0 = \tau_0 < \tau_1 < \cdots < \tau_{N-1} < \tau_N = \Delta,$$

there exists a discrete distribution $G(\cdot)$ with its mass at the points $\tau_i$, $i = 0,1,\ldots,N$, and such that

$$\int_{0}^{\Delta} h(\tau)dG(\tau) = \int_{0}^{\Delta} h(\tau)dG_{\ast}(\tau).$$
Further, if \( m(\cdot) \) is a monotone continuous function such that

\[
(3.17) \quad |m(t_i) - m(t_{i-1})| \leq \varepsilon, \quad i = 1, 2, \ldots, N
\]

then

\[
(3.18) \quad \left| \int_0^\Delta m(t) dG(t) - \int_0^\Delta m(t) dG_\varepsilon(t) \right| \leq 2\varepsilon.
\]

**Proof.**

\[
\int_0^\Delta h(t) dG(t) = \sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} h(t) dG(t) + h(t_0)G(t_0)
\]

\[
= \sum_{i=1}^{N} \left\{ b_i h(t_i) + (1-b_i)h(t_{i-1}) \right\} (G(t_i) - G(t_{i-1})) + h(t_0)G(t_0)
\]

where \( 0 \leq b_i \leq 1, \quad i = 1, 2, \ldots, N. \)

\[
= \sum_{i=0}^{N} h(t_i) \delta_i
\]

where

\[
\delta_0 = (1-b_1) [G(t_1) - G(t_0)] + G(t_0)
\]

\[
\delta_i = b_i [G(t_i) - G(t_{i-1})] + (1-b_{i+1}) [G(t_{i+1}) - G(t_i)], \quad i = 1, 2, \ldots, N-1;
\]

\[
\delta_N = b_N [1 - G(t_{N-1})].
\]

It is easily seen that \( \delta_i \geq 0, \quad i = 0, 1, \ldots, N; \) and \( \sum_{i=0}^{N} \delta_i = 1. \) Take \( G_\varepsilon(\cdot) \) to be the discrete distribution with mass \( \delta_i \) at \( t_i \), \( i = 0, 1, \ldots, N. \)

Then (3.16) holds.

Now

\[
\int_0^\Delta m(t) dG(t) = \sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} m(t) dG(t) + m(t_0)G(t_0)
\]

\[
= \sum_{i=1}^{N} \left\{ m(t_i) + e_i [m(t_{i-1}) - m(t_i)] \right\} (G(t_i) - G(t_{i-1}))+m(t_0)G(t_0)
\]

where \( 0 \leq e_i \leq 1, \quad i = 1, 2, \ldots, N. \)

Hence
\[
\left| \int_0^\Delta m(t) dG(t) - \int_0^\Delta m(t) dG^i(t) \right|
\]

\[
= \left| \sum_{i=1}^N \{m(t_i)[G(t_i)-G(t_{i-1})]-\delta_i\} - m(t_0)[G(t_0)-G(t_{i-1})] + e_i[m(t_{i-1})-m(t_i)][G(t_i)-G(t_{i-1})] \right|
\]

\[
\leq \left| \sum_{i=1}^{N-1} \{m(t_i)[(1-b_i)(G(t_i)-G(t_{i-1})) - (1-b_{i+1})(G(t_{i+1})-G(t_i))] + m(t_N)(1-b_N)(1-G(t_{N-1})) + m(t_0)[G(t_0)-\delta_0] \right|
\]

\[
+ \left| \sum_{i=1}^N e_i[m(t_{i-1})-m(t_i)][G(t_i)-G(t_{i-1})] \right|
\]

\[
\leq \sum_{i=1}^N (1-b_i)[m(t_i)-m(t_{i-1})][G(t_i)-G(t_{i-1})] + \varepsilon
\]

\[
\leq 2\varepsilon \quad \text{in view of (3.17)}.
\]

Section 4. Costly actions; an approximation theorem.

For any \(\varepsilon > 0\), let \(N_{\varepsilon}\) be a positive integer such that \(\theta^*/N_{\varepsilon} < \varepsilon\). Let \(t^k_i(r)\) be such that

\[(4.1) \quad \eta(i,k;t^k_i(r)) = \frac{r}{N_{\varepsilon}} \theta(i,k) \quad r = 0,1,\ldots,N_{\varepsilon}; \quad k = 1,2,\ldots,K; \quad i = 1,2,\ldots,L.\]

Note that since \(\eta(i,k;x)\) is a continuous, increasing function of \(x\), for each \((i,k)\), (4.1) has solutions. If, in any case, there is more than one solution, we take \(t^k_i(r)\) to be the largest solution. Further,

\[(4.2) \quad \eta(i,k;t^k_i(r)) - \eta(i,k;t^k_i(r-1)) < \varepsilon \]

for all \(r\), for any given \((i,k)\).

Let \(G_{\varepsilon}\) denote the finite class of hesitation distributions \(\{\eta_x(\cdot)\}\) where \(x \in \{t^k_i(r)\}: \quad r = 0,1,\ldots,N_{\varepsilon}; \quad k = 1,2,\ldots,K; \quad i = 1,2,\ldots,L\).

Suppose under some policy \(S\), at some stage, action \(k\) has been taken in state \(i\) and a hesitation distribution \(G(\cdot)\) is used. Then the probability of a transition before the next action is given by \(\int_0^\infty F_i^k(t)dG(t)\) and the
expected time before the next action is given by $\int_0^\infty \eta(i,k;t) dG(t)$. In view of (4.2) and Theorem 3.4, there exists a discrete distribution $G_\varepsilon(\cdot)$ with mass at the points $t_i^k(r)$, $r = 0, 1, \ldots, N_\varepsilon$, such that replacing the hesitation distribution $G(\cdot)$ by $G_\varepsilon(\cdot)$ does not change the probability of transition before the next action and the expected time to the next action is changed by at most $2\varepsilon$.

Since the probability of transition before the next action is not changed by replacing $G(\cdot)$ by $G_\varepsilon(\cdot)$, in view of the observation on page 6, Section 2, the distribution of the state at the time of next action also remains the same when $G(\cdot)$ is replaced by $G_\varepsilon(\cdot)$. Further, observe that $G_\varepsilon(\cdot)$ is a randomization over the class $G_{\varepsilon^2}$.

**Definition.** A policy $S_\varepsilon$ will be said to be an $\varepsilon$-approximation of a policy $S$ if $|I(S) - I(S_\varepsilon)| \leq \varepsilon$.

**Theorem 4.1.** For any policy $S$ and any sufficiently small $\varepsilon > 0$, we can find a policy $S_\varepsilon$, an $\varepsilon$-approximation of $S$, such that $S_\varepsilon$ only involves hesitation distributions that are randomizations over the class $G_{\varepsilon^2}$.

**Proof.** Let $n_1^{th}$ be the first action at which $S$ involves hesitation with positive probability and let $G(\cdot)$ be the hesitation distribution prescribed by $S$. Let $G_\varepsilon(\cdot)$ be a randomization over $G_{\varepsilon^2}$ such that replacing $G(\cdot)$ by $G_\varepsilon(\cdot)$ does not change the distribution of states at the $n_1 + 1^{th}$ action and the time spent by the system under the $n_1^{th}$ action is changed by at most $2\varepsilon^2$. Replace $G(\cdot)$ by $G_\varepsilon(\cdot)$ and after reaching the $n_1 + 1^{th}$ state, create by randomization the time that would have been spent under the $n_1^{th}$ action if the hesitation distribution were $G(\cdot)$ and given the $n_1 + 1^{th}$ state. Maintain this as part of the history of the system. With this partially artificial history, take $n_1 + 1^{th}$
action as prescribed by \( S \) and proceed to follow \( S \). If \( n_2 \) action is the next action that involves hesitation with positive probability, again replace the hesitation distribution by a randomization over \( G_2 \) as above and proceed similarly.

Repeat the same procedure at every step where hesitation is involved.

Let \( S' \) be the resulting modified policy. Note that part of the history maintained is artificial in that we are not recording the actual time spent by the system under any action that involves hesitation but creating by randomization the time that would be spent by the system if the hesitation distribution prescribed by \( S \) were to be used—conditioned on the state of the system at the time of the next action. Since replacing \( G(\cdot) \) by a proper randomization over \( G_2 \) on any occasion does not change the distribution of the state of the system at the next action, the distribution of the histories of the system under \( S' \) at the time of any action is the same as that under \( S \).

