To My Mother

MARKOV DECISION PROCESSES WITH RESTRICTED OBSERVATIONS

A Master's Thesis
Presented by
Zeynep Müge Avşar

to
the Graduate School of Natural and Applied Sciences
of Middle East Technical University
in Partial Fulfillment for the Degree of

MASTER OF SCIENCE

in

INDUSTRIAL ENGINEERING

MIDDLE EAST TECHNICAL UNIVERSITY
ANKARA
September, 1992

Approval of the Graduate School of Natural and Applied Sciences.

for Prof. Dr. Alpay Ankara
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Çağlar Güven
Chairman of the Department

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in Industrial Engineering.

Assist. Prof. Dr. Yasemin Serin
Supervisor

Examining Committee in Charge:

Assoc. Prof. Dr. Nesim Erkip
Assoc. Prof. Dr. Çağlar Güven
Assist. Prof. Dr. Yasemin Serin

ABSTRACT

MARKOV DECISION PROCESSES
WITH RESTRICTED OBSERVATIONS

AVŞAR, Zeynep Müge
M. S. in Industrial Engineering
Supervisor : Assist. Prof. Dr. Yasemin Serin
September, 1992, 88 pages

In this study, Markov Decision Processes, are analyzed under unobservability constraints and algorithms are developed to find the optimal policies with respect to the objective of minimizing the expected total discounted cost over finite planning horizon. Models are constructed for the nonstationary and stationary policies. Compared to the existing approaches to similar stochastic systems, the proposed algorithms are computationally appealing. This approach can also be considered as a state reduction method for large scale Markov Decision Processes. A bound on cost function is developed and the concept of "refining observations" is introduced.

Key words: Markov Decision Process under Constraints, Method of Feasible Directions.

Science Code: 605.02.02

ÖZ

KISITLI GÖZLEM ALTINDA MARKOV KARAR SÜREÇLERİ

AVŞAR, Zeynep Müge
Yüksek Lisans Tezi, Endüstri Mühendisliği Anabilim Dalı
Tez Yöneticisi: Y. Doç. Dr. Yasemin Serin
Eylül, 1992, 88 sayfa.

Bu tezde, Markov Karar Süreçleri, kısıtlı gözlenebilirlik altında incelenmiş ve sonlu planlama süreleri için iskonto edilmiş toplam beklenen maliyeti enazlayan politikalar bulmak üzere algoritmalar geliştirilmiştir. Sistem zamana bağlı ve zamandan bağımsız politikalar olmak üzere iki durum için modellenmiştir. Varolan metotların bir takım fiziksel gözlem sorunları olan stokastik sistemlere yaklaşımları yapılması gerekli hesaplamalar bazında karşılaştırıldığında, önerilen metotlar gelişme sağlamıştır. Maliyet fonksiyonu üzerinde bir sınır geliştirilmiş ve detaylı gözlem yapma üzerine kurulu bir prosedür sunulmuştur. Kullanılan yaklaşıma Markov Karar Süreçleri için bir durum indirgeme metodu olarak bakılabilir.

Anahtar Sözcükler: Kısıt Altında Makov Karar Süreçleri, Geçerli Yönler Metodu.

Bilim Dalı Sayısal Kodu: 605.02.02

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Yasemin Serin, for her guidance, suggestions, great patience, and for all the time and effort she devoted to this study.

My thanks are also due to my colleagues in the Department of Industrial Engineering for their support and encouragement.

Finally, my warmest love and gratitude extend to my family for their understanding and supporting approach.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Introduction

The Markov Decision Process, MDP, is a methodology established as a tool to control stochastic systems. A stochastic system is modeled as a MDP by classifying the representative conditions as states of the system and by determining the alternative actions that can be taken with different costs. During a finite or infinite planning horizon, condition of the system is observed by the decision maker either continuously or periodically and accordingly an action is taken. In this thesis, we concentrate on periodically observed systems defined by a finite number of states and actions over a finite planning horizon. At the beginning of every period, depending on the state of the system and the action taken, the system immediately incurs a cost and moves to another state to be observed at the beginning of the next period. The events starting with observing a state and ending with incurring a cost are all assumed to occur instantaneously at the beginning of the periods. The Markov property results from the current state being probabilistically dependent on only the previous state and action. The MDP model finds a decision rule for each state with respect to the objective of minimizing a cost function. A policy is a collection of decision rules that describes the actions to be taken at every state. A policy is stationary if it uses the same decision rule at every period. Our objective, in this thesis, is to minimize the expected total discounted cost. A policy is optimal if the corresponding expected total discounted cost is minimum. The present work can easily be applied to the case of minimizing the average cost.

In computing an optimal policy for a completely observable MDP, where the true state of the system is known to the decision maker at every period, policy iteration method of Howard(1971a, 1971b) for infinite planning horizon and stochastic dynamic programming for finite planning horizon (Hillier and Lieberman, 1974) are commonly utilized methods. Both of the infinite and finite horizon MDP's can be formulated as linear models. The feasible policies for this problem are functions mapping the finite state set onto the finite action set. The optimal policy is deterministic in both cases, i.e., the decision rule assigns an action to every state. For the infinite horizon case, there is a stationary optimal policy.

In this thesis, we study MDP under a set of state observability constraints over a finite horizon. Serin(1989) studied the same problem over an infinite horizon. The idea of introducing unobservability constraints serves for systems in which obtaining information on exact state is physically infeasible or undesirable. A communication network whose states change rapidly is an example for the former case. Routing decisions at a given node should not wait for the information about the state of the whole system, but the status of that and maybe some neighboring nodes. Even if the exact state information can be obtained, it becomes obsolete before it is used. Large scale MDP's are examples for the latter case. The decision maker may prefer to avoid making detailed observation in order to reduce observation cost. The two different nature of stochastic control processes characterized above result in state observation problems or preferences, and are extensively studied in the literature.

Unobservability constraints imposed by the physical nature of the system itself are studied under the heading of Partially Observable Markov Decision Processes, POMDP (Monahan, 1982). POMDP can be represented as a completely observable MDP whose state is the posterior probability distribution over the finite state set, but the resulting state space is infinite. Then, the optimal policy to the new MDP is expressed by a function mapping the posterior probability distribution space to the finite action set, which brings about the difficulty in the steps of

computation and implementation of optimal policy. The computation of the optimal policy is based on dividing the posterior distribution space into a finite number of regions. Each of these regions corresponds to a different action. This procedure is performed at every decision epoch. Then, in order to employ the optimal policy proposed by Smallwood and Sondik(1973) for finite horizon problem, at each period the decision maker has to compute the posterior probability distribution. The optimal policy is not stationary with respect to the original state set.

Serin and Kulkarni partition the finite state set into a number of mutually exclusive and exhaustive subsets. The unobservability is based on taking the message of a subset if the process visits one of the states in that subset. Then, representing condition of the system by subsets, feasible policies are expressed over finite space of subsets, which makes the computation step simpler than that of POMDP. However, infinite horizon study of Serin on stationary policies shows that optimal policy may be randomized, i.e., the decision rule at a state may be a probability distribution over the action set, rather than an action, which makes up a difficulty in the implementation step. In this thesis, we study the same approach over a finite horizon by relaxing the stationarity requirement. In this case, the optimal policy turns out to be deterministic.

Since the completely observable MDP can be formulated as a linear model, aggregation/disaggregation procedures for stochastic LP's by Mendelssohn(1983) and Birge(1985a, 1985b) serve for the purpose of finding optimal policy with a reduction in the computational burden for large scale MDP's. Partitioning the state set is also an aggregation procedure and makes the computation step manageable. Note that aggregation or partitioning procedures also function as approximation schemes. In this thesis, we introduce the concept of refinement, analogous to disaggregation, by relaxing the partitioning constraints in a stepwise manner. In order to guide the decision maker in answering "how to partition the state space", "how to refine the observations" questions, we also develop bounds on optimal expected discounted cost to be compared to the observation cost, which is supposed to be

estimated with respect to partitions. If observing in more detail costs more, then the decision maker has an opportunity to decide on the detail of the observation with the information of possible improvement in the objective value. The refinement process can continue up to observing the original states of the MDP, for which the objective value is minimum but the observation cost is probably maximum.

Finally, considering the structure of the MDP model formulated with respect to a given partition of state space, we end up with the last extreme in literature: MDP under constraints. Ross(1989a) studies a completely observable MDP under a number of linear cost constraints, so linearity of the model is preserved. However, the partitioning constraints in our case are nonlinear. The similarity is that in both cases introduction of constraints leads to randomization in the optimal policy. A more detailed explanation is given in Section 1.3.

In this thesis, we analyze a MDP, whose state space is partitioned, under stationary and nonstationary policies for the objective of minimizing the expected total discounted cost over finite planning horizon, give some results for infinite planning horizon and discuss the refinement concept which represents the detail of information to gather.

1.2 Definitions

As the notational conventions, P(.) shows the probability of event {.} and |S| denotes the number of elements in the set S. $\|\alpha\|$ denote the norm of the vector α.

We start with definitions related with the MDP that is observed at discrete time intervals to be in one of the N states and accordingly one of the M actions is taken. We consider the time intervals with equal length and refer them as periods also. A period is named (indexed) by the number of intervals from the beginning of that period until the end of the planning horizon, e.g., period 5 means there are 5 decision epochs to go until the end of the planning horizon.

X_t is the random variable denoting the state of the system when there are t periods to go until the end of the planning horizon. It takes values in the finite state space $S=\{1, 2, ..., N\}$. An action, denoted by the random variable A_t, is taken by the controller of the process as soon as state of the system is observed at the beginning of period t. The finite action set is $A=\{1, 2, ..., M\}$. The stochastic process $\{(X_t, A_t): t=1, ..., T\}$, taking values from set SxA for a planning horizon of T periods, is called the core process. As a function of the state visited and the action taken in period t, an immediate cost $C(X_t, A_t)$, is incurred instantaneously. The transition probability of being in state j in a period given that the system was in state i and the action a was taken in the previous period is given by $P_{ij}(a)$. We consider homogeneous processes, so the expected cost per period and the transition probabilities are independent of time, $E(C(X_t=i, A_t=a))=c_{ia}$ for every t and $P(X_{t-1}=j| X_t=i, A_t=a)=P_{ij}(a)$ for all $t=1, ...,T$ and $i\varepsilon S$, $a\varepsilon A$.

The initial condition of the system is represented by the initial probability distribution, $p_i=P(X_T=i)$ for all $i\varepsilon S$.

Let the function $\alpha_{iat}(T)$ denote a decision rule which indicates the probability of taking action a given that the system is in state i in period t, when the planning horizon is T periods, $\alpha_{iat}(T)=P_\alpha(A_t=a| X_t=i)$ for all $i\varepsilon S$, $a\varepsilon A$, $t=1, ..., T$. Then, a T-period policy $\alpha(T)$ is a sequence of decision rules for each of the T periods. Noting that $P_{\alpha(T)}(.)$ represents the probability of event {.} under policy $\alpha(T)$, the set of feasible policies for a MDP is defined below:

$$\mathcal{A} = \{\alpha(T)\varepsilon R^{NMT} : \sum_{a=1}^{M} \alpha_{iat}(T)=1 \text{ for all } i\varepsilon S, t=1, ..., T \text{ and } \alpha_{iat}(T) \geq 0 \text{ for all } i\varepsilon S, a\varepsilon A, t=1, ..., T\} \quad (1.1)$$

We call the decision rules corresponding to a state i and/or period t under policy α as the partial policy α of state i in period t. If for every $i\varepsilon S$ and $t\varepsilon\{1,...,T\}$, there is an action a such that $\alpha_{iat}=1$, then α

is called a deterministic policy. Otherwise, if for at least one state the corresponding partial policy is not deterministic, then α is called a randomized policy.

In order to define the MDP restricted by some observability conditions of system or preferences of decision maker, we need to revise the definitions above and introduce additional notation. Considering the restrictions, we partition the state space into a collection $\mathbb{S}=\{S_1, S_2, ..., S_K\}$ of disjoint subsets and observe the system as in one of these subsets, rather than in an individual state. The new process which is observed at every period to be in one of the K subsets is called the observation process. Let $O=\{1, 2, ..., K\}$. Since the observation process is defined over the set O, we define the new stochastic observation process $\{Z_t: t=1, ..., T\}$, Z_t being the random variable denoting the subset that the system visits in period t, i.e., $X_t \in S_{Z_t}$. Equivalently, $Z_t=k$ if and only if $X_t \in S_k$.

The restriction we impose on the MDP states that the actions taken should be the same for all the states in the same subset. This means that, observing Z_t rather than X_t is sufficient to decide on an action, because every state in a subset is assigned the same decision rule. Let k(i) be the index denoting the state subset to which state i belongs. The decision rule indicating the probability of taking action a given that the system is in one of the states of subset k in period t is given by the function $\alpha_{kat}(T)$, when the planning horizon is T periods.

$$\alpha_{iat}(T)=P(A_t=a|\, X_t=i)$$

$$=P(A_t=a|\, Z_t=k(i)) \quad \text{for all } i\in S,\ a\in A,\ t=1, ..., T \qquad (1.2)$$

Then,

$$P(A_t=a|\, X_t=i)=P(A_t=a|\, X_t=j) \qquad \text{for all } i,j\in S_k$$

$$=\alpha_{kat}$$

The set of feasible policies for restricted MDP is

$$\mathcal{A}_1 = \{\alpha(T)\varepsilon R^{KMT} : \sum_{a=1}^{M} \alpha_{kat}(T)=1 \text{ for all } k\varepsilon O,\ t=1, ..., T$$
$$\text{and } \alpha_{kat}(T) \geq 0 \text{ for all } k\varepsilon O,\ a\varepsilon A,\ t=1, ..., T\} \tag{1.3}$$

As the definition implies, $\mathcal{A}_1$ consists of nonstationary policies, i.e., different decision rules can be used in different periods. From this point on, we fix the total number of periods within the planning horizon to T and drop the argument T for notational simplification. We also assume that we are given a partition $\mathcal{S}$ of the state space and the MDP under consideration is restricted with respect to this partition.

Let γ be the discount factor, $0<\gamma<1$. Given that the system is in state i in period t, the expected discounted cost of employing policy α in the last t periods is defined as follows:

$$v_{it}(\alpha) = \sum_{a=1}^{M} \alpha_{k(i)at}\left(c_{ia} + \gamma\sum_{j=1}^{N} P_{ij}(a)v_{j(t-1)}(\alpha)\right) \tag{1.4}$$

for all $i\varepsilon S$, $t=1, ..., T$, $\alpha\varepsilon\mathcal{A}_1$, where $v_{j0}(\alpha)$ is constant, e.g., $v_{j0}(\alpha)=0$.

Let $c_{it}(\alpha)$ be the expected immediate cost incurred under policy α, given that the system is in state i at the beginning of the period t.

$$c_{it}(\alpha) = \sum_{a=1}^{M} \alpha_{k(i)at}\, c_{ia} \quad \text{for all } i\varepsilon S,\ t=1, ..., T,\ \alpha\varepsilon\mathcal{A}_1 \tag{1.5}$$

Let $P_{ij}(\alpha, t)$ be the probability of being in state j at the beginning of the next period, given that the system is in state i when there are t periods to go and policy α is used.

$$P_{ij}(\alpha, t) = P_{\alpha}(X_{t-1}=j \mid X_t=i)$$

$$= \sum_{a=1}^{M} \alpha_{k(i)at} P_{ij}(a) \quad \text{for all } t=2, \ldots, T \text{ and } \alpha \in \mathcal{A}_1 \qquad (1.6)$$

Let $y_{iat}(\alpha)$ be the discounted probability of being in state i and taking action a in period t, under policy α

$$y_{iat}(\alpha) = \gamma^{(T-t)} P_\alpha(X_t=i, A_t=a) \quad \text{for all } i \in S, t=1, \ldots, T, a \in A$$

and $\alpha \in \mathcal{A}_1$, and $w_{it}(\alpha)$ be the discounted probability of being in state i in period t, under policy α

$$w_{it}(\alpha) = \gamma^{(T-t)} P_\alpha(X_t=i) \quad \text{for all } i \in S, t=1, \ldots, T \text{ and } \alpha \in \mathcal{A}_1.$$

For notational simplification, from now on, we drop the argument α in $v(\alpha)$, $y(\alpha)$, $w(\alpha)$.

Using the above notation, we define the following vectors:

$\mathbf{p}'=(p_1, p_2, \ldots, p_N)$ is the initial probability distribution, where ' denotes transpose of the vector,

P(a) is the transition matrix under the action a,

$P(\alpha, t)$ is the transition matrix under the partial policy α for period t, for all $\alpha \in \mathcal{A}_1$.

$$\mathbf{v}_t = (v_{1t}, v_{2t}, \ldots, v_{Nt})'$$

$$\mathbf{c}_t(\alpha) = (c_{1t}(\alpha), c_{2t}(\alpha), \ldots, c_{Nt}(\alpha))'$$

$$\mathbf{w}_t = (w_{1t}, w_{2t}, \ldots, w_{Nt})'$$

for all $t=1, \ldots, T$ and $\alpha \in \mathcal{A}_1$.

$\mathbf{y} = (\ldots, y_{iat}, \ldots)'$ of dimension NMT

$\mathbf{V}' = (\mathbf{v}_T', \mathbf{v}_{T-1}', ..., \mathbf{v}_1')$ which is also referred as cost function conditioned on the initial state,

$$\mathbf{C}(\boldsymbol{\alpha})' = (\mathbf{c}_T(\boldsymbol{\alpha})', \mathbf{c}_{T-1}(\boldsymbol{\alpha})', ..., \mathbf{c}_1(\boldsymbol{\alpha})')$$

$$\mathbf{W}' = (\mathbf{w}_T', \mathbf{w}_{T-1}', ..., \mathbf{w}_1')$$

for all $\boldsymbol{\alpha} \varepsilon \mathcal{A}_1$.

By definition of $\mathcal{A}_1$, we allow using nonstationary policies. Then, if we want to concentrate on stationary policies, dropping the time index we form the set $\mathcal{A}_2$ of all feasible stationary policies for restricted MDP.

$$\mathcal{A}_2 = \{\boldsymbol{\alpha} \varepsilon R^{KM} : \sum_{a=1}^{M} \alpha_{ka} = 1 \text{ for all } k \varepsilon O \text{ and } \alpha_{ka} \geq 0 \text{ for all } k \varepsilon O, a \varepsilon A\} \tag{1.7}$$

Note that $\mathcal{A}_2 \subset \mathcal{A}_1 \subset \mathcal{A}$. Notation for stationary policy case is as follows:

$$v_{it} = \sum_{a=1}^{M} \alpha_{k(i)a} \left(c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_{j(t-1)} \right) \tag{1.8a}$$

$$c_i(\boldsymbol{\alpha}) = \sum_{a=1}^{M} \alpha_{k(i)a} c_{ia} \tag{1.8b}$$

$$P_{ij}(\boldsymbol{\alpha}) = \sum_{a=1}^{M} \alpha_{k(i)a} P_{ij}(a) \tag{1.8c}$$

$P(\boldsymbol{\alpha})$ is the transition matrix under policy $\boldsymbol{\alpha} \varepsilon \mathcal{A}_2$ for all periods $t=1, ..., T$.

$C(\alpha)' = (c(\alpha)', c(\alpha)', ..., c(\alpha)')$ of dimension NT for all $\alpha \varepsilon \mathcal{A}_2$.

Our objective is to minimize the expected total discounted cost function, $\Phi(\alpha)$, over $\mathcal{A}_1$ so

$$\Phi(\alpha^*) = \underset{\alpha \varepsilon \mathcal{A}_1}{\text{minimum}} \{\Phi(\alpha)\} \tag{1.9a}$$

where

$$\Phi(\alpha) = E_\alpha \left[\sum_{t=1}^{T} \gamma^{(T-t)} C(X_t, A_t) \right] \text{for all } \alpha \varepsilon \mathcal{A}_1 \tag{1.9b}$$

where E_α represents the expectation under policy α. We also consider the similar problem where the policies are restricted to $\mathcal{A}_2$. For the infinite horizon problems, taking limit as t goes to infinity,

$$v_i = \sum_{a=1}^{M} \alpha_{k(i)a} \left(c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a)\, v_j \right) \tag{1.10a}$$

where α is defined over $\mathcal{A}_2$, and

$$\Phi(\alpha) = \lim_{T \to \infty} E_\alpha \left[\sum_{t=1}^{T} \gamma^{(T-t)} C(X_t, A_t) \right] \tag{1.10b}$$

1.3 Related Studies

In this section, we present related studies under three topics to which the problem we consider seems to be connected, namely POMDP, MDP under constraints and state aggregation/disaggregation in MDP's.

Within the context of stochastic control processes, Partially Observed Markov Decision Process , POMDP, is a generalization of Markov

Decision Process, MDP, allowing uncertainty in state observation and information acquisition about the current state. The decision maker is aware of the probabilistic relationship between the core process $\{(X_t, A_t): t=1, ..., T\}$ and the observation process $\{Z_t: t=1, ..., T\}$; an observation has the message that the current state is j when the true state is i with probability $q_{ij}=P(Z_t=j| X_t=i)$ for all $i,j \varepsilon S$. Recall that the problem that we concentrate on in this thesis is a special form of as POMDP, where the observation process has the message of subset k for all the states in S_k.

In order to convert POMDP into an equivalent and completely observable MDP, the state of the process is represented by the posterior probability distribution π_t, $\pi_t=(\pi_{1t},\pi_{2t},...,\pi_{Nt})'$; which is also called the information vector, where π_{it} is the probability of being in state i in period t. Then, the state space is given as

$$\Pi=\left\{ \pi_t \varepsilon R^N : \sum_{i=1}^{N} \pi_{it} = 1 \text{ and } \pi_{it} \geq 0 \text{ for all } i \varepsilon S \right\} \qquad (1.11)$$

for all t=1, ..., T. The well-known theorem leading to detection of Markov property in POMDP is given below (Monahan, 1982).

Theorem: For any fixed sequence of actions $A_T, A_{T-1}, ..., A_t \varepsilon A$, the sequence of probabilities of being in state i in period t, $\pi_{it}=P(X_t=i)$, $\{\pi_t: t=1, ..., T\}$ is a Markov Process, that is,

$$P(\pi_t| \pi_T, \pi_{T-1}, ..., \pi_{t+1}, A_{t+1}) = P(\pi_t| \pi_{t+1}, A_{t+1}) \quad \text{for all } t=1, ..., T$$

The optimal policy for a completely observable MDP is found by probabilistic dynamic programming moving backward period by period since the following recursion is satisfied by the optimum expected discounted cost function of the core process $\{(X_t, A_t): t=1, ..., T\}$

$$v_{it} = \underset{a \varepsilon A}{\text{minimum}} \left\{ c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_{j(t-1)} \right\} \quad \text{for all } i \varepsilon S, t=1, ..., T \qquad (1.12)$$

This recursive relationship takes a similar form under the continuous state space definition of POMDP, as given by the following theorem, where $T(\pi|\ Z_t=j,\ A_{t+1}=a)$ is defined as the posterior probability distribution in the period t given that action a was taken in the previous period and the current observation gives the message of state j.

"Theorem: The finite horizon optimum expected discounted cost function of POMDP, $\mathbf{v}(\pi)=(v_1(\pi),\ \ldots,\ v_T(\pi))'$, satisfies the following recursion

$$v_t(\pi)=\min_{a\in A}\left\{\sum_{i=1}^{N}\pi_{it}\,c_{ia}+\gamma\sum_{j=1}^{N}P(Z_{t-1}=j|\ \pi_t=\pi\ ,\ A_t=a)\ v_{t-1}\Big(T(\pi|\ Z_{t-1}=j,\ A_t=a)\Big)\right\}$$

for all t=2, ..., T (1.13a)

and

$$v_1(\pi)=\underset{a\in A}{\text{minimum}}\left\{\sum_{i=1}^{N}\pi_{i1}\,c_{ia}\right\} \qquad (1.13b)"$$

This theorem implies that the optimum policy is deterministic over the state space Π.

For optimal control of POMDP over a finite planning horizon, Smallwood and Sondik(1973) show that the conditional expected discounted cost function, $v_t(\pi)$, when there are a finite number of periods until the end of the planning horizon, satisfying the recursion given by (1.13a) is a piecewise-linear and concave function of the posterior state probabilities of the core process, implying that the space of posterior probability distribution in that period can be divided into a finite number of convex regions, separated by hyperplanes, over which the conditional cost function is linear. Hence, determining the gradient of the conditional cost function at any point in each region is sufficient to determine the optimum action to be taken, if the posterior distribution falls into that region. Reconstructing the regions for each decision epoch, the optimal policy is found. However, as the time horizon becomes longer, the number of regions may grow exponentially. From Blackwell's

study of "Discounted Dynamic Programming" (1965), it is known that the infinite horizon analog of recursion in (1.13a) is as follows:

$$v(\pi)=\underset{a\in A}{\text{minimum}}\left\{\sum_{i=1}^{N}\pi_i c_{ia}+\gamma\sum_{j=1}^{N}P(Z=j|\,\pi\,,A=a)\,v\big(T(\pi|\,Z=j,A=a)\big)\right\} \quad (1.14)$$

where v is the expected discounted cost when π is the initial distribution vector and the horizon is infinitely long. Solving the problem optimally requires minimization operation stated by (1.14) over all points in the space Π, meaning that (1.14) itself is not sufficient to develop a procedure to find the optimum stationary policy. Sondik(1978) studies this problem, where he introduces finitely transient policies over state space Π as the stationary policies such that after a finite number of periods the information vector is not mapped into the points of discontinuity of the policy. Sondik shows that if the optimum policy is finitely transient, then the optimum cost function $v^*(\pi)$ is a piecewise linear and continuous function, the converse is also true. Thus, in such a case the space Π can be divided into subsets in each of which the optimum action is the same for all the points. Utilizing the properties of finitely transient policies, Sondik makes an approximation to the cost function and he develops a policy iteration algorithm.

Before proceeding with a summary of the models incorporating the theory of POMDP's, we should note that Monahan(1982) gives a complete literature review on the problem of controlling processes with incomplete state information.

Eckles(1968) presents an expression similar to Bellman's 'Principal of Optimality' for the calculation of optimal maintenance policies when the decision maker is not informed completely about the system. Utilizing the complete history of all decisions and outcomes, the decision maker can calculate age of the system at any period. Age is assumed to characterize condition of the system. The optimum maintenance policy is calculated by dynamic-programming.

Wang(1977) models the Markovian replacement system whose condition is represented by its finite deterioration state which is not directly observable. Similar studies in the same area are carried out by Ross(1970, 1971) and Rosenfield(1976a, 1976b) for quality control and replacement models. The former treats the posterior probability distribution as the state of the system and the latter represents the state by (i, t) meaning that t periods ago the machine was known to be in state i and no new information is available since then. These last two studies' primary concern is to characterize the structure of the optimal policy rather than solving the problems.

The other applications of POMDP are cost control in accounting by Kaplan(1969), the internal control of a corporate control system by Hughes(1977), the learning process by Karush and Dear(1967), the teaching process and the health-care system by Smallwood(1971a, 1971b), the intraseasonal decisions of fishing vessel operators by Lane(1989).

There are a number of studies on structure of POMDP problem; as the convexity of policy regions of state space Π by Lovejoy(1987b), monotonicity of conditional expected cost functions by Albright(1979), analysis of the model by a new state variable using unnormalized conditional law by Borkar(1991), analysis of two-state case by Sernik and Marcus(1991). Within these, the idea forming the basis of the Sawaki's study(1983) results in partitioning the state space exactly the same way used in the present study. Representing the state of POMDP by posterior probability distributions, a completely observable MDP is obtained, but the state space becomes continuous. Sawaki calls a MP piecewise linear if there exists a simple partition $\mathcal{S}=\{ S_1, S_2, ..., S_K\}$ of state space S such that v_{it}'s are equal for all $i \in S_k$, t=1, ..., T and all k=1, ..., K, and defines a piecewise constant policy α if the same action is taken at each state of every state subset. In this respect, Sawaki's transformation of POMDP into piecewise linear ones has advantages for computer applications. His conclusion follows by showing that if a MP is piecewise linear, then the optimum policy α^* is piecewise constant. Smith(1967, 1971) partitions the state set as the same way Sawaki does,

and uses an (implicit) enumeration method to find an admissible policy, which is stationary and deterministic over subset space and minimizes the expected average cost over an infinite horizon.