Hence we have

\[
ET_n(S') = ET_n(S) + \varepsilon_n
\]

where \(-2\varepsilon^2 \leq \varepsilon_n \leq 2\varepsilon^2\), \( n = 1, 2, \ldots \); and hence

\[
Ei_n(S') = Ei_n(S) + a_n \varepsilon_n, \quad n = 1, 2, \ldots
\]

where \( a_n \) is the earning rate of the \( n \)th action. Thus

\[
I_n(S') = \frac{\sum_{n=1}^{N} Ei_n(S')}{\sum_{n=1}^{N} ET_n(S')}
\]

\[
= \frac{\sum_{n=1}^{N} (Ei_n(S) + a_n \varepsilon_n)}{\sum_{n=1}^{N} (ET_n(S) + \varepsilon_n)}
\]

\[
= \left( \frac{\sum_{n=1}^{N} Ei_n(S) + \sum_{n=1}^{N} a_n \varepsilon_n}{\sum_{n=1}^{N} ET_n(S) + \sum_{n=1}^{N} \varepsilon_n} \right)
\]
\[
I_N(S^\dagger) - I_N(S) = \frac{\sum_{n=1}^{N} a_n \varepsilon_n - I_N(S) \sum_{n=1}^{N} \varepsilon_n}{\sum_{n=1}^{N} ET_n(S) + \sum_{n=1}^{N} \varepsilon_n} + \varepsilon
\]

where \( \overline{a} = \max_{i,k} |a_{i,k}| \) and \( \varepsilon = \sum_{n=1}^{N} \varepsilon_n / N \). Since \( \lim_{N \to \infty} \inf \left\{ \frac{\sum_{n=1}^{N} ET_n(S)}{N} \right\} \geq \mu_0 > 0 \) and \(-\varepsilon^2 \leq \varepsilon \leq \varepsilon^2\), it follows that, for sufficiently small \( \varepsilon \),

\[
|I(S^\dagger) - I(S)| \leq \varepsilon.
\]

Take \( S_{\varepsilon} = S^\dagger \).

Thus, \( S_{\varepsilon} \) is an \( \varepsilon \)-approximation of \( S \).

**Section 5. Costly actions; stationary policies.**

Consider any stationary policy \( S \). Let \( G_i(\cdot) \) be the hestiation distribution in state \( i \) under \( S \), \( i = 1, 2, \ldots, L \). In view of Lemma 2.3, we can assume \( q(i; G_i) > 0 \) for all \( i \). Let \( X_n(S) \) denote the state of the system after the \( n^{th} \)
transition; \( n = 0, 1, \ldots \). \((X_0(S) \text{ is the initial state of the system.})\) It is easy to see that \( \{X_n(S) : n = 0, 1, \ldots \} \) is a Markov Chain with a finite state space and transition probability matrix \( P = P(S) = (p_{ij}^k) \), depending on the choice of the actions in the various states. In view of one of our assumptions in Section 2, there exists at least one stationary policy \( S \) for which the resulting Markov Chain \( \{X_n(S)\} \) is irreducible. From the Appendix it then follows that, without loss of generality we may assume the Markov Chain \( \{X_n(S)\} \) to be irreducible for every good stationary policy \( S \).

In Section I we have defined \( I(S) \), our criterion of interest for any policy \( S \), by (1.1). Now define

\[
(5.1) \quad I^{(2)}(S) = \lim_{N \to \infty} \inf_{n=1}^{N} \frac{\sum_{n=1}^{N} E_i^{(2)}(S)}{\sum_{n=1}^{N} E_T^{(2)}(S)}
\]

where \( T^{(2)}_n(S) \) is the time from the \((n-1)\)th transition to the \(n\)th transition and \( i^{(2)}_n(S) \) is the income earned during this period. We then have

**Theorem 5.1.** For any stationary policy \( S \), \( I(S) = I^{(2)}(S) \).

**Proof.** Since we shall be considering a particular stationary policy \( S \), we shall drop the index \( k \) from all the parameters. Let \( G_i(\cdot) \) denote the hesitation distribution in state \( i \), \( i = 1, 2, \ldots, L \). Let \( P = (p_{ij}) \) be the transition probability matrix associated with \( \{X_n(S)\} \). We then have, from Theorem 1 of the Appendix,

\[
(5.2) \quad I(S) = \sum_{i=1}^{L} \frac{\pi_i a_i n(i;G_i) - c_i}{\sum_{i=1}^{L} \pi_i n(i;G_i)}
\]

and

\[
(5.3) \quad I^{(2)}(S) = \sum_{i=1}^{L} \frac{\pi_i a_i \theta(i;G_i) - c_i N(i;G_i)}{\sum_{i=1}^{L} \pi_i \theta(i;G_i)}
\]
where \( \eta(i;G_i), \delta(i;G_i) \) and \( N(i;G_i) \) are as defined in Section 3; \( \{ \pi_i \} \) are the stationary probabilities associated with \( P = (p_{ij}) \); \( \{ \pi^* \} \) are the stationary probabilities associated with \( P^* = (p^*_{ij}) \) and

\[
(5.4) \quad p^*_{ij} = p_{ij} q(i;G_i) + \delta_{ij} (1 - q(i;G_i));
\]

\[
\delta_{ij} = \begin{cases} 
0 & \text{if } i \neq j \\
1 & \text{if } i = j.
\end{cases}
\]

\( \{ \pi_i \} \) are given by the system of equations

\[
(5.5) \quad \begin{cases} 
\pi_i \geq 0 & i = 1, 2, \ldots, L \\
\sum_{i=1}^{L} \pi_i = 1 \\
\pi_j = \sum_{i=1}^{L} \pi_i p_{ij} & j = 1, 2, \ldots, L.
\end{cases}
\]

\( \{ \pi^* \} \) are given by the system of equations

\[
(5.6) \quad \begin{cases} 
\pi^*_i \geq 0 \\
\sum_{i=1}^{L} \pi^*_i = 1 \\
\pi^*_j = \sum_{i=1}^{L} \pi_i^* p_{ij} & j = 1, 2, \ldots, L.
\end{cases}
\]

It is easily seen that

\[
(5.7) \quad \pi^*_i = (\pi_i / q(i;G_i)) / \sum_{i=1}^{L} (\pi_i / q(i;G_i)). \quad i = 1, 2, \ldots, L.
\]

Thus
\[ I(S) = \sum_{i=1}^{L} \pi_i (a_i \eta(i;G_i) - c_i) / \sum_{i=1}^{L} \pi_i \eta(i;G_i) \]

\[ = \sum_{i=1}^{L} \pi_i (a_i \theta(i;G_i) - c_i N(i;G_i)) / \sum_{i=1}^{L} \pi_i \theta(i;G_i) \]

\[ = I^{(2)}(S). \]

**Theorem 5.2.** In \( S_0 \) we need only consider policies involving one-point hesitation distributions.

**Proof.** Let \( S \) be any stationary policy and let \( I(S) = I \). Let \( G_i(\cdot) \) be the hesitation distribution in state \( i \) under the policy \( S; i = 1, 2, \ldots, L \). We then have, dropping the index \( k \) from all the parameters,

\[ I(S) = I = \sum_{i=1}^{L} \pi_i (a_i \theta(i;G_i) - c_i N(i;G_i)) / \sum_{i=1}^{L} \pi_i \theta(i;G_i). \]

Let \( x_i, i = 1, 2, \ldots, L \), be such that

\[ \theta(i;x_i)(a_i - I) - c_i N(i;x_i) \]

\[ = \max_{x \in [0, \infty]} \theta(i;x)(a_i - I) - c_i N(i;x). \]

Let \( \hat{S} \) be the policy that differs from \( S \) only in that the hesitation distributions \( G_i(\cdot) \) are replaced by \( G_{x_i}(\cdot), i = 1, 2, \ldots, L \). We then have

\[ I(S') = \sum_{i=1}^{L} \pi_i (a_i \theta(i;G_{x_i}) - c_i N(i;G_{x_i})) / \sum_{i=1}^{L} \pi_i \theta(i;x_i). \]

It can be easily verified that

\[ \sum_{i=1}^{L} \pi_i ((a_i - I)[\theta(i;x_i) - \theta(i;G_i)] - c_i [N(i;x_i) - N(i;G_i)]) \]

\[ = I(S') - I. \]
Now for each $i, \; i = 1, 2, \ldots, L,$

$$(a_i - I)[\theta(i; x_i) - \theta(i; G_i)] - c_i[N(i; x_i) - N(i; G_i)]$$

$$= (a_i - I) \left\{ \theta(i; x_i) - \frac{\int_0^\infty \theta(i; t)F_i(t)dG_i(t)}{\int_0^\infty F_i(t)dG_i(t)} \right\} - c_i \left\{ \frac{1}{F_i(x_i)} - \frac{1}{\int_0^\infty F_i(t)dG_i(t)} \right\}$$

$$\int_0^\infty F_i(t) \left[ (a_i - I)\theta(i; x_i) - c_i N(i; x_i) \right] - \left[ (a_i - I)\theta(i; t) - c_i N(i; t) \right] \; dG_i(t)$$

$$> 0 \quad \text{unless } dG_i(x_i) = 1$$

and \quad $dG_i(t) = 0$ for $t \neq x_i$ \hspace{1cm} (in view of (5.8)).