The objective to approximate completely observable large scale MDP brings about the construction of computationally feasible bounds for both completely observable MDP and POMDP.

Approximation scheme developed by Lovejoy(1986) reduces the computational burden of policy iteration algorithm for MDP's with large number of states and/or actions. Lovejoy replaces the original MDP with a separable approximate MDP and that way generates bounds on the optimal policy for all states. By separability assumptions, both the state and action spaces are partitioned into the same number of subsets, the subsets that are affected by the currently visited subset are disjoint. These assumptions are quite restrictive and form a special case of the MDP problems that we concentrate on in this thesis. Considering some additional conditions, Lovejoy decomposes the MDP problem into several subproblems with smaller state and control spaces, and shows that at each iteration of the successive approximations of policy iteration algorithm, the calculations can be performed in a decomposed manner and converge to a separable, continuous optimal cost function and generate a separable optimum policy.

The underlying idea of Runggaldier's approach(1991) is to approximate the POMDP by a sequence of simpler approximating Dynamic Programs such that for each approximating problem it is possible to compute an optimal policy, and there exists an approximating problem such that the corresponding optimal policy, when suitably extended, is epsilon-optimal. The approximation is based on approximating the cost, transition, observation functions by a number of uniform step functions.

Lovejoy(1991) also studies bounds for POMDP. Since the usual transformation of POMDP into a completely observable MDP results in an uncountable state space, Lovejoy approximates the state space by a

finite grid of points and obtains approximate nonstationary and stationary policies.

A study of Monahan(1980), "Optimal Stopping in a POMDP with Costly Information" is based on the idea of incurring more cost to make a more detailed observation about the true current state of the core process, which is basically the same as the partition-dependent observation cost we introduce for the refinement discussion in Chapter III. In both Monahan's and our model, assumption is that the decision to make a more detailed observation, the observation itself and afterwards employement of an action are all instantaneous. At that point, it should be noted that for such instantaneous more detailed observations, our process may be restricted by observability constraints imposed as a nature of the system.

Iterative aggregation/disaggregation procedures by Mendelssohn(1983) and Birge(1985a, 1985b) serves for solving large scale completely observable MDP's optimally. At each iteration, an aggregate master problem and a sequence of smaller subproblems are solved. Each subproblem concentrates on a finite state, finite action MDP with a reduced state space, which is the point similar to partitioning a completely observable large scale MDP for the purpose of decreasing computational burden. The procedures use linear programming formulation of MDP and employs ideas for aggregation of LP's developed by Vakhutinskii and Dudkin(1973), Agafanov and Makarova(1976), Zipkin(1977,1980a, 1980b).

K. W. Ross(1989a) models the MDP as a linear program under the long-run average cost criterion subject to a number, say H, of cost constraints. He shows that there exists an optimal stationary policy with randomization in at most H states. For the case of single cost constraint, the deterministic, nonstationary round-robin type policies and steering policies are addressed as alternative optima. Due to nonlinearity of our model, in this thesis we do not study on the number of possible randomizations in the optimal policy.

A problem is called multilinear programming problem if its variables can be classified so that when all are fixed except the ones in one class, the resulting problem is a linear program. The models constructed in this study are multilinear models. The methodology and the computational experience seem to be in accordance with the results in Drenick(1992). Further study is required to state the exact relationship.

CHAPTER II

MDP WITH RESTRICTED OBSERVATIONS: NONSTATIONARY POLICIES

In this chapter, we study the problem of minimizing expected total discounted cost under unobservability constraints with nonstationary policies over finite planning horizon. We devote the first section to problem formulation. The last section is the presentation of two solution algorithms.

2.1. Finite Horizon Model

Consider a MDP $\{(X_t, A_t): t=1, ..., T\}$ with state space $S=\{1, ..., N\}$ and action space $A=\{1, ..., M\}$ over a T-period planning horizon. Suppose there does not exist any unobservability constraint, so the observation process is the same as the core process. A decision rule is defined by the probability α_{iat} to take action a if the state of the system is i when there are t periods until the end of the planning horizon. A collection of decision rules form a policy

$$\boldsymbol{\alpha}=(\alpha_{111}, \alpha_{121}, ..., \alpha_{NM1}, ..., \alpha_{11T}, ..., \alpha_{NMT})\varepsilon\mathcal{A}\subset R^{NMT}$$

where $\mathcal{A}$ is the set of feasible policies as defined in (1.1). Our optimality criterion is minimization of the expected total discounted cost $\Phi(\boldsymbol{\alpha})$ over the T-period planning horizon, i.e.,

$$\Phi(\boldsymbol{\alpha}) = E_{\alpha}\left[\sum_{t=1}^{T} \gamma^{(T-t)} C(X_t, A_t)\right] \tag{2.1}$$

$\Phi(\alpha)$ is well defined, since

$$\Phi(\alpha) \geq \frac{1-\gamma^{T}}{1-\gamma} \underset{i\varepsilon S,\, a\varepsilon A}{\text{minimum}} \left\{c_{ia}\right\} \tag{2.2}$$

and S and A are finite (Ross, 1983).

Compactness of the space $\mathcal{A}$ and continuity of the cost function given in (2.1) imply existence of an optimal policy α^* (Derman, 1970), i.e.,

$$\Phi(\alpha^*) = \underset{\alpha\varepsilon\mathcal{A}}{\text{minimum}} \left\{\Phi(\alpha)\right\} \tag{2.3}$$

By the expected total discounted cost of the system starting in state i and evolving for T periods when a given policy is employed and recalling that the probability of being in state i initially is given by p_i, we can rewrite the expected discounted cost function as (Ross, 1983)

$$\Phi(\alpha) = \sum_{i=1}^{N} p_i\, v_{iT} \tag{2.4}$$

The optimal policy for this problem can be found by the probabilistic dynamic programming moving backward period by period and using the recursive relationship

$$v^*_{it} = \underset{a\varepsilon A}{\text{minimum}} \left\{ c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a)\, v^*_{j(t-1)} \right\} \tag{2.5}$$

satisfied for all $i\varepsilon S$ and $t=1, \ldots, T$, assigning a constant value to v^*_{j0}'s (Hillier and Lieberman, 1974), e.g.,

$$v^*_0 = 0 \tag{2.6}$$

If the probability p_i of starting in state i is zero, the action assigned to the state i in the first period is arbitrary (Ross, 1989a).

Using the recursive relationship given in (2.5), we can state the expected total discounted cost minimization problem as follows:

Problem F:

$$\text{Maximize} \sum_{i=1}^{N} p_i v_{iT} \tag{2.7a}$$

subject to

$$v_{it} \leq c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_{j(t-1)} \quad \text{for all } i \in S,\ t=1, \ldots, T,\ a \in A \tag{2.7b}$$

$$v_{i0} = 0 \quad \text{for all } i \in S,\ a \in A \tag{2.7c}$$

$$v_{it} \text{ unrestricted} \quad \text{for all } i \in S,\ t=1, \ldots, T \tag{2.7d}$$

and the optimum policy α^* is given by

$$\alpha^*_{iat} = \begin{cases} 1 & \text{if } a = \underset{b \in A}{\operatorname{argmin}} \left\{ c_{ib} + \gamma \sum_{j=1}^{N} P_{ij}(b)\, v^*_{j(t-1)} \right\} \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

for all $i \in S$, $a \in A$, $t=1, \ldots, T$, where v^*_{jt}'s form the optimum solution to Problem F (Ross, 1983).

Let y_{iat} be the dual variable corresponding to the constraint (2.7b). Then, the dual of the Problem F is

Problem F_D:

$$\text{Minimize} \sum_{i=1}^{N} \sum_{a=1}^{M} c_{ia} \left\{ \sum_{t=1}^{T} y_{iat} \right\} \tag{2.9a}$$

subject to

$$\sum_{a=1}^{M} y_{iaT} = p_i \quad \text{for all } i \varepsilon S \tag{2.9b}$$

$$\sum_{a=1}^{M} \left\{ y_{iat} - \gamma \sum_{j=1}^{N} y_{ja(t+1)} P_{ji}(a) \right\} = 0 \text{ for all } i \varepsilon S,\ t=1, \ldots, T-1 \tag{2.9c}$$

$$y_{iat} \geq 0 \quad \text{for all } i \varepsilon S,\ a \varepsilon A,\ t=1, \ldots, T \tag{2.9d}$$

The dual variable y_{iat} can be interpreted as the discounted probability of being in state i and taking action a in period t, i.e., the optimum solution to Problem F_D is

$$y^*_{iat} = \gamma^{(T-t)} P_{\alpha^*}(X_t=i,\ A_t=a) \quad \text{for all } i \varepsilon S,\ a \varepsilon A,\ t=1, \ldots, T \tag{2.10}$$

Then, the constraint set (2.9c) shows the relation between the probability of being in state i at the beginning of time period t and probabilities of being in all possible states at the beginning of the previous period (t+1).

The Problem F_D is a linear programming formulation of the MDP. The optimum policy α^* is given by

$$\alpha^*_{iat} = \frac{y^*_{iat}}{\sum_{a=1}^{M} y^*_{iat}} \quad \text{for all } i \varepsilon S,\ a \varepsilon A,\ t=1, \ldots, T \tag{2.11}$$

$$= P_{\alpha^*}(A_t=a \mid X_t=i)$$

At a basic optimal solution, y^*_{iat} can take a positive value for at most one action while others are zero for each $i \in S$ and $t=1, \ldots, T$, which is in accordance with the implication of recursion (2.5), i.e., 'the deterministic optimum policy'. If it is not possible to be in state i at some period, some arbitrary action is assigned to that state (Ross, 1989a). Observe that summation of y^*_{iaT} over all actions is positive if the initial probability p_i is positive.

Now, we may define w_{it} as the discounted probability of being in state i at period t under a given policy α, i.e.,

$$w_{it} = \gamma^{(T-t)} P_\alpha(X_t=i) \quad \text{for all } i \in S,\ t=1, \ldots, T \tag{2.12}$$

Note that

$$w_{it} = \sum_{a=1}^{M} y_{iat} \quad \text{for all } i \in S,\ t=1, \ldots, T \tag{2.13}$$

From another point of view,

$$\gamma^{(T-t)} P_\alpha(X_t=i, A_t=a) = \left(\gamma^{(T-t)} P_\alpha(X_t=i)\right) P_\alpha(A_t=a \mid X_t=i)$$

where $P_\alpha(A_t=a \mid X_t=i) = \alpha_{iat}$. Then,

$$y_{iat} = w_{it}\, \alpha_{iat} \quad \text{for all } i \in S,\ a \in A,\ t=1, \ldots, T \tag{2.14}$$

Now, we are ready to consider the above MDP under unobservability constraints. Suppose that $\{Z_t: t=1, \ldots, T\}$ is the observation process defined over subset space $O=\{1,\ldots, K\}$ characterized by a partition $\mathbb{S}=\{S_1, \ldots, S_K\}$. If α is a policy with respect to partition $\mathbb{S}$, then the probability of taking action a at some period t is the same for all the states in the same subset. Then, in terms of formulation, unobservability constraints with respect to partition $\mathbb{S}$ are introduced by imposing

$$\alpha_{iat}=\alpha_{jat} \quad \text{for all i, j pair in the same subset}$$

to the feasible policy space $\mathcal{A}$, and the nonstationary policy space with respect to partition $\mathbf{S}$, $\mathcal{A}_1$, is obtained. A policy with respect to partition $\mathbf{S}$ is

$$\boldsymbol{\alpha}=(\alpha_{111}, \alpha_{121}, ..., \alpha_{KM1}, ..., \alpha_{11T}, ..., \alpha_{KMT})\in\mathcal{A}_1\subset R^{KMT}$$

where α_{kat} is the probability of taking action a when the system is observed to be in subset k in period t, i.e.,

$$\alpha_{kat}=P(A_t=a \mid Z_t=k)$$

$$=P(A_t=a \mid X_t\in S_k) \quad \text{for all } k\in O,\ a\in A,\ t=1, ..., T.$$

We refer to a completely observable MDP as an unrestricted MDP and the MDP under unobservability constraints as the restricted MDP with respect to partition $\mathbf{S}$. Note that unrestricted MDP is a MDP with respect to partition $\{S_1, ..., S_N\}$, where $S_i=\{i\}$ for all $i\in S$.

We present the expected discounted cost minimization problem for MDP with respect to partition $\mathbf{S}$ in Theorem II.2, where we define $\mathbf{R}$, $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{C}(\boldsymbol{\alpha})$ as vectors and $B(\boldsymbol{\alpha})$ as a square matrix, all of dimension NT.

$$\mathbf{R}'=(\mathbf{p}', 0, 0, ..., 0) \tag{2.15a}$$

$$\mathbf{V}'=(v_{1T}, ..., v_{NT}, ..., v_{N1})$$

$$=(\mathbf{v}_T', \mathbf{v}_{T-1}', ..., \mathbf{v}_1') \tag{2.15b}$$

$$\mathbf{W}'=(w_{1T}, ..., w_{NT}, ..., w_{N1})$$

$$=(\mathbf{w}_T', \mathbf{w}_{T-1}', ..., \mathbf{w}_1') \tag{2.15c}$$

$$\mathbf{C}(\boldsymbol{\alpha})'=(c_{1T}(\boldsymbol{\alpha}), ..., c_{NT}(\boldsymbol{\alpha}), ..., c_{11}(\boldsymbol{\alpha}), ..., c_{N1}(\boldsymbol{\alpha}))$$

$$=(c_T(\alpha)', c_{T-1}(\alpha)', \ldots, c_1(\alpha)') \tag{2.15d}$$

where $c_{it}(\alpha)$ is the expected immediate cost incurred in period t given that the system is in state i and under policy α,

$$B(\alpha)_{it,jn} = \begin{cases} 1 & \text{if } i=j \text{ and } t=n \\ -\gamma P_{ij}(\alpha, t) & \text{if } t-1=n \\ 0 & \text{otherwise} \end{cases} \tag{2.15e}$$

for all i, $j \in S$ and t,n=1, ..., T, where $P_{ij}(\alpha, t)$ is the transition probability from state i to j under the employment of partial policy α of period t. Recall from definitions that

$$c_{it}(\alpha) = \sum_{a=1}^{M} \alpha_{k(i)at}\, c_{ia} \quad \text{for all } i \in S,\ t=1, \ldots, T,\ \alpha \in \mathcal{A}_1$$

$$P_{ij}(\alpha, t) = P_\alpha(X_{t-1}=j \mid X_t=i)$$

$$= \sum_{a=1}^{M} \alpha_{k(i)at}\, P_{ij}(a) \quad \text{for all } i,j \in S,\ t=2, \ldots, T \text{ and } \alpha \in \mathcal{A}_1$$

In Lemma II.1, we point out some characteristics of the matrix $B(\alpha)$ and we use them while analyzing the restricted MDP problem.