Hence the theorem.

**Theorem 5.3.** There exists an optimal policy in the class of all stationary policies.

**Proof.** In view of Theorem 5.2, we can restrict our attention to stationary policies involving one-point hesitation distributions only. Let $S$ be any stationary policy involving hesitation distribution $G_{i}(\cdot)$ in state $i$, \linebreak $i = 1, 2, \ldots, L$. Let $\bar{S}$ denote only the choice of actions in various states \linebreak dictated by $S$. Then, dropping the index $k$ from the parameters, we have

$$I(S) = I(\bar{S}; \bar{t}_1, \bar{t}_2, \ldots, \bar{t}_L)$$

$$= \frac{\sum_{i=1}^{L} \left\{ \pi_i (a_i \int_0^{\bar{t}_i} (1 - F_i(x))dx - c_i) / F_i(\bar{t}_i) \right\}}{\sum_{i=1}^{L} \left\{ \pi_i \int_0^{\bar{t}_i} (1 - F_i(x))dx / F_i(\bar{t}_i) \right\}}.$$
Since all the $F(\cdot)$ are assumed to be continuous, for each $\bar{S}$, as a function of $^1t_1,^2t_2,\ldots,^L_t$, $I(\bar{S};^1t_1,^2t_2,\ldots,^L_t)$ is continuous on $(0,\infty) \times (0,\infty) \times \cdots \times (0,\infty)$ and is bounded above. Since, if any $^it_i=0$, $I(S)=\infty$, $I(\bar{S};^1t_1,^2t_2,\ldots,^L_t)$, as a function of $^1t_1,^2t_2,\ldots,^L_t$, attains its maximum. Further, since there is only a finite number of different $\bar{S}$, the theorem follows.

Remark. The assumption of the continuity of the distribution functions is used in the proof of the above theorem. The following example shows that this continuity assumption cannot be dropped.

Example. Consider a system with two states and one action in each state. Let

$$F_1(x) = \begin{cases} 1 - e^{-x} & 0 \leq x < \infty \\ 0 & x \leq 1 \\ 1 - e^{-1} & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

$$F_2(x) = \begin{cases} 1 - e^{-x} & 0 \leq x < 1 \\ 1 - e^{-2} & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

$a_2 > a_1$, $c_1 > c_2$, $p_{12} = 1 = p_{21}$.

We have $\Theta(F_1) = \Theta(F_2) = 1$.

Since there is only one action in each state, any policy $S$ is specified by $(^1t_1,^2t_2)$ if $G_{^1t_1}(\cdot)$ and $G_{^2t_2}(\cdot)$ are the hesitation distributions specified by $S$. Thus

$$I(S) = I(^1t_1,^2t_2).$$

Since $F_1(\cdot)$ is exponential, $I(^1t_1,^2t_2) \leq I(\infty;^2t_2)$ for all $^2t_2$.

We have $I(\infty,\infty) = (a_1+a_2) - (c_1+c_2)/2 = 1$, say.

Now $\Theta(2;^1t) = \int_0^1 (1 - F_2(x))dx/F_2(\cdot)$

$$= \begin{cases} 1 & \text{if } ^1t \leq 1 \text{ or } ^1t \geq 2 \\ 1 + (^1t-1)e^{-1}/(1-e^{-1}) & \text{if } 1 < ^1t < 2. \end{cases}$$
Hence \( I(\infty, t) = \begin{cases} I & \text{if } t \leq 1 \text{ or } t \geq 2 \\ \frac{I + m(a_2(t-1) - c_2)}{1 + m(t-1)} & \text{if } 1 < t < 2 \end{cases} \)

where \( m = e^{-1/2(1-\text{e}^{-1})} \).

Since \( a_2 > a_1 \), \( \frac{d}{dt} I(\infty, t) > 0 \) for \( 1 < t < 2 \).

Thus \( \text{Sup } I(\infty, t) = \lim_{t \to 2^-} I(\infty, t) = \frac{I + m(a_2 - c_2)}{1 + m} > I \)

since \( c_1 > c_2 \).

Hence no optimal stationary policy exists.

Section 6. Costly actions; existence of an optimal policy.

Theorem 6.1. (Derman). Let \( w_{i,k} \) and \( w_{1,k} \) be two sets of expected rewards under action \( k \) in state \( i \), \( k = 1,2,\ldots,K; \ i = 1,2,\ldots,L \). (\( K < \infty \), \( L < \infty \)). Let \( \hat{w}_n \) and \( \hat{w}_n' \) be the respective expected rewards ascribed to \( n^{th} \) action for a fixed policy \( S \). Consider the reward criterion

\[
\psi(S(i)) = \liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{N} w_{n} \hat{w}_{n} \quad i = 1,2,\ldots,L
\]

when the initial state of the system is \( i \). Then there exists a non-randomized stationary policy \( S^* \) such that

\[
\psi(S^*(i)) = \max_S \psi(S(i)) \quad i = 1,2,\ldots,L
\]

where the maximum is taken over all \( S \) that take actions only at the instants of transitions.

Remark. The above theorem is a slight modification of Theorem 3 of Derman [1962]. \( \psi(S(i)) \) may depend on the initial state \( i \) for a general policy \( S \) but \( \psi_S(i) \) is independent of \( i \).
We are now ready to give the

Proof of Theorem 2.1. For every $\varepsilon > 0$, let $S_{\varepsilon} \subseteq S$ denote the class of policies involving hesitation distributions that are randomizations over $G_{\varepsilon}$. By virtue of Theorem 6.1, there exists a non-randomized stationary policy $S^*_\varepsilon$ that is optimal in $S_{\varepsilon}$. Let $S_0$ be the optimal policy in $S_0$. (Theorem 5.3). Note that $S_0$ is independent of $\varepsilon$ and $S^*_\varepsilon$ is in $S_{\varepsilon}$.

Let $I^* = \sup_{S \in S} I(S)$.

Clearly,

(6.1) $I^* \geq I(S_0) > I(S^*_\varepsilon)$ for all $\varepsilon > 0$.

Let $\{S_m : m = 1, 2, \ldots\}$ be a sequence of policies in $S$ such that

(6.2) $I(S_m) \uparrow I^*$.

For a given $\varepsilon > 0$, we can find an $M$ such that for $m > M$, we have

(6.3) $I(S_m) \geq I^* - \varepsilon$.

Let $S^*_m \in S_{\varepsilon}$ be an $\varepsilon$-approximation of $S_m$ (Theorem 4.1). Hence

(6.4) $I(S^*_m) \geq I(S_m) - \varepsilon$.

We thus have

(6.5) $I(S^*_\varepsilon) \geq I(S^*_m) \geq I(S_m) - \varepsilon$.

Now from (6.1), (6.5) and (6.3) we get

(6.6) $I^* \geq I(S_0) \geq I(S^*_\varepsilon) \geq I^* - 2\varepsilon.$
Since $\epsilon$ is arbitrary, from (6.6) we get

\[(6.7) \quad I^* = I(S_0)\].

Take $S^* = S_0$.

Hence the theorem.

Remark. We have seen earlier (Section 5) that if the assumption of continuity of the waiting time distributions is dropped, an optimal stationary policy may not exist, and hence, an optimal policy may not exist. However, part of the inequality (6.6) is still true, i.e.,

\[(6.6') \quad I^* \geq I(S^*_\epsilon) \geq I^* - 2\epsilon\].

Thus, it follows that, even if an optimal policy does not exist, we can find a non-randomized stationary policy that is "almost optimal".

Proof of Corollary 2.2. In view of Theorem 2.1, the optimal policy is a non-randomized stationary policy, say $S^*$, involving only one-point hesitation distributions, say, $G_i(\cdot)$, $i = 1, 2, \ldots, L$. From Theorem 1 of the Appendix and Theorem 5.1, we have

\[(6.8) \quad I^* = I(S^*) = \sum_{i=1}^{L} \pi_i \left(a_i, \theta(i; t_1) - c_i N(i; t_1) \right) \left/ \sum_{i=1}^{L} \pi_i \theta(i; t_1) \right.\].