Lemma II.1: For each $\alpha \in \mathcal{A}$,

a) $B(\alpha)$ is invertible.

b) $B(\alpha)^{-1}_{it,jn}$ is given below:

$$\begin{cases} 1 & \text{if } i=j \text{ and } t=n \\ \gamma^{(t-n)}\{P(\alpha, t)P(\alpha, t-1) \ldots P(\alpha, n+1)\}_{ij} & \text{if } t > n \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

for all i, $j \in S$ and t,n=1, ..., T.

Proof: Proof of this lemma is given in Appendix A.

Theorem II.2: The optimum policy α^* with respect to partition $\mathcal{S}$ for a MDP is given by the solution of the following problem.

Problem D_1:

$$\text{Minimize} \quad \mathbf{R}' \mathbf{V} \tag{2.17a}$$

subject to

$$B(\alpha)\, \mathbf{V} = C(\alpha) \tag{2.17b}$$

$$\sum_{a=1}^{M} \alpha_{kat} = 1 \quad \text{for all } k \in O,\ t=1, \ldots, T \tag{2.17c}$$

$$\alpha_{kat} \geq 0 \quad \text{for all } k \in O,\ a \in A,\ t=1, \ldots, T \tag{2.17d}$$

Proof: We start with the unrestricted MDP in Problem F_D. Introducing the unobservability constraints with respect to partition $\mathcal{S}$, we obtain Problem D_1.

Replacing all y_{iat}'s in Problem F_D by $w_{it}\, \alpha_{iat}$'s, we may restate Problem F_D as a nonlinear model. However, in order to construct the restricted MDP model with respect to partition $\mathcal{S}$, those replacements are made according to the relation

$$y_{iat} = w_{it}\, \alpha_{k(i)at} \quad \text{for all } i \in S,\ a \in A,\ t=1, \ldots, T \tag{2.18}$$

which are, in fact, the unobservability constraints with respect to partition $\mathcal{S}$ since α is defined over the policy space $\mathcal{A}_1$. Then, (2.9c) becomes

$$\sum_{a=1}^{M} w_{it}\, \alpha_{k(i)at} - \gamma \sum_{j=1}^{N} \sum_{a=1}^{M} w_{j(t+1)}\, P_{ji}(a)\, \alpha_{k(j)a(t+1)} = 0$$

and the MDP with respect to partition $\mathbf{S}$ is given as

Problem PD_1:

$$\text{Min} \sum_{i=1}^{N} \sum_{t=1}^{T} w_{it} c_{it}(\alpha) \tag{2.19a}$$

subject to

$$w_{iT} = p_i \quad \text{for all } i \varepsilon S \tag{2.19b}$$

$$w_{it} - \gamma \sum_{j=1}^{N} w_{j(t+1)} P_{ji}(\alpha, t+1) = 0 \text{ for all } i \varepsilon S,\ t=1, \ldots, T-1 \tag{2.19c}$$

$$\sum_{a=1}^{M} \alpha_{kat} = 1 \quad \text{for all } k \varepsilon O,\ t=1, \ldots, T \tag{2.19d}$$

$$\alpha_{kat} \geq 0 \quad \text{for all } k \varepsilon O,\ a \varepsilon A,\ t=1, \ldots, T \tag{2.19e}$$

$$w_{it} \geq 0 \quad \text{for all } i \varepsilon S,\ t=1, \ldots, T \tag{2.19f}$$

The constraint sets (2.19b) and (2.19c) can be written as $B(\alpha)' \mathbf{W} = \mathbf{R}$. Since $B(\alpha)$ is invertible as shown by Lemma II.1a, using

$$\mathbf{W}' = \mathbf{R}' B(\alpha)^{-1} \tag{2.20}$$

Problem PD_1 takes the following form:

Problem PD_1:

$$\text{Min } \mathbf{R}' B(\alpha)^{-1} C(\alpha) \tag{2.21a}$$

subject to

$$\sum_{a=1}^{M} \alpha_{kat} = 1 \quad \text{for all } k \varepsilon O,\ t=1, \ldots, T \tag{2.21b}$$

$$\alpha_{kat} \geq 0 \quad \text{for all } k \varepsilon O,\ a \varepsilon A,\ t=1, \ldots, T \tag{2.21c}$$

Defining $\mathbf{V}=B(\alpha)^{-1}C(\alpha)$, Problem D_1 follows.

From product $\mathbf{R}' B(\alpha)^{-1}$, the discounted probability of being in state i in period t under policy α can be written as

$$w_{it} = \sum_{j=1}^{N} \sum_{n=T}^{1} R_{jn} B(\alpha)^{-1}_{jn,it}$$

$$= \sum_{j=1}^{N} p_j B(\alpha)^{-1}_{jT,it}$$

For t=1, ..., T-1,

$$w_{it} = \sum_{j=1}^{N} p_j \left(\gamma^{(T-t)} P(\alpha, T) \ldots P(\alpha, t+1) \right)_{ji}$$

$$= \gamma^{(T-t)} \sum_{j=1}^{N} p_j \left(P(\alpha, T) \ldots P(\alpha, t+1) \right)_{ji}$$

$$= \gamma^{(T-t)} P_\alpha(X_t=i) \quad \text{for all } i \in S,\ \alpha \in \mathcal{A}_1 \tag{2.22a}$$

and for t=T,

$$w_{iT} = \sum_{j=1}^{N} p_j B(\alpha)^{-1}_{jT,iT}$$

$$= p_i$$

$$= P_\alpha(X_T=i) \quad \text{for all } i \in S,\ \alpha \in \mathcal{A}_1 \tag{2.22b}$$

The summation of the discounted probabilities in a period is given as

$$\sum_{i=1}^{N} w_{it} = \gamma^{(T-t)} \quad \text{for all } i \in S \tag{2.23}$$

Since the subset $\mathcal{A}_1$ defined by (2.17c) and (2.17d) is closed and bounded, it is compact, and the objective function is a continuous function of α. So, the optimum cost is

$$\Phi(\alpha^*) = \underset{\alpha \in \mathcal{A}_1}{\text{minimum}} \left\{ \mathbf{R}' \, B(\alpha)^{-1} \, C(\alpha) \right\} \tag{2.24}$$

where

$$\left(B(\alpha)^{-1} C(\alpha)\right)_{it} = c_{it}(\alpha) + \sum_{n=1 \text{ if } t>1}^{t-1} \sum_{j=1}^{N} \gamma^{(t-n)} \left(P(\alpha, t) \ldots P(\alpha, n+1)\right)_{ij} c_{jn}(\alpha)$$

$$= v_{it} \quad \text{for all } i \in S, \; t = 1, \ldots, T \tag{2.25}$$

Then,

$$\Phi(\alpha^*) = \underset{\alpha \in \mathcal{A}_1}{\text{minimum}} \left\{ \mathbf{R}' \, \mathbf{V} \right\}$$

$$= \underset{\alpha \in \mathcal{A}_1}{\text{minimum}} \left\{ \sum_{i=1}^{N} p_i \, v_{iT} \right\} \tag{2.26}$$

Note that Problem D_1 in (2.17), Problem PD_1 in (2.19) and Problem PD_1 in (2.21) are all different statements of the same model. In an open form, (2.17) can be written as

Problem D_1:

$$\Phi(\alpha^*) = \text{Min} \sum_{i=1}^{N} p_i \, v_{iT} \tag{2.27a}$$

subject to

$$v_{it}=\sum_{a=1}^{M} \alpha_{k(i)at}\left\{ c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)v_{j(t-1)} \right\} \text{ for all } i\varepsilon S,\ t=1, ..., T \qquad (2.27b)$$

$$v_{i0} = 0 \quad \text{for all } i\varepsilon S \qquad (2.27c)$$

$$\sum_{a=1}^{M} \alpha_{kat} = 1 \quad \text{for all } k\varepsilon O,\ t=1, ..., T \qquad (2.27d)$$

$$\alpha_{kat} \geq 0 \quad \text{for all } k\varepsilon O,\ a\varepsilon A,\ t=1, ..., T \qquad (2.27e)$$

$$v_{it} \text{ unrestricted for every } i\varepsilon S,\ t=1, ...,T \qquad (2.27f)$$

If we consider the structure of $B(\alpha)^{-1}$ given in Lemma II.1b or the recursive relationship (2.27b) between v_{it} 's Problem D_1 can also be stated as follows:

Problem D_1:

$$\text{Minimize } \mathbf{p}'\left(\mathbf{c}_T(\alpha)+\sum_{t=1}^{T-1} \gamma^{(T-t)} P(\alpha, T)...P(\alpha, t+1)\mathbf{c}_t(\alpha) \right) \qquad (2.28a)$$

subject to

$$\sum_{a=1}^{M} \alpha_{kat} = 1 \quad \text{for all } k\varepsilon O,\ t=1, ..., T \qquad (2.28b)$$

$$\alpha_{kat} \geq 0 \quad \text{for all } k\varepsilon O,\ a\varepsilon A,\ t=1, ..., T \qquad (2.28c)$$

From that point on, except otherwise stated, by Problem D_1 we refer to the model in (2.28).

Hence, we have a problem with a nonlinear objective function to be minimized over a set of linear constraints. Note that in the objective function, every type of decision variable appears at most to the first power in multiplication terms. Such nonlinear programming problems involving the sums of products of variables of at most first degree in objective function or constraints are called multilinear programming problems(Drenick, 1992). The term 'multilinear' comes from the fact that the variables can be combined into sets so that if all the variables are fixed except the ones in one set, the resulting problem is an LP. For the statement of Problem D_1 in (2.28), partial policy (α_{11t}, α_{12t}, ..., α_{KMt}) corresponding to period t forms a variable set, resulting in a total of T variable sets. In literature, the solution approaches to multilinear models are problem-specific. In a recent study, Drenick(1992) shows that many characteristics of linear duality theory are preserved in multilinear problems. Drenick states that "a globally optimal solution to a multilinear programming problem, if it exists, lies on the boundary of its feasible region, this is true also of a locally optimal solution". For the present case, this observation implies that at least for one subset k - action a - period t, the corresponding probability α^*_{kat} is zero. But, multilinearity leads to a stronger result for Problem D_1.

Whether the optimal policy is randomized or deterministic is important for implementation purposes. In this respect, the result stated in Theorem II.3 is important.

Theorem II.3: There exists a deterministic global optimal policy to Problem D_1.

Proof: Problem D_1 is bounded, and has an optimal solution. Suppose the optimum policy α^* in periods 1, 2, ..., t-1, t+1, ..., T is known and we are to find the optimum partial policy for period t, (α_{11t}, ..., α_{1Mt}, α_{21t}, ..., α_{KMt}). Then, Problem D_1 becomes the following linear program:

$$\text{Min} \sum_{i=1}^{N} w^*_{it} \sum_{a=1}^{M} \alpha_{k(i)at} \left(c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a)\, v^*_{j(t-1)} \right) + \text{constant} \qquad (2.29a)$$

subject to

$$\sum_{a=1}^{M} \alpha_{kat} = 1 \quad \text{for all } k \in O \tag{2.29b}$$

$$\alpha_{kat} \geq 0 \quad \text{for all } k \in O,\ a \in A \tag{2.29c}$$

where w^* and v^* are constants defined by optimum policy of periods T, T-1, ..., t+1 and t-1, ..., 2, 1, respectively, and the constant term is the optimum expected discounted cost for the first (T-t-1) periods.

The extreme points of the feasible policy space defined by (2.29b) and (2.29c) correspond to deterministic policies. When the objective function is stated as

$$\sum_{k=1}^{K} \sum_{a=1}^{M} \alpha_{kat} \sum_{i \in S_k} w^*_{it} \left(c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a)\, v^*_{j(t-1)} \right) \tag{2.30}$$

it can be easily observed that there is a deterministic optimal policy for period t.

$$\alpha^*_{kat} = \begin{cases} 1 & \text{if } a = \underset{b \in A}{\operatorname{argmin}} \left\{ \sum_{i \in S_k} w^*_{it} \left(c_{ib} + \gamma \sum_{j=1}^{N} P_{ij}(b)\, v^*_{j(t-1)} \right) \right\} \\ 0 & \text{otherwise} \end{cases} \tag{2.31}$$

for all $k \in O$. Assigning this deterministic partial policy to period t and repeating the same discussion for all periods one by one, a deterministic optimal policy is obtained.

2.2. Solution Method

Since Problem D_1 is defined over a finite planning horizon, with finite action and state spaces, a deterministic optimal solution can

be found by complete enumeration. But, as the number of actions or subsets or length of the planning horizon increases, complete enumeration becomes cumbersome.

For solving Problem D_1, dynamic programming can not be used due to the structure of the recursive relationship in (2.27b). For a policy to be feasible with respect to partition $\mathcal{S}$, the decision rules must be the same for every state of a subset. Consequently, the optimality equation takes the form given in (2.31), unlike (2.5).

In order to obtain a good solution to this problem, we use the method of feasible directions (Bazaraa and Shetty, 1979). In the rest of this section, we assume we have a fixed policy $\boldsymbol{\alpha}$ which is restricted with respect to partition $\mathcal{S}$. A vector $\boldsymbol{\beta}$ is a feasible direction for Problem D_1 at $\boldsymbol{\alpha}$, if

$$\boldsymbol{\beta} \neq \mathbf{0} \tag{2.32}$$

and $\exists\ \theta(\boldsymbol{\beta})>0 \ni$

$$(\boldsymbol{\alpha} + \theta\boldsymbol{\beta}) \varepsilon \boldsymbol{\lambda}_1 \quad \text{for all } \theta \varepsilon (0, \theta(\boldsymbol{\beta})] \tag{2.33}$$

We pick a feasible policy and search for feasible descent directions for the problem at that point. If we can find such a direction, a line search may provide us a policy with some improvement in the objective function value. Then, we repeat the procedure at the improved policy and continue that way until we are stuck in finding feasible descent directions. Note that such points are either local minimum or saddle point.