From Corollary 3.3, it follows that

\[(6.9) \quad I^* = I(S^*) = \sum_{i=1}^{L} \pi_i (a_i, \theta(i) - c_i N(i; t_1)) \left/ \sum_{i=1}^{L} \pi_i \theta(i) \right.\].

Let $S$ be a policy that differs from $S^*$ only in that it involves no hesitation. Then

\[(6.10) \quad I(S) = \sum_{i=1}^{L} \pi_i (a_i, \theta(i) - c_i) \left/ \sum_{i=1}^{L} \pi_i \theta(i) \right.\].
Hence, unless \( N(i; t_i) = 1 \) for all \( i \), the optimality of \( S^* \) will be contradicted.

This completes the proof.

**Remark.** The case when all the waiting time distributions are exponential with mean 1 may be regarded as the continuous analogue of the models considered by Blackwell and Derman.

We have seen above (Corollary 2.2) that when the waiting time distributions are exponential, we can eliminate from consideration policies that involve hesitation. A question that arises is whether the same remains true when the distributions are not necessarily exponential. The following example answers the above question in the negative.

**Example.** Consider a system with two states, with one action in each state and such that

\[
F_1(x) = \begin{cases} 
  x & 0 \leq x \leq 1 \\
  0 & x > 1. 
\end{cases}
\]

\[
F_2(x) = 1 - e^{-x} \quad 0 \leq x < \infty
\]

\( a_1 = 4, \quad a_2 = 2; \quad c_1 = c_2 = 1 \)

\( p_{12} = 1 = p_{21} \).

There is only one stationary policy \( S \) not involving any hesitation and we have

\[
I(S) = \frac{\pi_1 (a_1 \theta(1) - c_1) + \pi_2 (a_2 \theta(2) - c_2)}{\pi_1 \theta(1) + \pi_2 \theta(2)}.
\]

Since \( p_{12} = p_{21} = 1 \), we have \( \pi_1 = \pi_2 = 1/2 \). Also \( \theta(1) = 1/2, \theta(2) = 1 \).

Thus \( I(S) = 4/3 \).
Let $S'$ be the stationary policy that involves the hesitation distribution $G(\cdot)$ in state 1 and no hesitation in state 2.

$$G(\cdot) = \int_0^t (1-x)dx/t = 1 - t/2$$

Hence $I(S') = 43/32 > 4/3$.

**Theorem 6.2.** Let $S$ be an optimal stationary policy with hesitation distributions $G_i(\cdot)$, $i = 1, 2, \ldots, L$, and let $I(S) = I$. Then if $a_i \geq (\leq)I$ and $F_i(\cdot)$ is such that the expected time to transition in state $i$ cannot be increased (decreased) by hesitation, then $t_i = \infty$.

**Proof.** Let $S'$ be a policy that differs from $S$ only in that it involves no hesitation in state $i$. Let $I' = I(S')$. Then

$$I - I' = \frac{\pi_i(0(i) - \theta(i; t_i))(I-a_i) - c_i \pi_i(N(i; t_i)-I)}{\pi_i 0(i) + \sum_{j \neq i} \pi_i \theta(j; t_j)}$$

Now, if $a_i \geq (\leq)I$, $\theta(i; t_i) \leq (\geq) \theta(i)$ and $t_i < \infty$, we shall have $I - I' < 0$, a contradiction. Hence the theorem.

**Remark.** The result of the above theorem also follows from the fact that the hesitation distributions involved in the optimal policy are such that they maximise, for each $i$, $\theta(i; t)(a_i - I) - c_i N(i; t)$.

**Corollary 6.3.** Let $S$ be an optimal stationary policy with hesitation distributions $G_i(\cdot)$, $i = 1, 2, \ldots, L$, and let $I(S) = I$. Then,

a) if $F_i(\cdot)$ is I.F.R. (D.F.R.) with density $f_i(\cdot)$ and $a_j \leq (\geq)I$, $t_j = \infty$;

b) if $F_j(\cdot)$ is exponential, $t_j = \infty$; $j = 1, 2, \ldots, L$. 

(Théorem 5.2)
Proof. The Corollary is an immediate consequence of Theorem 6.2, Lemma 3.2 and Corollary 3.3.

A numerical example.

Consider a machine which at any time can be in one of two states:
state 1 - the working state; state 2 - the failed state. In state 1 we have three actions (3 types of maintenance) and in state 2 we have two actions (2 types of repair service) available to us. Suppose

\[ F_1^1(x) = x/4, \quad 0 \leq x \leq 4; \quad a_1^1 = 5; \quad c_1^1 = 2. \]
\[ F_1^2(x) = 1 - e^{-x/2}, \quad 0 \leq x < \infty; \quad a_1^2 = 5; \quad c_1^2 = 3. \]
\[ F_1^3(x) = x/6, \quad 0 \leq x \leq 6; \quad a_1^3 = 8; \quad c_1^3 = 4. \]
\[ F_2^1(x) = 1 - e^{-x/3}, \quad 0 \leq x < \infty; \quad a_2^1 = -1; \quad c_2^1 = 1. \]
\[ F_2^2(x) = 1 - e^{-x-xe^{-x}}, \quad 0 \leq x < \infty; \quad a_2^2 = -2; \quad c_2^2 = 1. \]
\[ p_{12} = p_{12}^2 = p_{12}^3 = p_{21}^1 = p_{21}^2 = 1. \]

We note that \( \theta(1,1) = 2, \theta(1,2) = 2, \theta(1,3) = 3, \theta(2,1) = 3 \) and \( \theta(2,2) = 2. \)

Let \( S(i,j,x,y) \) denote the policy that takes actions \( i \) and \( j \) in states 1 and 2 and involves the hesitation distributions \( G_x(\cdot) \) and \( G_y(\cdot) \) in states 1 and 2 respectively. We have \( I(S(2,1,\infty,\infty)) \leq I(S(2,1,x,y)) \) for all \( x,y. \)

\[ I(S(2,1,\infty,\infty)) = \frac{(5x2-3) + (-1x3-1)}{2 + 3} = \frac{3}{5}. \]

Hence \( I^* \geq 3/5. \)

Since \( F_2^2(\cdot) \) is I.F.R. and \( a_2^2 < I^* \), and \( F_2^1(\cdot) \) is exponential, the optimal policy will not involve any hesitation in state 2. Hence we shall only consider policies \( S(i,j,x,\infty) \), with \( x=\infty \) when \( i=2 \), which we may denote by \( S(i,j,x). \)
Let \( I(i,j,x) = I(S(i,j,x)) \).
Let \( I(i,j,x^*) = \max_x I(i,j,x) \).
We have

\[
I(1,1,x) = (16 - 8/x - 5x/2)/(7 - x/2) \quad 0 < x \leq 4
\]
\[
I(1,1,4) = 0.80
\]
\[
x^* = (4\sqrt{137} - 8)/19 = 2.0431
\]
\[
I(1,1,x^*) = 1.1670.
\]
\[
I(1,2,x) = (15 - 8/x - 5x/2)/(6 - x/2) \quad 0 < x \leq 4
\]
\[
I(1,2,4) = 0.75
\]
\[
x^* = (4\sqrt{94} - 8)/15 = 2.0521
\]
\[
I(1,2,x^*) = 1.2005
\]
\[
I(2,1,\infty) = 0.60
\]
\[
I(2,2,\infty) = 0.50
\]
\[
I(3,1,x) = (44 - 24/x - 4x)/(9 - x/2) \quad 0 < x \leq 6
\]
\[
I(3,1,6) = 2.6667
\]
\[
x^* = 6(\sqrt{22} - 1)/7 = 3.1632
\]
\[
I(3,1,x^*) = 3.2028
\]
\[
I(3,2,x) = (43 - 24/x - 4x)/(8 - x/2) \quad 0 < x \leq 6
\]
\[
I(3,2,6) = 3.00
\]
\[
x^* = 8(\sqrt{15} - 1)/7 = 3.2834
\]
\[
I(3,2,x^*) = 3.5476.
\]

Hence the optimal policy is \( S(3,2, 3.2834) \) and \( I^* = 3.5476 \). It may be noted that if we do not consider any hesitation, the optimal policy is \( S(3,2) \) and \( I^* = 3.00 \).
Section 7. Costless actions.

In this section we consider the stochastic system introduced in Section 2 when the actions are costless. We restrict our attention to the class of policies $S^\alpha$, defined in Section 1, for a fixed $\alpha > 0$. Of course, we are interested in $\alpha$ sufficiently close to zero.

As a consequence of this restriction, we have

\[(7.1) \quad ET_n(S) \geq \eta_*(\alpha)\]

for every $S \in S^\alpha$ and for all $n$, where

\[(7.2) \quad \eta_*(\alpha) = \min_{l,k} \eta(l,k;\alpha).\]

Equation (7.1) implies that $\liminf_{N \to \infty} \left\{ \sum_{n=1}^{N} \frac{ET_n(S)}{N} \right\} \geq \eta_*(\alpha)$ for every $S$ in $S^\alpha$.

By replacing (4.1) by

\[(7.3) \quad \eta(i,k; t^k_r(r)) = \eta(i,k;\alpha) + \frac{r}{N} (0(i,k) - \eta(i,k;\alpha))\]

\[r = 0,1,\ldots,N; \quad k = 1,2,\ldots,K; \quad i = 1,2,\ldots,L.\]

and the $\mu_0$ by $\eta_*(\alpha)$, it can be seen that Theorem 4.1 is applicable even when the actions are costless and that $S^\alpha$ is in the class $S^\alpha$. Next if we replace (5.8) by

\[(7.4) \quad \theta(i;x_i)(a_i-I) = \max_{x \in \alpha} \theta(i;x)(a_i-I),\]

Theorem 5.2 can be seen to hold in the present case also. Theorem 5.3, with $\alpha \leq t_i \leq \infty$, $i = 1,2,\ldots,L$, is also valid and we thus have

**Theorem 7.1.** For every $\alpha > 0$, there exists a non-randomized stationary policy $S^*_\alpha$ that is optimal in $S^\alpha$. Further, $S^*_\alpha$ involves only one-point hesitation distributions.
Proof. The proof is very similar to the proof of Theorem 2.1 given in Section 6.

In general, the optimal policy depends on \( \alpha \). An interesting question is under what conditions the optimal policy will be independent of \( \alpha \), for \( \alpha \) sufficiently close to zero. We indicate below a few situations where the optimal policy is independent of \( \alpha \).

1) If all the waiting time distributions are exponential, then hesitation can be eliminated and as such the optimal policy will be independent of \( \alpha \).

2) If \( \max_k a_i^k = a_i^* = a_i^* \) for all \( i \), then the optimal policy will be independent of \( \alpha \).

3) From (7.4) we see that the hesitation distributions \( G_\alpha^*(\cdot) \) involved in 
   \[ S_\alpha^* \text{ maximise } \theta(i; x)(a_i - I_i^*) \text{ for } x \in [\alpha, \infty), \]
   where \( I_i^* = I(S_\alpha^*) \); i.e. the \( x_i \)
either maximise or minimise \( \theta(i; x) \) depending upon whether \( a_i > I_i^* \) or \( I_i^* \);
   \( i = 1, 2, \ldots, L. \)

   Thus if the waiting time distributions are such that \( \theta(i, k; x) \), for all \( i \)
and \( k \), attains neither its maximum nor minimum at the origin, then for \( \alpha \)
sufficiently close to zero, the optimal policy is independent of \( \alpha \).

4) For some \( \alpha \) sufficiently close to zero, let \( S_\alpha^* \) be the stationary policy
   that is optimal in \( S_\alpha^* \) and let \( I_i^* = I(S_i^*) \). Suppose the actions involved
   in the policy \( S_\alpha^* \) are such that for \( a_i < I_i^* \), \( F_i(\cdot) \) is I.F.R. and for
   \( a_i > I_i^* \), \( F_i(\cdot) \) is D.F.R. Then \( S_\alpha^* \) involves no hesitation and, as such,
   \( I_i^* \) and \( S_\alpha^* \) are independent of \( \alpha \).

5) Suppose for each combination of \( i \) and \( k \), \( \theta(i, k; x) \), as a function of \( x \),
either first increases and then decreases or first decreases and then
increases. Suppose \( G_i(\cdot) \), \( i = 1, 2, \ldots, L \), are the hesitation distribu-
tions involved in \( S_\alpha^* \). If, for some \( \alpha \) sufficiently close to zero,
\[ \min_{i=1}^L t_i^\alpha > \alpha, \]
then \( S_\alpha^* \) is independent of \( \alpha \).
We shall indicate the truth of the statement (5) by considering one possible situation. Suppose $\theta(i;x)$, corresponding to the action prescribed by $S^*_\alpha$ in state $i$, first increases and then decreases (to $\theta(i)$ as $x \to \infty$). From Figure (1), it is clear that if $t_1(\alpha_0) > \alpha_0$, then $t_1(\alpha_0)$ maximises $\theta(i;x)$ and hence $t_1(\alpha)$, for any $\alpha < \alpha_0$, will be $t_1(\alpha_0)$.

![Figure (1)](image)

Now consider a stationary policy $S$ such that $a_1$ is the highest earning rate and $F_i(\cdot)$ is I.F.R. with density $f_i(\cdot)$. Then, if $S=\alpha$, the hesitation distribution involved in state 1 (from (7.4) and Lemma 3.2) will be $G_\alpha(\cdot)$. Now if $\alpha = 0$, this policy would involve instantaneous hesitation repeatedly in state 1 and if the system starts in state 1, then $E_{1n}(S) = 0$ and $E_{1n}(S) = 0$ for $n = 1, 2, \ldots$. As such, for this policy, our criterion of interest, $I(S)$, would not be defined at all. Hence the necessity to restrict our attention to the class of policies $S^\alpha$, for $\alpha > 0$.

It may be observed that continuous instantaneous hesitation is equivalent to taking an action that has an exponential waiting-time distribution.
with mean $1/F'(0)$. So one could introduce this artificial action. However, the problem would still remain if $F'(0) = 0$.

Derman's [1962] linear programming method or Howard's [1963] policy improvement method can be used to obtain the optimal stationary policy. We note that for a given $\alpha > 0$, the hesitation distribution $G_i(\cdot)$ used by $S^*_\alpha$ in state $i$ is such that it either maximises or minimises $\theta(i;x)$, $i = 1, 2, \ldots, L$. Now consider a new problem with the same state space as our original problem but corresponding to every action $k$ in state $i$ of the original problem, the new problem has two actions $k'$ and $k''$ such that

\[ a_{ik} = a_{i k'} = a_{i''} \quad \text{for all } i \text{ and } k, \]
\[ p_{ij} = p_{ij'} = p_{ij''} \quad \text{for all } i, j \text{ and } k, \]

\[ \theta(i,k') = \max_{x \in [\alpha, \infty]} \theta(i,k;x) \quad \text{for all } i \text{ and } k, \]

and

\[ \theta(i,k'') = \min_{x \in [\alpha, \infty]} \theta(i,k;x) \quad \text{for all } i \text{ and } k. \]

If we do not allow any hesitation for the new problem, then the optimal policy for the new problem is the same as the optimal policy for our original problem. The optimal policy for the new problem can be obtained by using either the Derman linear programming technique or the Howard policy improvement method.

Section 8. Costly observations.

So far it was implicitly assumed that the system could be continuously observed at no cost. We shall now study the same problems as before when there is a positive cost $c_0$ for each observation. We shall use the same model as described in Section 2 but we shall allow the $c_{ik}$ to be zero.
Instead of the hesitation distributions of the previous sections, we shall now have the observation distributions—the distributions of the times to the next observation. We shall restrict our attention to the class of policies \( R^A \) such that for any policy \( S \) in \( R^A \) the observations form a sequence and the observation distributions prescribed by \( S \) are distributions on \([0, \Delta]\), where \( \Delta < \infty \). Any policy \( S \) in \( R^A \) specifies, after each observation, the action to be taken and the time to the next observation, possibly randomized and depending on the past history of the system.

Another restrictive assumption which we shall now make is that if, after an action has been taken after an observation, the system makes a transition before the next observation, then the system is temporarily absorbed in the new state and earns nothing until the next observation and action.