Lemma II.4: $\boldsymbol{\beta}$ is a feasible direction at $\boldsymbol{\alpha}$, if

$$\begin{aligned} \beta_{kat} &\geq 0 \qquad \text{for } \alpha_{kat} = 0 \\ \beta_{kat} &\leq 0 \qquad \text{for } \alpha_{kat} = 1 \end{aligned} \tag{2.34a}$$

and

$$\sum_{a=1}^{M} \beta_{kat} = 0 \quad \text{for all } k \varepsilon O,\ t=1, \ldots, T \tag{2.34b}$$

Proof: Since $(\alpha+\theta\beta)\varepsilon\mathcal{A}_1$, we should have

$$0 \le \alpha_{kat} + \theta\beta_{kat} \le 1 \quad \text{for all } k\varepsilon O,\ a\varepsilon A,\ t=1, \ldots, T \text{ for } \theta\varepsilon(0,\ \theta(\beta)],$$

implying (2.34a). $(\alpha+\theta\beta)\varepsilon\mathcal{A}_1$ also implies

$$\sum_{a=1}^{M} \left(\alpha_{kat} + \theta\beta_{kat}\right) = 1 \quad \text{for all } k\varepsilon O,\ t=1, \ldots, T$$

$$\sum_{a=1}^{M} \alpha_{kat} + \theta\sum_{a=1}^{M} \beta_{kat} = 1$$

where

$$\sum_{a=1}^{M} \alpha_{kat} = 1 \quad \text{since } \alpha\varepsilon\mathcal{A}_1$$

Then,

$$\sum_{a=1}^{M} \beta_{kat} = 0 \quad \text{for all } k\varepsilon O,\ t=1, \ldots, T$$

which is (2.34b).

In order to bound the set of feasible directions, we use normalization constraints. However, due to the structural properties of Problem D_1, use of two different normalization constraints brings about two solution algorithms. Algorithm I iterates between deterministic policies because the corresponding normalization constraint guarantees policy improvement with a step size of one unit. On the other hand, the constraint used for Algorithm II causes the algorithm to consider also the randomized policies.

From now on, we use the term "the direction vector $\boldsymbol{\beta}$ makes changes in the partial policy of period t". It means that at least two components of $\boldsymbol{\beta}$ corresponding to period t, say β_{kat} and β_{kbt} for a, b∈A are nonzero for some k, so that the partial policy corresponding to period t in $\boldsymbol{\alpha}+\theta\boldsymbol{\beta}$ is different than that of $\boldsymbol{\alpha}$.

Since Problem D_1 is a minimization problem, during the search for a minimum the directions that we concentrate on are descent directions. Negative directional derivative of Φ at a point $\boldsymbol{\alpha}$ in the direction of $\boldsymbol{\beta}$ is sufficient for $\boldsymbol{\beta}$ to be a descent direction.

Hence, we now aim to find a feasible direction $\boldsymbol{\beta}$ for which $\nabla\boldsymbol{\Phi}(\boldsymbol{\alpha})'\,\boldsymbol{\beta}<0$, i.e.,

$$\sum_{k=1}^{K}\sum_{a=1}^{M}\sum_{t=1}^{T}\frac{\partial\Phi(\boldsymbol{\alpha})}{\partial\alpha_{kat}}\beta_{kat}<0.$$

The gradient vector of the objective function is of dimension KTM as given below:

$$\nabla\boldsymbol{\Phi}(\boldsymbol{\alpha})=\left(\ldots,\frac{\partial\Phi(\boldsymbol{\alpha})}{\partial\alpha_{kat}},\ldots\right)' \tag{2.35}$$

From (2.27a), the partial derivative of $\Phi(\boldsymbol{\alpha})$ with respect to α_{kat} is

$$\frac{\partial\Phi(\boldsymbol{\alpha})}{\partial\alpha_{kat}}=\sum_{i=1}^{N}p_i\frac{\partial v_{iT}}{\partial\alpha_{kat}} \tag{2.36}$$

Differentiating (2.27b) with respect to α_{kat}, we get

$$\frac{\partial v_{in}}{\partial \alpha_{kat}} = \begin{cases} c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_{j(t-1)} & \text{if } i \in S_k \text{ and } t=n \\ \gamma \sum_{j=1}^{N} P_{ij}(\alpha, n) \frac{\partial v_{j(n-1)}}{\partial \alpha_{kat}} & \text{if } 1 \le t \le (n-1) \\ 0 & \text{otherwise} \end{cases} \tag{2.37}$$

for all n, t=1, ..., T and $k \in O$, $a \in A$. Using (2.36),

$$\frac{\partial \Phi(\alpha)}{\partial \alpha_{kaT}} = \sum_{i \in S_k} p_i \left\{ c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_{j(T-1)} \right\}$$

where $p_i = w_{iT}$,

$$\frac{\partial \Phi(\alpha)}{\partial \alpha_{ka(T-1)}} = \sum_{i=1}^{N} p_i \gamma \sum_{j \in S_k} P_{ij}(\alpha, T) \left\{ c_{ja} + \gamma \sum_{m=1}^{N} P_{jm}(a) v_{m(T-2)} \right\}$$

where $\gamma \sum_{i=1}^{N} p_i P_{ij}(\alpha, T) = w_{j(T-1)}$,

$$\frac{\partial \Phi(\alpha)}{\partial \alpha_{ka(T-2)}} = \sum_{i=1}^{N} p_i \gamma \sum_{j=1}^{N} P_{ij}(\alpha, T) \gamma \sum_{m \in S_k} P_{jm}(\alpha, T-1) \left\{ c_{ma} + \gamma \sum_{\alpha=1}^{N} P_{m\alpha}(a) v_{\alpha(T-2)} \right\}$$

where $\gamma \sum_{j=1}^{N} \sum_{i=1}^{N} p_i P_{ij}(\alpha, T) \; P_{jm}(\alpha, T-1) = \gamma \sum_{j=1}^{N} w_{j(T-1)} P_{jm}(\alpha, T-1)$

$$= w_{m(T-2)}.$$

In general,

$$\frac{\partial\Phi(\alpha)}{\partial\alpha_{kat}} = \sum_{i\in S_k} w_{it}\left\{ c_{ia} + \gamma\sum_{j=1}^{N} P_{ij}(a)v_{j(t-1)} \right\} \tag{2.38}$$

for all $t=1, \ldots, T$ and $k\in O$, $a\in A$. The expression in (2.38) is a weighted sum of the test quantities used in (2.5) in stochastic dynamic programming where the weights are the discounted probabilities. Note that, for any given policy α, corresponding $\mathbf{w}$ and $\mathbf{v}$ can easily be computed using (2.19b), (2.19c) and (2.27b), (2.27c), respectively.

Before we explain how a feasible descent direction is selected, it is necessary to give some results about the step size θ, which are crucial for the solution procedure. Suppose we have a feasible descent direction β. If we substitute the new policy vector $\alpha+\theta\beta$ in the objective function of Problem D_1, we obtain a function of θ. We refer $\Phi(\alpha+\theta\beta)$ by $f(\theta)$ for the given α and β, to be minimized over $(0, \theta(\beta)]$.

The behaviour of $f(\theta)$ as a function of the step size θ, at a given point and along a given direction, is analyzed in Proposition II.5.

Proposition II.5: Suppose a feasible direction β makes changes in partial policies of n periods. Then, $f(\theta)$ is a polynomial of order at most n over $(0, \theta(\beta)]$.

Proof: Let the feasible direction β make changes in partial policy of n periods. If we substitute the new policy vector $\rho=\alpha+\theta\beta$ in objective function (2.28a) of Problem D_1, θ appears only in the transition matrices and the cost vectors of these n periods as

$$c_{it}(\rho) = \sum_{a=1}^{M} \left(\alpha_{k(i)at} + \theta\beta_{k(i)at}\right) c_{ia}$$

$$= c_{it}(\alpha) + \theta\sum_{a=1}^{M} \beta_{k(i)at}\, c_{ia}$$

$$P_{ij}(\rho, t) = \sum_{a=1}^{M} \left(\alpha_{k(i)at} + \theta\beta_{k(i)at}\right) P_{ij}(a)$$

$$= P_{ij}(\alpha, t) + \theta \sum_{a=1}^{M} \beta_{k(i)at} P_{ij}(a)$$

Hence, if t is one of these n periods, each entry in $P(\alpha+\theta\beta, t)$ is linear in θ. The transition matrices corresponding to other periods are constants. Hence, the highest order term in (2.28a) is

$$\gamma^{(T-1)} \mathbf{p}' \left(P(\rho, T)\, P(\rho, T-1) \ldots P(\rho, 2)\, \mathbf{c}_1(\rho)\right)$$

from which Proposition II.5 follows.

Corollary II.6: If $\nabla\Phi(\alpha)'\beta < 0$ for a feasible direction β changing the partial policy of one period only, then $f(\theta)$ is a linear function of θ over $(0, \theta(\beta)]$.

Thus, minimum $f(\theta)$ is at $\theta=\theta(\beta)$. Similarly, if $\nabla\Phi(\alpha)'\beta = 0$ for such a β, then $f(\theta) = \Phi(\alpha)$ for all $\theta\varepsilon(0, \theta(\beta)]$.

If the directional derivative at a given point α is negative for some feasible direction β, then there exist other policies, with lower expected total discounted cost values, to which we can reach by proceeding along that direction. If the minimum value of directional derivative at α is nonnegative, i.e., if the necessary Kuhn-Tucker conditions are satisfied, one of the following cases holds:

Case 1) If $\nabla\Phi(\alpha)'\beta = 0$ for some feasible β, then α is either a saddle point or a local minimum. Note that this case may also result from a zero gradient vector $\nabla\Phi(\alpha)$.

Case 2) If $\nabla\Phi(\alpha)'\beta > 0$ for all feasible β at α, then α is a local minimum.

2.2.1 Algorithm I

The first algorithm we develop iterates between deterministic policies, using the fact that there exists a deterministic global optimal policy to Problem D_1. In order to guarantee improvement at each iteration from one deterministic policy to another, a descent direction is selected in such a way that the policy improvement is achieved through changes in the partial policy of only one period, although there may be other periods implying improvement, i.e., contributing the directional derivative with a negative value. From Corollary II.6, proceeding along such a direction causes improvement at a constant rate. Then, if the search procedure starts with a deterministic policy, iterations occur between deterministic policies by taking a step of size one at each iteration.

As in the case of policy iteration algorithm of Howard(1971a, 1971b) for unrestricted MDP, we may proceed along a steepest descent direction for solving Problem D_1, if there exists any. In order to find the steepest descent direction at a given deterministic policy α when $\alpha_{kd(k,t)t}=1$ for $d(k,t)\epsilon A$, we need to solve the problem of minimizing the directional derivative under the constraints defining a feasible direction at α and the step size is fixed at $\Theta=1$.

Problem SD1(α):

$$\text{Min} \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{a=1}^{M} \beta_{kat} \frac{\partial \Phi(\alpha)}{\partial \alpha_{kat}} \tag{2.39a}$$

subject to

$$\sum_{a=1}^{M} \beta_{kat} = 0 \quad \text{for all } k\epsilon O,\ t=1, \ldots, T \tag{2.39b}$$

$$\beta_{kd(k,t)t} \leq 0 \tag{2.39c}$$

$$0 \leq \beta_{kat} \leq 1 \quad \text{for all } k\epsilon O,\ t=1, \ldots, T,\ a\epsilon A-\{d(k,t)\} \tag{2.39d}$$

β makes changes in one period only (2.39e)

$$\beta = 0 \tag{2.39f}$$

β_{kat} unrestricted (2.39g)

An optimal solution to (2.39) that is also the steepest descent direction under the constraint (2.39e) can be identified as given in Theorem II.7. Let

$$r(k,t) = \underset{a \in A \ni a \neq d(k,t)}{\text{minimum}} \left\{ \frac{\partial \Phi(\alpha)}{\partial \alpha_{kat}} \right\} - \frac{\partial \Phi(\alpha)}{\partial \alpha_{kd(k,t)t}} \tag{2.40}$$

Theorem II.7: The optimal solution to (2.39) which is the steepest descent direction at the deterministic point α, under the constraint that allows change in partial policy of only one period, is

$$\beta^*_{kat} = \begin{cases} 1 & \text{if } t = t^* \text{ and } a = \underset{b \in A \ni b \neq d(k,t)}{\text{argmin}} \left\{ \frac{\partial \Phi(\alpha)}{\partial \alpha_{kbt}} \right\} \\ -1 & \text{if } t = t^* \text{ and } a = d(k,t) \\ 0 & \text{otherwise} \end{cases} \tag{2.41}$$

where if there exists a k-t pair such that $r(k,t) < 0$, then

$$t^* = \underset{t=1,\ldots,T}{\text{argmin}} \left\{ \sum_{k=1}^{K} \min \{r(k,t), 0\} \right\} \tag{2.42}$$

If $r(k,t) \geq 0$ for all $k \in O$, $t = 1, \ldots, T$, then

$$t^* = \underset{t=1,\ldots,T}{\text{argmin}} \left\{ \min_{k \in O} \{r(k,t)\} \right\} \tag{2.43}$$

and the corresponding minimum directional derivative is

$$\sum_{k\in O} \min \left\{ \underset{a\in A \ni a\neq d(k,t^*)}{\text{minimum}} \left\{ \sum_{i\in S_k} w_{it^*} \left\{ c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_{j(t^*-1)} \right.\right.\right.$$

$$\left.\left.\left. - c_{id(k,t^*)} - \gamma \sum_{j=1}^{N} P_{ij}(d(k,t^*)) v_{j(t^*-1)} \right\} \right\}, 0 \right\} \quad \text{if } \exists \text{ a k-t pair} \ni r(k,t)<0$$

and

$$\underset{k\in O}{\text{minimum}} \left\{ r(k,t^*) \right\} \quad \text{if } r(k,t)\geq 0 \text{ for all } k\in O,\ t=1, \ldots, T \tag{2.44}$$

Proof: Let the current policy α be deterministic with $\alpha_{kd(k,t)t}=1$ for all $k\in O$, $t=1, \ldots, T$, $d(k,t)\in A$. The problem in (2.39) is separable, i.e., it can be written as KT subproblems, one for each subset in each period. The optimal value of such a subproblem is $r(k, t)$ as defined in (2.40) which selects an action better than the current action $d(k, t)$ if there is any, and the optimal solution to such a subproblem is

$$\beta_{kat} = \begin{cases} 1 & \text{if } a = \underset{b\in A \ni b\neq d(k,t)}{\text{argmin}} \left\{ \dfrac{\partial \Phi(\alpha)}{\partial \alpha_{kbt}} \right\} \\ -1 & \text{if } a = d(k, t) \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } k\in O,\ t=1,\ldots,T \tag{2.45a}$$

If $r(k,t)<0$ for some pair, then the resulting minimum contribution from period t is

$$\sum_{k\in O \ni r(k,t)<0} r(k,t) \tag{2.45b}$$

and (2.42) follows. If $r(k,t)\geq 0$ for all, then the objective is minimum when t is given as (2.43) since $\boldsymbol{\beta}=\mathbf{0}$. Then, the optimal objective value is

$$\sum_{k\in O} \text{minimum} \left\{ r(k,t^*), 0 \right\} \quad \text{if } \exists \text{ a k-t pair} \ni r(k,t)<0$$

$$\underset{k\in O}{\text{minimum}}\{r(k,t^*)\} \qquad \text{if } r(k,t)\geq 0 \text{ for all } k\in O,\ t=1,\ldots,T$$

and the proof is complete.