For any policy \( S \), the criterion of interest is \( I(S) \) defined by (1.2). As before, we are interested in the existence and nature of policies \( S^* \) such that

\[
I(S^*) = \sup_{S \in R^A} I(S).
\]

**Lemma 8.1.** Any policy \( S \) for which

\[
\liminf_{N \to \infty} \left\{ \frac{\sum_{n=1}^{N} EY_n(S)}{N} \right\} < \mu_0^*,
\]

where

\[
\mu_0^* = \begin{cases} 
\Delta (c_0 + c^*_x - a^*_x n^*_x(\Delta))/(c_0 + c^*_x - a^*_x n^*_x(\Delta)) & \text{if } a^*_x < 0, \\
(c_0 + c^*_x)/(a^*_x + \frac{c^*_0 + c^*_x}{\Delta}) & \text{if } a^*_x \geq 0, \\
(c_0 + c^*_x)/(a^*_x + \frac{c^*_0 - a^*_x n^*_x(\Delta)}{\Delta}) & \text{if } a^*_x < 0 \leq a^*_x,
\end{cases}
\]

is not optimal. In the above equation, \( a^*_x, a^*_x, c^*_x, c^*_x \) are as defined in Section 2 and
\[ n^*(\Delta) = \max_{i,k} \eta(i,k;\Delta) \]

(8.2) \[ n_*(\Delta) = \min_{i,k} \eta(i,k;\Delta). \]

Proof. Let \( \hat{S} \) be a policy that observes the system after every \( \Delta \) units of time. Then

\[ I_N(\hat{S}) = \sum_{n=1}^{N} \frac{E_i(n,\hat{S})}{N\Delta} \]

\[ = \begin{cases} \frac{-(c_0+c_*)}{\Delta} & \text{if } a_* \geq 0 \\ \frac{(a_*n_*(\Delta)-c_*-c_0)}{\Delta} & \text{if } a_* < 0. \end{cases} \]

Thus

(8.3) \[ I(S) \leq \begin{cases} \frac{-(c_0+c_*)}{\Delta} & \text{if } a_* \geq 0 \\ \frac{(a_*n_*(\Delta)-c_*-c_0)}{\Delta} & \text{if } a_* < 0. \end{cases} \]

Now, for any policy \( S \),

\[ I_N(S) = \sum_{n=1}^{N} \frac{E_i(n,S)}{\sum_{n=1}^{N} \sum E_y(n,S)} \]

\[ = \begin{cases} \frac{a_* - (c_0+c_*)}{\left(\sum_{n=1}^{N} E_y(n,S)/N\right)} & \text{if } a_* \geq 0, \\ \frac{(a_*n_*(\Delta) - c_0 - c_*)}{\left(\sum_{n=1}^{N} E_y(n,S)/N\right)} & \text{if } a_* < 0. \end{cases} \]

Hence

(8.4) \[ I(S) \leq \begin{cases} \frac{a_* - (c_0+c_*)}{\lim \inf_{N \to \infty} \left\{ \sum_{n=1}^{N} E_y(n,S)/N \right\}} & \text{if } a_* \geq 0, \\ \frac{(a_*n_*(\Delta) - c_0 - c_*)}{\lim \inf_{N \to \infty} \left\{ \sum_{n=1}^{N} E_y(n,S)/N \right\}} & \text{if } a_* < 0. \end{cases} \]

Now, if \( \lim \inf_{N \to \infty} \left\{ \sum_{n=1}^{N} E_y(n,S)/N \right\} < v_0 \), from (8.4), (8.3) and (8.1) we get
\[ I(S) < I(S'). \]

Hence the Lemma.

In view of the above lemma, we need only consider policies \( S \) for which
\[
\lim_{N \to \infty} \inf \left\{ \frac{\sum_{n=1}^{N} E Y_n(S)}{N} \right\} \geq \nu_0^*.
\]

Suppose under some policy \( S \), at some stage we are in state \( i \) and action \( k \) has been taken and the next observation is planned after time \( Y \) whose distribution is \( G(\cdot) \). Then

\[ (8.5) \quad p_{ij}^k = \text{probability that the system is in state } j \text{ at the time of the next observation} \]
\[ = p_{ij}^k P[T_i^k \leq Y] + \delta_{ij}^k P[T_i^k > Y] \]
\[ = p_{ij}^k q(i,k;G) + \delta_{ij}^k (1 - q(i,k;G)), \]

where
\[ \delta_{ij}^k = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \]

The expected income before the next observation will be
\[ = a_i^k \min \{ T_i^k, Y \} - c_i^k - c_0 \]
\[ = a_i^k \eta(i,k;G) - c_i^k - c_0. \]

Let \( \varepsilon > 0 \) be given. Let \( N_\varepsilon \) be a positive integer such that \((N_\varepsilon - 1)\varepsilon < \Delta = N_\varepsilon \varepsilon \). Take

\[ (8.6) \begin{cases} t_i = i\varepsilon & \text{for } i = 0, 1, \ldots, N_\varepsilon - 1 \\
 t_{N_\varepsilon} = \Delta. \end{cases} \]
Then from Theorem 3.4 it follows that there exists a discrete distribution
\( G_\varepsilon(\cdot) \), which has its mass only at the points \( t_i, i = 0, 1, \ldots, N_\varepsilon \), such that

\[
(8.7) \quad q(i, k; G) = \int_0^\Delta F_i^k(t) dG(t) = \int_0^\Delta F_i^k(t) dG_\varepsilon(t) = q(i, k; G_\varepsilon)
\]

(8.8) \( |\eta(i, k; G) - \eta(i, k; G_\varepsilon)| \leq 2\varepsilon \) and

\[
(8.9) \quad \left| \int_0^\Delta tdG(t) - \int_0^\Delta tdG_\varepsilon(t) \right| \leq 2\varepsilon.
\]

Let \( G^{\varepsilon}_2 \) now denote the finite class of hesitation distributions \( \{G_{\varepsilon}^+_i(\cdot); \ i = 0, 1, \ldots, N_\varepsilon \} \). Hence we have

**Theorem 8.2.** For any policy \( S \) and for any sufficiently small \( \varepsilon > 0 \), there exists a policy \( S^{\varepsilon}_\varepsilon \), an \( \varepsilon \)-approximation of \( S \), such that \( S^{\varepsilon}_\varepsilon \) involves observation distributions that are randomizations over the class \( G^{\varepsilon}_2 \).

The proof is essentially the same as the proof of Theorem 4.1, with the observation distributions replacing the hesitation distributions.

Now consider any stationary policy \( S \). Let \( X_n(S) \) denote the state of the system at the time of the \( n \)th observation, \( n = 1, 2, \ldots \). It is easily seen that \( \{X_n(S); n = 1, 2, \ldots \} \) is a Markov Chain with transition probability matrix \( P = (p_{ij}^k) \) where the \( p_{ij}^k \) depend on the choice of the action \( k \) in state \( i \) and the observation distribution \( G_i(\cdot) \) in state \( i \) and are given by expressions like (8.5). As before, we can restrict our attention to stationary policies \( S \) for which the Markov Chain \( \{X_n(S); n = 1, 2, \ldots \} \) is irreducible. Then, from Theorem 1 of the Appendix, we get

\[
(8.10) \quad I(S) = \frac{\sum_{i=1}^L \{\pi_i a_i^k \eta(i, k; G_i) - c_i^k - c_0)/q(i, k; G_i)\}}{\sum_{i=1}^L \{\pi_i \theta(G_i)/q(i, k; G_i)\}}
\]
where \( \{\pi_i\} \) are the stationary probabilities associated with \( P = (p_{ij}^K) \).

**Theorem 8.3.** In the class of stationary policies we need only consider policies involving one-point observation distributions.

The proof is similar to the proof of Theorem 5.2. However, the \( x_i \), introduced in (5.8) are now defined by

\[
(8.11) \quad V_i(x_i) = \max_{0 \leq x \leq \Delta} V_i(x), \quad i = 1, 2, \ldots, L,
\]

where

\[
(8.12) \quad V_i(x) = (a_i \eta(i;x) - c_i - c_0 - l_0)/F_i(x), \quad i = 1, 2, \ldots, L.
\]

**Theorem 8.4.** There exists an optimal policy in the class of stationary policies.

The proof is similar to the proof of Theorem 5.3.

**Theorem 8.5.** There exists a non-randomized stationary policy \( S^* \) that is optimal in \( R^\Delta \). Further, \( S^* \) involves only one-point observation distributions.

The proof is similar to the proof of Theorem 2.1 given in Section 6.

The optimal policy \( S^* \) in \( R^\Delta \) may possibly depend upon \( \Delta \). To indicate this, let us denote the optimal policy in \( R^\Delta \) by \( S^*_\Delta \) and let \( I(S^*_\Delta) = I^*_\Delta \). We shall now show that for sufficiently large \( \Delta \), \( S^*_\Delta \) is independent of \( \Delta \).

Consider a sequence of stationary policies \( \{S_n\} \) such that \( S_n \) is in \( R^\Delta_n \) and in some state \( i \) involves the use of the observation distribution \( G^\Delta_n(\cdot) \), \( n = 1, 2, \ldots \). From (8.10) we see that

\[
I(S_n) \to 0 \text{ as } \Delta_n \to \infty.
\]

In view of this, we shall assume that, for sufficiently large \( \Delta \), \( I^*_\Delta = I(S^*_\Delta) > 0 \).

Next observe that if \( G^\Delta_x(\cdot) \) are the observation distributions involved in \( S^*_\Delta \), then the \( x_i \) maximize
\[ V_i(x) = (a_i n(i; x) - c_i - c_0 - I^*_{\Delta} x) / F_i(x) \quad i = 1, 2, \ldots, L. \]

To indicate the possible dependence of \( x_i \) on \( \Delta \), we shall denote them by \( x_i(\Delta) \). For some sufficiently large \( \Delta_0 \), let

\[ A_i = \{ x_i(\Delta) : \Delta \geq \Delta_0 \}, \]

\[ x_i^* = \text{Sup} \{ x : x \in A_i \}, \quad i = 1, 2, \ldots, L, \]

and

\[ x^* = \max_i x_i^*. \]

Since, for any \( \Delta < \infty \), \( I_{\Delta}^* \leq x^* \) and \( V_i(x) \to -\infty \) as \( x \to \infty \), it follows that \( x_i^* < \infty \), \( i = 1, 2, \ldots, L \). Let \( \Delta_2 > \Delta_1 > \max (x_i^*, \Delta_0) \).

Since \( R_{\Delta_1} \subset R_{\Delta_2} \), we have \( I_{\Delta_1}^* \leq I_{\Delta_2}^* \).

Note that the observation distributions involved in \( S_{\Delta_2}^* \) are distributions on \([0, x^*]\) and hence \( S_{\Delta_2}^* \Delta_1 \) is in \( R_1 \) also. Hence we have \( I_{\Delta_2}^* \leq I_{\Delta_1}^* \), so that

\[ I_{\Delta_2}^* = I_{\Delta_1}^*. \]

Thus, for sufficiently large \( \Delta \), \( I_{\Delta}^* \) is independent of \( \Delta \) and so is \( S_{\Delta}^* \). Hence the restriction that the observation distributions are distributions on \([0, \Delta]\) is not a serious restriction. However, if we do not assume \( \Delta < \infty \), then for a policy \( \mathcal{S} \) for which \( Y_n(\mathcal{S}) = \infty \) with probability 1 for some \( n = n_0 \), say, \( I_{\mathcal{S}}(\mathcal{S}) \) will not be defined for \( n > n_0 \) and hence, our criterion of interest, \( I(\mathcal{S}) \), will not be defined.

Section 9. Monotonicity properties of the optimal policies and the equivalence of different criteria.

In this section we give two theorems which hold in the three cases considered—the costly actions case, the costless actions case and the costly
observations case. In the costly actions case we consider the class of policies \( S \); in the costless actions case, the class \( S^a \) and in the costly observations case, the class \( R^A \).

**Definition.** A vector \( U = (u_1, u_2, \ldots, u_m) \) is said to be greater than (\( \geq \)) a vector \( V = (v_1, v_2, \ldots, v_m) \) if \( u_i \geq v_i, \quad i = 1, 2, \ldots, m \).

Let \( C = (c_0, c_1, c_1^1, c_2^2, \ldots, c_1^K, \ldots, c_L^K) \) denote the cost vector and \( a = (a_1, a_1, \ldots, a_1^K, \ldots, a_L^K) \) denote the earning rate vector. Let \( S^*_C \) denote the optimal stationary policy when \( C \) is the cost vector and, for any \( S \), let \( I_C(S) \) denote \( I(S) \), when \( C \) is the cost vector. The subscript \( a \) is used analogously.

**Theorem 9.1.**

a) \( I_C^*(S^*_C) \) is a decreasing function of \( C \).

b) \( I^*_a(S^*_a) \) is an increasing function of \( a \).

The Theorem follows from the fact that for a given stationary policy \( S \), \( I(S) \) is a decreasing function of \( C \) and an increasing function of \( a \), and from optimality considerations.

In Theorem 5.1, we showed, in the costly actions case, that for any stationary policy \( S \), \( I(S) = I^{(2)}(S) \). Now denote \( I(S) \) by \( I^{(1)}(S) \) and define

\[
I^{(3)}(S) = \lim \inf_{N \to \infty} \frac{\sum_{n=1}^{N} i_n(S)}{\sum_{n=1}^{N} T_n(S)}
\]

\[
I^{(4)}(S) = \lim \inf_{N \to \infty} \frac{\sum_{n=1}^{N} i_n^{(2)}(S)}{\sum_{n=1}^{N} T^{(2)}(S)}
\]

\[
I^{(5)}(S) = \lim \inf_{T \to \infty} \frac{E I^T(S)}{T}
\]

\[
I^{(6)}(S) = \lim \inf_{T \to \infty} \frac{I^T(S)}{T}
\]
where $I^T(S)$ is the income earned up to time $T$ under the policy $S$. Note that $I^{(3)}$, $I^{(4)}$ and $I^{(6)}$ are random variables.

Note that for the costly observations case, the transitions cannot be observed. For this case, we shall only consider the criteria $I^{(1)}$, $I^{(3)}$, $I^{(5)}$, and $I^{(6)}$ and whenever we say transitions, we shall take it to mean actions.

We then have:

**Theorem 9.2.** If $S$ in any stationary policy such that the associated Markov Chain $\{X_n(S)\}$ is irreducible, then with probability 1,

$$I^{(1)}(S) = I^{(2)}(S) = I^{(3)}(S) = I^{(4)}(S) = I^{(5)}(S) = I^{(6)}(S)$$

for all the three cases under consideration.

**Proof.** We have already shown (Theorem 5.1) that for the costly actions case, $I^{(1)}(S) = I^{(2)}(S)$. The same proof holds for the costless actions case.

Hence

$$I^{(1)}(S) = I^{(2)}(S).$$

From Theorem 1 of the Appendix, it follows that

$$\begin{cases} I^{(1)}(S) = I^{(3)}(S) \text{ with probability 1} \\ I^{(2)}(S) = I^{(4)}(S) \text{ with probability 1}. \end{cases}$$

Now consider $I^T(S)$. Suppose the system starts in state $i$. Considering in terms of transitions, let us define a cycle as the period between successive returns to state $i$. Since $S$ is stationary and the associated Markov Chain is irreducible, the successive cycles are independently and identically distributed and with probability one, the number of transitions in a cycle will be finite. Let $v_j(u_j)$ denote income (length) of each cycle.
when the initial state of the system is $i, i = 1, 2, \ldots, L$. Let $N(T)$ denote the number of complete cycles in time $T$; $V_n(i)$, the income earned in the $n^{th}$ cycle when the initial state of the system is $i$, and $W_n(i)$, the income earned during the $(n+1)^{th}$ incomplete cycle starting in state $i$, $i = 1, 2, \ldots, L$. We then have

$$I^{(5)}(S) = \lim_{T \to \infty} \inf \left\{ \frac{\sum_{n=1}^{N(T)} V_n(i) + W_n(i)}{T} \right\}$$

$$I^{(6)}(S) = \lim_{T \to \infty} \inf \left\{ \frac{\sum_{n=1}^{N(T)} V_n(i) + W_n(i)}{T} \right\}.$$  

(9.3)

Note that the length of any transition is finite with probability one and so is the number of actions to a transition. Hence the income earned in any cycle is finite with probability one. Hence

$$\lim_{T \to \infty} \frac{W_n(i)}{N(T)} = 0 \quad \text{with probability one, for all } i.$$  

Thus

$$I^{(5)}(S) = \lim_{T \to \infty} \inf \left\{ \frac{\sum_{n=1}^{N(T)} V_n(i)}{T} \right\}$$

$$I^{(6)}(S) = \lim_{T \to \infty} \inf \left\{ \frac{\sum_{n=1}^{N(T)} V_n(i)}{T} \right\}.$$  

(9.4)

From John and Miller [1963], it then follows that

$$I^{(5)}(S) = u_i / \mu_i$$

$$I^{(6)}(S) = u_i / \mu_i \quad \text{with probability one.}$$  

(9.5)

We have

$$\mu_i = \theta(i; G_i) + \sum_{j \neq i} \pi_j \theta(j; G_j) / \pi_i \quad \text{for the costly and costless actions cases,}$$

$$= \theta(G_i) + \sum_{j \neq i} \{ \pi_j \theta(G_j) / q(j; G_j) \} / (\pi_i / q(i; G_i)) \quad \text{for the costly observations case.}$$
\[ u_i = \{a_i \theta(i; G_i) - c_i N(i; G_i)\} + \sum_{j \neq i} \pi_j \{a_j \theta(j; G_j) - c_j N(j; G_j)\}/\pi_i \]

for the costly actions case,

\[ = a_i \theta(i; G_i) + \sum_{j \neq i} \pi_j a_j \theta(j; G_j)/\pi_i \]

for the costless actions case,

\[ = (a_i n(i; G_i) - c_i - c_0) + \sum_{j \neq i} \{\pi_j (a_j n(j; G_j) - c_j - c_0)/q(j; G_j)\} / (\pi_i/q(i; G_i)) \]

for the costly observations case.