Now, we can give Algorithm I.

Algorithm I

Step 0) Initialization: Choose a deterministic initial policy $\alpha\in\mathcal{A}_1$.

Step 1) Policy Evaluation: Compute the expected total discounted cost $\Phi(\alpha)$.

Step 2) Policy Improvement: From (2.42) and (2.43), find the period t^* implying the highest improvement among all periods and the minimum directional derivative from (2.44). If it is positive, then stop, $\Phi(\alpha)$ is a local minimum. If it is equal to zero, then stop, $\Phi(\alpha)$ is either a local minimum or a saddle point; if negative, go to Step 3.

Step 3) For every subset k, if

$$\underset{\substack{a\in A\ni\\ a\neq d(k,t^*)}}{\text{minimum}}\left\{\sum_{i\in S_k} w_{it^*}\left(c_{ia}+\gamma\sum_{j=1}^{N}P_{ij}(a)v_{j(t^*-1)}-c_{id(k,t^*)}-\gamma\sum_{j=1}^{N}P_{ij}(d(k,t^*))v_{j(t^*-1)}\right)\right\}$$

is negative, let

$$\alpha_{kat^*}=\begin{cases}1 & \text{if } a=\underset{b\in A\ni b\neq d(k,t^*)}{\text{argmin}}\left\{\sum_{i\in S_k} w_{it^*}\left(c_{ib}+\gamma\sum_{j=1}^{N}P_{ij}(b)v_{j(t^*-1)}\right)\right\}\\ 0 & \text{otherwise}\end{cases}$$

for all $k\in O$. Go to Step 1.

2.2.2. Algorithm II

Using feasible descent directions that change the partial policy of only one period at an iteration may cause the algorithm to terminate after a large number of iterations. Another disadvantage of Algorithm I is the risk of termination with a deterministic local optimum or saddle point in spite of the fact that there may exist a randomized local optimum or a saddle point with a lower expected cost. The reason is that, Algorithm I does not take randomized policies into account. Along the line between two deterministic policies of two successive iterations of Algorithm I, there can not be any point satisfying necessary Kuhn-Tucker conditions, because the expected cost function decreases linearly. However, there can be randomized policies which do not lie on any such line.

In this section, we propose another algorithm, Algorithm II, for solving Problem D_1 which allows changes in partial policy of every period in an iteration and proceeds along the steepest descent directions. Recall from Proposition II.5 that directions making changes in more than one period cause the expected cost function to be a nonlinear function of step size. Then, for minimizing the cost function along such directions, the policy improvement step must include a line search, which is the computational burden of this algorithm and may slow the algorithm in terms of the computation time required until termination. On the other hand, it may decrease the number of iterations. Relaxing the restriction on the direction of Algorithm I, Algorithm II is given the chance of detecting randomized local optima or saddle points. The line search procedure of solving

$$\underset{\theta\varepsilon(0,\theta(\beta^*)]}{\text{minimum}} \{f(\theta)\}$$

requires more effort than nonstationary case for which θ is fixed as $\theta=1$. In the present study, line search is carried out by evaluating $f(\theta)$ which is a polynomial of degree at most T, at 100 points and a minimum is selected as the optimal step size.

The steepest descent direction for Algorithm II is the solution of Problem SD1 where (2.39c) and (2.39d) are replaced with

$$\beta_{kat} \geq 0 \qquad \text{for } \alpha_{kat} = 0$$

and

$$\beta_{kat} \leq 0 \qquad \text{for } \alpha_{kat} = 1$$

and (2.39e) is replaced with

$$\sum_{a=1}^{M} \left| \beta_{kat} \right| \leq \kappa \quad \text{for each k and t} \tag{2.46}$$

κ is a constant. If

$$r(k, t) = \underset{a \in A \ni \alpha_{kat} < 1}{\text{minimum}} \left\{ \frac{\partial \Phi(\alpha)}{\partial \alpha_{kat}} \right\} - \underset{a \in A \ni \alpha_{kat} > 0}{\text{maximum}} \left\{ \frac{\partial \Phi(\alpha)}{\partial \alpha_{kat}} \right\} \tag{2.47}$$

is negative, then the contribution of k-t pair to the directional derivative is given a weight of one. For the restricted MDP problem with stationary policy over infinite horizon, Serin(1989) utilizes unit norm and accordingly introduces the steepest descent direction, which will be summarized in Chapter III. From Serin(1989), it immediately follows that the direction to be selected at an iteration of Algorithm II is as characterized by Corollary II.8.

Corollary II.8: The steepest descent direction for Algorithm II is

$$\beta^*_{kat}=\begin{cases}\frac{1}{2} & \text{if } r(k,t)<0 \text{ and } a=\underset{b\in A\ \exists\ \alpha_{kbt}<1}{\text{argmin}}\left\{\frac{\partial\Phi(\alpha)}{\partial\alpha_{kbt}}\right\}\\ -\frac{1}{2} & \text{if } r(k,t)<0 \text{ and } a=\underset{b\in A\ \exists\ \alpha_{kbt}>0}{\text{argmax}}\left\{\frac{\partial\Phi(\alpha)}{\partial\alpha_{kbt}}\right\}\\ 0 & \text{otherwise}\end{cases} \tag{2.48}$$

if there is a k-t pair such that $r(k,t)<0$. If $r(k,t)\geq 0$ for all $k\in O$, $t=1, \ldots, T$, $\boldsymbol{\beta}^*$ gives the nonnegative minimum directional derivative as

$$\beta^*_{kat}=\begin{cases}\frac{1}{2} & \text{if } k=k^* \text{ and } t=t^* \text{ and } a=\underset{b\in A\ \exists\ \alpha_{kbt}<1}{\text{argmin}}\left\{\frac{\partial\Phi(\alpha)}{\partial\alpha_{kbt}}\right\}\\ -\frac{1}{2} & \text{if } k=k^* \text{ and } t=t^* \text{ and } a=\underset{b\in A\ \exists\ \alpha_{kbt}>0}{\text{argmax}}\left\{\frac{\partial\Phi(\alpha)}{\partial\alpha_{kbt}}\right\}\\ 0 & \text{otherwise}\end{cases} \tag{2.49}$$

where $(k^*,t^*)=\underset{k\in O,\ t=1,\ldots,T}{\text{argmin}}\{r(k,t)\}$.

Computation of maximum feasible step size at a randomized policy is not straightforward as in the case of Algorithm I. The following lemma shows how to find the feasibility limit on the step size for a given direction at a given point.

Lemma II.9: The maximum feasible step size $\Theta(\boldsymbol{\beta}^*)$ along the feasible direction $\boldsymbol{\beta}^*$ is given by

$$\underset{\substack{k\in O\\ t=1,\ldots,T}}{\text{minimum}}\left\{\underset{a\in A\ \exists\ \beta^*_{kat}>0}{\text{minimum}}\left\{2\left(1-\alpha_{kat}\right)\right\},\ \underset{a\in A\ \exists\ \beta^*_{kat}<0}{\text{minimum}}\left\{2\alpha_{kat}\right\}\right\} \tag{2.50}$$

Proof: From the definition of $\theta(\boldsymbol{\beta}^*)$, $(\boldsymbol{\alpha}+\theta\boldsymbol{\beta}^*)\varepsilon\boldsymbol{\lambda}_1$ for all $\theta\varepsilon(0,\theta(\boldsymbol{\beta}^*)]$. It implies that

$$0 \le \alpha_{kat} + \theta\beta^*_{kat} \le 1 \qquad \text{for all } k\varepsilon O,\ a\varepsilon A,\ t=1, \ldots, T$$

Then, we have the following inequalities for all $k\varepsilon O$, $t=1, \ldots, T$:

$$\theta \ge \frac{-\alpha_{kat}}{\beta^*_{kat}} \quad \text{if } \beta^*_{kat} > 0 \tag{2.51a}$$

$$\theta \le \frac{1-\alpha_{kat}}{\beta^*_{kat}} \quad \text{if } \beta^*_{kat} > 0 \tag{2.51b}$$

$$\theta \le \frac{\alpha_{kat}}{|\beta^*_{kat}|} \quad \text{if } \beta^*_{kat} < 0 \tag{2.51c}$$

$$\theta \ge \frac{\alpha_{kat}-1}{|\beta^*_{kat}|} \quad \text{if } \beta^*_{kat} < 0 \tag{2.51d}$$

where (2.51a) and (2.51d) are redundant. Since $1-\alpha_{kat} \ge \alpha_{kbt}$ for $a,b\varepsilon A$, using the definition of $\boldsymbol{\beta}^*$, the tightest restriction on θ is given as

$$\theta \le \underset{\substack{k\varepsilon O \\ t=1,\ldots,T}}{\text{minimum}} \left\{ \underset{a\varepsilon A\ \exists\ \beta^*_{kat}>0}{\text{minimum}} \left\{2\left(1-\alpha_{kat}\right)\right\}, \underset{a\varepsilon A\ \exists\ \beta^*_{kat}<0}{\text{minimum}} \left\{2\alpha_{kat}\right\} \right\} \tag{2.52}$$

Corollary II.10: If the feasible direction $\boldsymbol{\beta}$ at a point $\boldsymbol{\alpha}$, makes changes in the partial policy of one subset, m, in only one period, n, such that $\beta_{mun}>0$, $\beta_{mdn}<0$ and $\beta_{kat}=0$ otherwise, then the maximum feasible step size $\theta(\boldsymbol{\beta})$ is given by $2\alpha_{mdn}$.

As mentioned before, it is possible that Algorithm II may stop with a randomized policy although there is a deterministic global optimum. The last part in this subsection gives a procedure to obtain a deterministic policy which is equivalent to a global optimal randomized policy. If the algorithm stops with a randomized policy, then there must

exist a deterministic equivalent, if it is global optimal. If there is no equivalent deterministic policy, then the current randomized policy is not global optimal. If there is, it is possibly but not necessarily global optimal. Before giving details of this procedure, a property of the directional derivative at a randomized policy is given in Lemma II.11.

Lemma II.11: Consider a point $\boldsymbol{\alpha}$ at which for each feasible direction the corresponding directional derivative is nonnegative. If the policy $\boldsymbol{\alpha}$ is randomized, then the minimum directional derivative at $\boldsymbol{\alpha}$ is zero.

Proof: Since we assume

$$\nabla\Phi(\boldsymbol{\alpha})'\,\boldsymbol{\beta} \geq 0 \quad \text{for all feasible } \boldsymbol{\beta} \tag{2.53}$$

is satisfied at $\boldsymbol{\alpha}$, it is sufficient to prove that there exists a feasible direction $\boldsymbol{\beta}$ such that the corresponding directional derivative $\nabla\Phi(\boldsymbol{\alpha})'\,\boldsymbol{\beta}$ is zero.

Let the partial policy corresponding to some subset m in period n be randomized. $\boldsymbol{\alpha}$ satisfies necessary Kuhn-Tucker conditions since the minimum directional is nonnegative. Let μ_{kat} and λ_{kt} be the lagrangian multipliers corresponding to the constraints (2.28c) and (2.28b), respectively. Since the objective function (2.28a) and the constraint functions are all differentiable and continuous, the necessary Kuhn-Tucker conditions for Problem D_1 can be given as follows:

$$\frac{\partial\Phi(\boldsymbol{\alpha})}{\partial\alpha_{kat}} + \lambda_{kt} - \mu_{kat} = 0 \tag{2.54a}$$

$$-\,\mu_{kat}\,\alpha_{kat} = 0 \tag{2.54b}$$

$$\mu_{kat} \geq 0 \tag{2.54c}$$

$$\lambda_{kt} \text{ unrestricted} \tag{2.54d}$$

for all $k \in O$, $a \in A$, $t = 1, \ldots, T$.

Let a direction $\boldsymbol{\beta}$ make changes in policy of only state subset m in period n and for actions such that $0<\alpha_{man}<1$. When the relation (2.54a) is used, the corresponding directional derivative

$$\nabla\Phi(\boldsymbol{\alpha})'\,\boldsymbol{\beta} = \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{a=1}^{M} \beta_{kat}\,\frac{\partial\Phi(\boldsymbol{\alpha})}{\partial\alpha_{kat}}$$

takes the following form:

$$\nabla\Phi(\boldsymbol{\alpha})'\,\boldsymbol{\beta} = \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{a=1}^{M} \beta_{kat}\,(\mu_{kat} - \lambda_{kt})$$

$$= \sum_{k=1}^{K}\sum_{t=1}^{T}\left(\sum_{a=1}^{M} \beta_{kat}\,\mu_{kat} - \lambda_{kt}\sum_{a=1}^{M}\beta_{kat}\right) \tag{2.55}$$

where $\sum_{a=1}^{M}\beta_{kat} = 0$ by definition of $\boldsymbol{\beta}$ from (2.34b). Then,

$$\nabla\Phi(\boldsymbol{\alpha})'\,\boldsymbol{\beta} = \sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{a=1}^{M} \beta_{kat}\,\mu_{kat} \tag{2.56}$$

Since β_{kat} is nonzero only for m, a, n such that $0<\alpha_{man}<1$

$$\nabla\Phi(\boldsymbol{\alpha})'\,\boldsymbol{\beta} = \sum_{a\in A\ \ni\ 0<\alpha_{man}<1} \beta_{man}\,\mu_{man} \tag{2.57}$$

On the other hand, from (2.54b), μ_{man} is zero for all $a\in A \ni 0<\alpha_{man}<1$. So, the result follows.

Coming back to the procedure, let α be a global optimal randomized policy. Our purpose, now, is to obtain a deterministic policy which is equivalent to α. Let

$$SR(\alpha) = \{t=1, ..., T: 0<\alpha_{kat}<1 \text{ for some } k\varepsilon O \text{ and } a\varepsilon A\} \qquad (2.58a)$$

i.e., α makes randomization in period t if $t\varepsilon SR(\alpha)$, and

$$H_t(\alpha) = \{k\varepsilon O : 0 < \alpha_{kat} < 1 \text{ for some } a\varepsilon A\} \qquad (2.58b)$$

for all t=1, ..., T, i.e., policy α is randomized for subset k in period t if $k\varepsilon H_t(\alpha)$. Note that for every $t\varepsilon SR(\alpha)$, $H_t(\alpha)$ is not empty since α is randomized. Select β so that it changes partial policy of only subset, m, for $m\varepsilon H_n(\alpha)$ and $n\varepsilon SR(\alpha)$. There must exist u, $d\varepsilon A \ni 0<\alpha_{mdn}<1$ and $0<\alpha_{mun}<1$ for some $m\varepsilon O$, n=1, ..., T. Let $\beta_{mun}>0$, $\beta_{mdn}<0$ and $\beta_{kat}=0$ otherwise. Then, the directional derivative $\nabla\Phi(\alpha)'\beta$ is zero from Lemma II.11. From Corollary II.6 and Corollary II.10,

$$f(\theta)=\Phi(\alpha) \quad \text{for} \quad \theta=2\alpha_{mdn}$$

So, the policy ρ, obtained by proceeding along the given direction β with a step size of $\theta(\beta)$, is an alternative optimum to α and

$$\rho_{man} = \alpha_{man} \quad \text{for all } a\varepsilon A-\{u, d\} \qquad (2.59)$$

$$\rho_{mun} = \alpha_{mun} + \frac{\theta}{2}$$

$$= \alpha_{mun} + \alpha_{mdn} \qquad (2.60)$$

$$\rho_{mdn} = \alpha_{mdn} - \frac{\theta}{2}$$

$$= 0 \qquad (2.61)$$

$$\rho_{kat} = \alpha_{kat} \quad \text{for all } k\varepsilon O-\{m\}, t\varepsilon\{1, ..., T\}-\{n\}, a\varepsilon A \qquad (2.62)$$

Note that since α is global optimum and ρ is an alternative to α, (2.53) still holds at ρ. We, now, concentrate on the following two cases for the purpose of obtaining a deterministic policy as an alternative to randomized global optimal policy.

Case 1) If $\alpha_{mun} + \alpha_{mdn} = 1$, i.e., partial policy ρ is deterministic for subset m in period n with $\rho_{mun} = 1$ and $\rho_{mdn} = 0$, then in period n the number of subsets with randomized policy is decreased by one, i.e., since $m \notin H_n(\rho)$ $|H_n(\rho)| = |H_n(\alpha)| - 1$.

Case 2) If $\alpha_{mun} + \alpha_{mdn} < 1$, i.e., partial policy ρ is still randomized for subset m in period n, then in order to reach an alternative policy τ, which is deterministic for subset m in period n with $|H_n(\tau)| = |H_n(\alpha)| - 1$, we need to continue proceeding along directions having the properties of β. Afterwards, we concentrate on another subset $l \varepsilon H_n(\tau)$; if we face with case (1), then we obtain another alternative optimum to α with a deterministic partial policy for subset l in period n. Continuing that way for every subset for which the partial policy α in period n is randomized, we reach an alternative optimum η which is deterministic in period n, i.e., $H_n(\eta)$ is empty and $|SR(\eta)| = |SR(\alpha)| - 1$. If $SR(\eta)$ is empty, then η is a completely deterministic alternative optimum policy to α. On the other hand, if $SR(\eta)$ is not empty, i.e., there are still some periods with randomized partial policies, then we need to continue with another period in the set $SR(\eta)$ until it becomes an empty set.

This procedure is based on proceeding along directions with zero directional derivative. This may result in cycling. For that reason, we do not integrate this procedure into Algorithm II.

Now, Algorithm II can be presented.

Algorithm II

Step 0) Initialization: Choose an initial policy $\alpha \varepsilon \mathcal{A}_1$.

Step 1) Policy Evaluation: Compute the expected total discounted cost $\Phi(\boldsymbol{\alpha})$.

Step 2) Policy Improvement: From (2.48) or (2.49), find the minimum directional derivative. If it is positive, then stop, $\Phi(\boldsymbol{\alpha})$ is a local minimum. If it is equal to zero, then stop, $\Phi(\boldsymbol{\alpha})$ is either a local minimum or a saddle point; if negative, go to Step 3.

Step 3) Compute direction $\boldsymbol{\beta}^*$ by (2.48) and the maximum feasible step size by (2.50).

Step 4) Make a line search on $f(\theta)$ over $\theta\varepsilon(0, \theta(\boldsymbol{\beta}^*)]$. Pick a policy corresponding to some step size θ_0 with $\Phi(\boldsymbol{\alpha}+\theta_0\boldsymbol{\beta}^*)<\Phi(\boldsymbol{\alpha})$, then take $\boldsymbol{\alpha}+\theta_0\boldsymbol{\beta}^*$ as the new policy and go to Step 1.

Example II.1: In order to check validity of the proposed algorithms for unrestricted MDP, we solve an example from Hillier and Lieberman (1974: 560-561).

N=4 states, M=3 actions, K=4 subsets, S= {{1}, {2}, {3}, {4} }, T=3 periods

$$P(1)=\begin{bmatrix} 0 & 0.875 & 0.0625 & 0.0625 \\ 0 & 0.75 & 0.125 & 0.125 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad P(2)=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad P(3)=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$\mathbf{p}'=[0.2, 0.4, 0.3, 0.1]$ and $\gamma=0.9$

$c_{11}=0$, $c_{12}=4$, $c_{13}=6$; $c_{21}=1$, $c_{22}=4$ $c_{23}=6$; $c_{31}=3$, $c_{32}=4$, $c_{33}=6$; $c_{41}=\infty$, $c_{42}=\infty$, $c_{43}=6$

For infinite cost values, we take 100. Let policy 1,2 and 3 denote the following:

1 ... $(\alpha_{k1t}, \alpha_{k2t}, \alpha_{k3t})=(1,0,0)$
2 ... $(\alpha_{k1t}, \alpha_{k2t}, \alpha_{k3t})=(0,1,0)$
3 ... $(\alpha_{k1t}, \alpha_{k2t}, \alpha_{k3t})=(0,0,1)$

Initial policy is (1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1), where the first four entries show the policies selected for every subset in the first period of the planning horizon, t=3, so semicolons separate periods.

A special form of Algorithm I making changes in one subset at an iteration, gives

Table 2.1 Textbook Example

Iteration	Policy (α)	$\Phi(\alpha)$
1	(1, 1, 1, 1; 1, 1, 1, 1; 1, 1, 1, 1)	72.47
2	(1, 1, 1, 1; 1, 1, 1, 3; 1, 1, 1, 1)	24.86
3	(1, 1, 1, 1; 1, 1, 1, 3; 1, 1, 1, 3)	14.63
4	(1, 1, 1, 3; 1, 1, 1, 3; 1, 1, 1, 3)	4.805
5	(1, 1, 2, 3; 1, 1, 1, 3; 1, 1, 1, 3)	4.07
6	(1, 1, 2, 3; 1, 1, 2, 3; 1, 1, 1, 3)	3.996

which is the same result given in the textbook.

Example II.2: Consider the following three state MDP with M=2 actions and K=2 subsets.

$$S_1=\{1\} \text{ and } S_2=\{2, 3\}$$

$$P(1)=\begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.4 & 0.2 & 0.4 \end{bmatrix} \text{ and } P(2)=\begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.1 & 0.3 & 0.6 \\ 0.2 & 0.1 & 0.7 \end{bmatrix}$$

$$\mathbf{p}'=[0.2, 0.5, 0.3] \text{ and } \gamma=0.8$$

$$c_{11}=2,\ c_{12}=3 \text{ and } c_{21}=19,\ c_{22}=2 \text{ and } c_{31}=3,\ c_{32}=24$$

Let policy 1 and 2 denote the following:

1 ... $(\alpha_{k1t}, \alpha_{k2t}) = (1,0)$
2 ... $(\alpha_{k1t}, \alpha_{k2t}) = (0,1)$

a) A summary of first iteration, by using Algorithm I for a 4-period planning horizon, is given below:

Initial policy is (2, 2; 2, 2; 2, 2; 2, 2), $\Phi(\alpha)$=37.39.

Partial derivatives of $\Phi(\alpha)$ are given in Table 2.2:

Table 2.2 Example for Algorithm I

Action a	Period t	Partial derivatives of $\Phi(\alpha)$ with respect to α_{kat} for k equals to 1	2
1	1	0.23	2.88
2	1	0.34	7.25
1	2	0.99	7.68
2	2	1.15	16
1	3	2.40	15.8
2	3	2.48	26.12
1	4	4.66	28.7
2	4	4.73	32.66

Table 2.3 Example for Algorithm I

Period t	r(k,t) for k equals to 1	2	$\nabla\Phi(\alpha)'\beta$	
1	-0.11	-4.37	-4.48	no change
2	-0.15	-8.34	-8.49	no change
3	-0.08	-10.31	-10.39	change
4	-0.072	-3.95	-4.022	no change

Then t*=3. Continuing that way,

Table 2.4 Example for Algorithm I

Iteration	Policy (α)	$\Phi(\alpha)$
1	(2,2; 2,2; 2,2; 2,2)	37.39
2	(2,2; 1,1; 2,2; 2,2)	26.99
3	(2,2; 1,1; 1,1; 2,2)	23.91
4	(2,2; 2,1; 1,1; 2,2)	23.85

b) A summary of first iteration, by using Algorithm II for a 4-period planning horizon, is given below:

Initial policy is (2, 1; 2, 1; 2, 1; 2, 1), $\Phi(\alpha)$=30.53.

Partial derivatives of $\Phi(\alpha)$ are given in Table 2.5:

Table 2.5 Example for Algorithm II

Action a	Period t	Partial derivatives of $\Phi(\alpha)$ with respect to α_{kat} for k equals to 1	2
1	1	0.32	4.63
2	1	0.48	3.52
1	2	2.15	9.4
2	2	2.1	6.76
1	3	4.18	15.46
2	3	4.07	11.56
1	4	4.52	26.11
2	4	4.42	20.4

Table 2.6 Example for Algorithm II

Period t	r(k,t) for k equals to 1	2	$\nabla\Phi(\alpha)'\beta$	
1	-0.16	-1.11	-1.27	change

Table 2.6 (Cont'd) Example for Algorithm II

Period t	r(k,t) for k equals to 1	2	$\nabla\Phi(\alpha)'\beta$	
2	0.06	-2.63	-2.63	change
3	0.12	-3.9	-3.9	change
4	0.1	-5.72	-5.72	change

Then, Algorithm II makes a line search to make changes partial policy of subset 2 in all periods and of subset 1 in period 1. Continuing that way,

Table 2.7 Example for Algorithm II

	Policy α in the period				
Iteration	4	3	2	1	$\Phi(\alpha)$
1	(0,1,1,0)	(0,1,1,0)	(0,1,1,0)	(0,1,1,0)	30.53
2	(0,1,.67,.33)	(0,1,.67,.33)	(0,1,.67,.33)	(.33,.67,.67,.33)	28.29
3	(.33,.67,.34,.66)	(.33,.67,1,0)	(.33,.67,1,0)	(.66,.34,1,0)	26.73
4	(0,1,.01,.99)	(0,1,1,0)	(0,1,.67,.33)	(.99,.01,.67,.33)	24.66
5	(0,1,0,1)	(.01,.99,1,0)	(.01,.99,.68,.32)	(1,0,.68,.32)	24.60
6	(0,1,0,1)	(.13,.87,1,0)	(.13,.87,.8,.2)	(1,0,.8,.2)	24.53
7	(0,1,0,1)	(0,1,1,0)	(0,1,.93,.07)	(1,0,.67,.33)	24.45
8	(0,1,0,1)	(0,1,1,0)	(.07,.93,1,0)	(1,0,.6,.4)	24.36
9	(0,1,0,1)	(0,1,1,0)	(.67,.33,1,0)	(1,0,0,1)	23.78
10	(0,1,0,1)	(0,1,1,0)	(1,0,1,0)	(1,0,0,1)	23.70

c) A special form of Algorithm I, making changes in one subset at an iteration for a 4-period planning horizon gives

Table 2.8 Example for Algorithm I

Iteration	Policy (α)	$\Phi(\alpha)$
1	(2,1; 2,1; 2,1; 2,1)	30.53
2	(2,2; 2,1; 2,1; 2,1)	24.82
3	(2,2; 2,1; 2,1; 2,2)	24.09
4	(2,2; 2,1; 1,1; 2,2)	23.85

Table 2.8 (Cont'd) Example for Algorithm I

Iteration	Policy (α)	$\Phi(\alpha)$
5	(2,2; 2,1; 1,1; 1,2)	23.7

d) In Table 2.11, we give the best policies we reach for 4 to 18 period planning horizons where policy 1, 2, 3 and 4 are given as

1 ... 1 for subset 1 and 1 for subset 2
2 ... 2 for subset 1 and 2 for subset 2
3 ... 2 for subset 1 and 1 for subset 2
4 ... 1 for subset 1 and 2 for subset 2

Note the convergence of the optimal costs and the policies.