\((G_i(\cdot), i = 1, 2, \ldots, L)\) are the hesitation (observation) distributions involved in the policy \(S\).

Thus, in all the three cases,

\[(9.6) \quad u_i/\mu_i = I^{(1)}(S) \quad \text{for all } i.\]

The theorem follows from (9.1), (9.2), (9.5) and (9.6).

We have proved the above theorem for stationary policies \(S\). It is conjectured that the theorem holds for non-stationary policies also, whenever all the criteria are well defined. Thus, if for some policy \(S\), some criteria are not defined, we could reasonably use some other criterion that is well defined.
Appendix.

Let \( \{X_t, \ t = 0,1, \ldots\} \) be a Markov Chain with a finite state space \( I \) and transition probability matrix \( P = (p_{ij}) \). We then have

**Theorem A.1.** (K. L. Chung). If the Markov Chain \( \{X_t\} \) is irreducible and \( f(\cdot) \) is a function defined on the states, then

\[
(A.1) \quad \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} f(X_t) = \sum_{j \in I} \pi_j f(j) = M \quad \text{(say),}
\]

where \( \{\pi_j\} \) are the stationary probabilities associated with \( P \) i.e. the \( \{\pi_j\} \) are given by

\[
(A.2) \quad \begin{cases} 
\pi_j \geq 0 & j \in I \\
\sum_{j \in I} \pi_j = 1 \\
\pi_j = \sum_{i \in I} \pi_i p_{ij} & j \in I.
\end{cases}
\]

Equation (A.1) implies that

\[
(A.3) \quad \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} Ef(X_t) = \sum_{j \in I} \pi_j f(j).
\]

Now consider a Markov Chain \( \{X_t\} \) which is not irreducible but \( A \subseteq I \) is a class of positive states and the rest of the states are transient. Then we can write

\[
(A.4) \quad \frac{1}{T} \sum_{t=0}^{T} f(X_t) = \frac{1}{T} \sum_{t=0}^{\tau-1} f(X_t) + \frac{T-\tau}{T} \left( \frac{1}{T-\tau} \sum_{t=\tau}^{T} f(X_t) \right)
\]

where \( \tau \) is the first time the Markov Chain enters the class \( A \). If \( \tau > T \), we take the first sum on the right side of the above equation to be from 0 to \( T \) and the second term to be zero.

Now for a given \( \tau \), we have for Theorem A.1,
\[ P \lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau}^{T} f(X_t) = \sum_{j \in A} \pi_j^* f(j) = M_A \text{ (say), where } \{\pi_j^*\} \text{ are the stationary probabilities associated with the reduced Markov Chain with state space } A. \]

Note that \( M_A \) is independent of \( \tau \).

Further, since \( \tau \) is finite with probability one,

\[ P \lim_{T \to \infty} \frac{T-\tau}{T} = 1 \quad \text{and} \quad P \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{\tau-1} f(X_t) = 0. \]

Hence we have

(A.5) \[ P \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} f(X_t) = \sum_{j \in A} \pi_j^* f(j) = M_A. \]

Now consider a Markov Chain \( \{X_t\} \) such that its finite state space \( I \) consists of two disjoint positive subclasses \( A \) and \( B \) and the rest of the states are transient. Then (A.4) is still valid provided \( \tau \) is now defined as the first time the Markov Chain is absorbed either in the class \( A \) or class \( B \). Since \( \tau \) is finite with probability one, we have

\[ P \lim_{T \to \infty} \frac{T-\tau}{T} = 1 \quad \text{and} \quad P \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{\tau-1} f(X_t) = 0. \]

Now, for a given \( \tau \) and given that the Markov Chain is absorbed in the class \( A \) at time \( \tau \), we have

(A.6) \[ P \lim_{T \to \infty} \frac{1}{T-\tau} \sum_{t=\tau}^{T} f(X_t) = \sum_{j \in A} \pi_j^* f(j) = M_A. \]

Similarly, for a given \( \tau \) and given that the Markov Chain is absorbed in class \( B \) at time \( \tau \), we have

(A.7) \[ P \lim_{T \to \infty} \frac{1}{T-\tau} \sum_{t=\tau}^{T} f(X_t) = \sum_{j \in B} \pi_j^* f(j) = M_B \text{ (say)} \]

where \( \{\pi_j^*\} \) are the stationary probabilities associated with the reduced Markov Chain with state space \( B \).
Let $p_A$ ($p_B$) be the probability that the Markov Chain is absorbed in class A (B). Then, we have, from (A.6) and (A.7)

(A.8) \[ \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} f(X_t) = M_A \quad \text{with probability } p_A \]

\[ = M_B \quad \text{with probability } p_B. \]

Now, for our problem, it has been remarked earlier (Section 5) that with every stationary policy $S$, there is an associated Markov Chain $\{X_n(S)\}$. We have defined $I^{(2)}(S)$ as

\[ I^{(2)}(S) = \lim_{N \to \infty} \inf_{n=1}^{N} \frac{\sum_{i} E_i^{(2)}(S)}{\sum_{n=1}^{N} E_k^{(2)}(S)} \]

which can be written as

\[ I^{(2)}(S) = \lim_{N \to \infty} \inf_{n=1}^{N} \frac{\sum_{i} E_i f(X_n(S))/N}{\sum_{n=1}^{N} E_i g(X_n(S))/N} \]

(A.9) \[ I^{(2)}(S) = \lim_{N \to \infty} \inf_{n=1}^{N} \frac{\sum_{i} E_i f(X_n(S))/N}{\sum_{n=1}^{N} E_i g(X_n(S))/N} \]

If $\{X_n(S)\}$ is an irreducible Markov Chain, then using (A.3) in the numerator and denominator of (A.9) we get

(A.10) \[ I^{(2)}(S) = \frac{\sum_{j \in I} \pi_j f(j)}/\sum_{j \in I} \pi_j g(j) = M^* \quad \text{(say)}. \]

If $\{X_n(S)\}$ is a Markov Chain such that $A \subset I$ is a class of positive states and the rest of the states are transient, then, as in (A.5), we get

(A.11) \[ I^{(2)}(S) = \frac{\sum_{j \in A} \pi_j^* f(j)}/\sum_{j \in A} \pi_j^* g(j) = M^*_j \quad \text{(say)}. \]

Next if a stationary policy $S$ is such that the Markov Chain $\{X_n(S)\}$ has two disjoint classes $A$ and $B$ of positive states and the rest of the states are disjoint, we have as in (A.8),

(A.12) \[ I^{(2)}(S) = p_A M^*_A + p_B M^*_B \quad \text{(where } M^*_B \text{ is analogous to } M_B \text{ in (A.7))}. \]
Hence \( I^{(2)}(S) \leq \max (M_A^*, M_B^*) \).

Suppose \( M_A^* > M_B^* \).

In view of one of our assumptions (Section 2), there exists at least one stationary policy \( S_0 \) such that \( \{X_n(S_0)\} \) is an irreducible Markov Chain. Let \( S' \) be a policy that is identical with \( S_0 \) until the Markov Chain enters the class A and then \( S' \) is identical with \( S \). For this policy \( S' \), we have, as in (A.10),
\[
I^{(2)}(S') = M_A^* > I^{(2)}(S).
\]
Since \( I^{(1)}(S) = I^{(2)}(S) \) for any stationary policy \( S \), we have \( I^{(1)}(S) < I^{(1)}(S') \).

Thus, any stationary policy \( S \) which results in the associated Markov Chain having more than one positive class need not be considered. Hence, without any loss of generality, we may restrict our attention to stationary policies such that the associated Markov Chains are irreducible.

It may be remarked here that the assumption (Section 2) which implies that there is at least one stationary policy \( S \) with the associated Markov Chain \( \{X_n(S)\} \) being irreducible, is not a very restrictive assumption. If it is not satisfied, then the Markov Chain associated with every stationary policy will have multiple chains and our results can be easily adapted to take care of this situation. In particular, \( I(S) \), for any stationary policy \( S \), might depend on the initial state of the system.
References


This paper is concerned with optimal policies for controlling a stochastic system with a finite state space and a finite action space when 1) the actions are costly, 2) the actions are costless, and 3) the observations are costly. The transitions of the system occur at random time intervals and we allow hesitation—with positive probability actions may be taken between transitions. A form of the long run average income is used as a criterion for comparing policies. It is shown that in each of the three cases considered, there exists a non-randomized stationary policy that is optimal.