Example II.3: The same as Example II.2 except the cost figures;

$c_{11}=2$, $c_{12}=3$ and $c_{21}=4$, $c_{22}=2$ and $c_{31}=3.5$, $c_{32}=2.4$

a) Algorithm I for a 4-period planning horizon gives

Table 2.9 Example for Algorithm I

Iteration	Policy (α)	$\Phi(\alpha)$
1	(1,1; 1,1; 1,1; 1,1)	9.92
2	(1,2; 1,1; 1,1; 1,1)	8.55
3	(1,2; 1,2; 1,1; 1,1)	7.65
4	(1,2; 1,2; 1,2; 1,1)	7.02

b) A special form of Algorithm I, making changes in one subset at an iteration for a 4-period planning horizon gives

Table 2.10 Example for Algorithm I

Iteration	Policy (α)	$\Phi(\alpha)$
1	(2,1; 2,1; 2,1; 2,1)	10.61
2	(2,2; 2,1; 2,1; 2,1)	9.19
3	(2,2; 2,2; 2,1; 2,1)	8.27
4	(2,2; 2,2; 2,2; 2,1)	7.60
5	(2,2; 2,2; 2,2; 2,2)	7.07
6	(1,2; 2,2; 2,2; 2,2)	6.85
7	(1,2; 1,2; 2,2; 2,2)	6.70
8	(1,2; 1,2; 1,2; 2,2)	6.57
9	(1,2; 1,2; 1,2; 1,2)	6.47

Table 2.11 Policies for Increasing Horizon Length

		Policies selected for period																	
T	$\Phi(\alpha^*)$	T	T-1	T-2	...												...	2	1
4	23.703	2	3	1	4														
5	26.812	2	3	1	2	1													
6	29.626	2	3	1	2	3	1												
7	31.615	2	3	1	2	1	2	1											
8	33.244	2	3	1	2	3	1	2	1										
9	34.685	2	3	1	2	3	1	2	3	1									
10	35.703	2	3	1	2	3	1	2	1	2	1								
11	36.537	2	3	1	2	3	1	2	3	1	2	1							
12	37.274	2	3	1	2	3	1	2	3	1	2	3	1						
13	37.796	2	3	1	2	3	1	2	3	1	2	1	2	1					
14	38.222	2	3	1	2	3	1	2	3	1	2	3	1	2	1				
15	38.600	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1			
16	38.867	2	3	1	2	3	1	2	3	1	2	3	1	2	1	2	1		
17	39.086	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	1	
18	39.279	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1

CHAPTER III

MDP WITH RESTRICTED OBSERVATIONS: STATIONARY POLICIES

In this chapter, we study the MDP restricted with respect to a partition $\mathbf{S}$ under stationary policies. The objective is again minimization of the expected total discounted cost. The first two sections are devoted to the finite and infinite planning horizon models. In the last section, the concept of refinement is introduced.

3.1 Finite Horizon Model

In Chapter II, we obtain Problem D_1 by introducing partition constraints to the unrestricted MDP model. Now, in addition to that, we restrict policies to stationary ones, i.e., we impose the following constraints

$$\alpha_{kat} = \alpha_{ka} \quad \text{for all } t=1, \ldots, T \text{ and } k \in O, a \in A$$

to the policy space $\mathcal{A}_1$, and the stationary policy space with respect to partition $\mathbf{S}$, namely $\mathcal{A}_2$, is obtained. Then, the transformation (2.14) in obtaining restricted problem from unrestricted MDP takes the following form:

$$y_{iat} = w_{it}\, \alpha_{k(i)a} \quad \text{for all } t=1, \ldots, T \text{ and } i \in S, a \in A \tag{3.1}$$

Before giving the model, we recall some definitions: the expected immediate cost incurred in a period given that the system is in state i and under the policy α is

$$c_i(\alpha) = \sum_{a=1}^{M} \alpha_{k(i)a} c_{ia} \quad \text{for all } i \in S,\ \alpha \in \mathcal{A}_2$$

and the transition probability to reach state j given that the system is in state i and under the policy α is

$$P_{ij}(\alpha) = \sum_{a=1}^{M} \alpha_{k(i)a} P_{ij}(a) \quad \text{for all } i, j \in S \text{ and } \alpha \in \mathcal{A}_2$$

$C(\alpha)$ is a vector of dimension NT as

$$C(\alpha)' = (c_1(\alpha), \ldots, c_N(\alpha), \ldots, c_1(\alpha), \ldots, c_N(\alpha)).$$

The problem of finding a stationary policy to minimize expected total discounted cost can be written as

Problem D_2:

$$\text{Minimize} \quad \sum_{i=1}^{N} c_i(\alpha) \sum_{t=1}^{T} w_{it} \tag{3.2a}$$

subject to

$$w_{iT} = p_i \quad \text{for all } i \in S \tag{3.2b}$$

$$w_{it} - \gamma \sum_{j=1}^{N} w_{j(t+1)} P_{ji}(\alpha) = 0 \quad \text{for all } i \in S,\ t = 1, \ldots, T-1 \tag{3.2c}$$

$$\sum_{a=1}^{M} \alpha_{ka} = 1 \quad \text{for all } k \in O \tag{3.2d}$$

$$\alpha_{ka} \geq 0 \quad \text{for all } k \in O,\ a \in A \tag{3.2e}$$

$$w_{it} \geq 0 \quad \text{for all } i \in S,\ t = 1, \ldots, T \tag{3.2f}$$

Similar to Lemma II.1, by partitioning $B(\alpha)$ into submatrices, $B(\alpha)^{-1}$ is given as

$$B(\alpha)^{-1}_{it,jn}=\begin{cases}1 & \text{if } i=j \text{ and } t=n\\ \gamma^{(t-n)}\ P(\alpha)^{(t-n)}_{ij} & \text{if } t>n\\ 0 & \text{otherwise}\end{cases} \tag{3.3}$$

As in Problem D_1, the product $B(\alpha)^{-1}\ C(\alpha)$ gives the conditional cost function $V=(v_{1T}, ..., v_{NT}, ..., v_{11}, ..., v_{N1})'$.

$$\left(B(\alpha)^{-1}C(\alpha)\right)_{it}=\sum_{j=1}^{N}\left(I+...+\gamma^{(t-1)}P(\alpha)^{(t-1)}\right)_{ij}c_j(\alpha)$$

$$=v_{it} \tag{3.4}$$

Unlike nonstationary policy case, deterministic global optimal policy can not be guaranteed for Problem D_2. For that reason, as a solution method we revise Algorithm II that allows randomized policies.

Starting with the gradient vector of the objective function in (3.2a),

$$\nabla\Phi(\alpha)'=(\ ...,\ \frac{\partial\Phi(\alpha)}{\partial\alpha_{ka}},\ ...\ ) \tag{3.5}$$

is of dimension KM where

$$\frac{\partial\Phi(\alpha)}{\partial\alpha_{ka}}=\sum_{t=1}^{T}\sum_{i\in S_k}w_{it}\left\{c_{ia}+\gamma\sum_{j=1}^{N}P_{ij}(a)\ v_{j(t-1)}\right\} \tag{3.6}$$

for all $k\in O$, $a\in A$.

Similar to Lemma II.4, the set of feasible directions, is defined for every direction β that satisfies the following conditions:

$$\beta_{ka} \geq 0 \quad \text{if } \alpha_{ka} = 0$$

$$\beta_{ka} \leq 0 \quad \text{if } \alpha_{ka} = 1 \tag{3.7a}$$

and

$$\sum_{a=1}^{M} \beta_{ka} = 0 \quad \text{for all } k \varepsilon O \tag{3.7b}$$

Now, we use a normalization constraint on feasible directions that allows to make changes in two actions in one subset and one period only.

Then, the steepest descent direction at a point $\boldsymbol{\alpha}$ is found by solving Problem SD2($\boldsymbol{\alpha}$).

Problem SD2($\boldsymbol{\alpha}$):

$$\text{Minimize} \quad \sum_{k=1}^{K} \sum_{a=1}^{M} \beta_{ka} \frac{\partial \Phi(\boldsymbol{\alpha})}{\partial \alpha_{ka}} \tag{3.8a}$$

subject to

$$\sum_{a=1}^{M} \beta_{ka} = 0 \quad \text{for all } k \varepsilon O \tag{3.8b}$$

$$\beta_{ka} \geq 0 \quad \text{for } \alpha_{ka} = 0 \tag{3.8c}$$

$$\beta_{ka} \leq 0 \quad \text{for } \alpha_{ka} = 1 \tag{3.8d}$$

$$\boldsymbol{\beta} \text{ makes changes in one subset only} \tag{3.8e}$$

$$\boldsymbol{\beta} \neq \mathbf{0} \tag{3.8f}$$

$$\beta_{ka} \text{ unrestricted} \tag{3.8g}$$

Under the constraint (3.8e), the steepest descent direction at $\boldsymbol{\alpha}$ which is an optimal solution of Problem SD2($\boldsymbol{\alpha}$) is

$$\beta^*_{ka}=\begin{cases}\frac{1}{2} & \text{if } k=k^* \text{ and } a=\underset{b\in A\ \exists\ \alpha_{kb}<1}{\operatorname{argmin}}\left\{\frac{\partial\Phi(\alpha)}{\partial\alpha_{kb}}\right\}\\ -\frac{1}{2} & \text{if } k=k^* \text{ and } a=\underset{b\in A\ \exists\ \alpha_{kb}>0}{\operatorname{argmax}}\left\{\frac{\partial\Phi(\alpha)}{\partial\alpha_{kb}}\right\}\\ 0 & \text{otherwise}\end{cases} \tag{3.9a}$$

where

$$r(k)=\underset{a\in A\ \exists\ \alpha_{ka}<1}{\text{minimum}}\left\{\frac{\partial\Phi(\alpha)}{\partial\alpha_{ka}}\right\}-\underset{a\in A\ \exists\ \alpha_{ka}>0}{\text{maximum}}\left\{\frac{\partial\Phi(\alpha)}{\partial\alpha_{ka}}\right\} \tag{3.9b}$$

$$k^*=\underset{k\in O}{\operatorname{argmin}}\{r(k)\} \tag{3.9c}$$

From Corollary II.10, the maximum feasible step size along $\boldsymbol{\beta}^*$, $\theta(\beta^*)$, is $2\alpha_{k^*d}$ where

$$d=\underset{a\in A\ \exists\ \alpha_{k^*a}>0}{\operatorname{argmax}}\left\{\frac{\partial\Phi(\alpha)}{\partial\alpha_{k^*a}}\right\} \tag{3.9d}$$

For a given policy $\boldsymbol{\alpha}$ and feasible direction $\boldsymbol{\beta}$, substituting the new policy vector $\boldsymbol{\rho}=\boldsymbol{\alpha}+\theta\boldsymbol{\beta}$ in objective function of Problem D_2, step size θ appears in the transition matrix $P(\boldsymbol{\alpha}+\theta\boldsymbol{\beta})$ and the cost vector $\mathbf{c}(\boldsymbol{\alpha}+\theta\boldsymbol{\beta})$. Using an argument similar to the proof of Proposition II.5, it can be shown that $f(\theta)=\Phi(\boldsymbol{\alpha}+\theta\boldsymbol{\beta})$ is a polynomial of order at most T.

The solution algorithm for Problem D_2 is given below:

Algorithm III

Step 0) Initialization: Choose an initial policy $\boldsymbol{\alpha}\in\boldsymbol{\Lambda}_2$.

Step 1) Policy Evaluation: Compute the expected total discounted cost $\Phi(\alpha)$.

Step 2) Policy Improvement: Compute the directional derivative $r(k)$ for every subset $k \varepsilon O$. If $r(k^*)$ is positive, then stop, $\Phi(\alpha)$ is a local minimum; if it is equal to zero, then stop, $\Phi(\alpha)$ is either a local minimum or a saddle point; if negative, go to step 3.

Step 3) Compute direction β^* by (3.9a) and the maximum feasible step size using (3.9d).

Step 4) Make a line search on $f(\theta)$ over $\theta \varepsilon (0, \theta(\beta^*)]$. Pick a policy corresponding to some step size θ_0 with $\Phi(\alpha+\theta_0\beta^*)<\Phi(\alpha)$, then take $\alpha+\theta_0\beta$ as the new policy and go to step 1.

Example III.1: The same as Example II.3. Algorithm III for a 10-period planning horizon gives

Table 3.1 Example for Algorithm III

Iteration	Policy (α)	$\Phi(\alpha)$
1	(.5,.5,.5,.5)	12.76
2	(.5,.5,0,1)	10.28
3	(1,0,0,1)	9.84

Note that (1,0,0,1) is the policy that is used in all the periods except the last for nonstationary policy case.

Example III.2 : The same as Example II.2. Algorithm III for a 10-period planning horizon gives

Table 3.2 Example for Algorithm III

Iteration	Policy (α)	$\Phi(\alpha)$
1	(.5,.5,.5,.5)	43.31
2	(.5,.5,.665,.335)	42.07
3	(.5,.5,.6717,.3283)	42.069
4	(.5,.5,.6782,.3218)	42.066

Table 3.2 (Cont'd) Example for Algorithm III

Iteration	Policy (α)	$\Phi(\alpha)$
5	(.5,.5,.68,.32)	42.065
6	(1,0,.68,.32)	42.033

3.2 Infinite Horizon Model

When the length of the planning horizon approaches to infinity, revising the definition of the expected total discounted cost given that the system is initially at state i, v_i, the restricted MDP model with respect to a partition $\mathbf{S}$ is given as

Problem D_3:

$$\Phi(\alpha^*) = \text{Minimum} \sum_{i=1}^{N} p_i v_i \tag{3.10a}$$

subject to

$$v_i = \sum_{a=1}^{M} \alpha_{k(i)a} \left(c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_j \right) \quad \text{for all } i \varepsilon S \tag{3.10b}$$

$$\sum_{a=1}^{M} \alpha_{ka} = 1 \quad \text{for all } k \varepsilon O \tag{3.10c}$$

$$\alpha_{ka} \geq 0 \quad \text{for all } k \varepsilon O,\ a \varepsilon A \tag{3.10d}$$

$$v_i \text{ unrestricted} \quad \text{for all } i \varepsilon S \tag{3.10e}$$

which is the problem studied by Serin(1989). Under the normalization restriction of Algorithm III, the steepest descent direction can be found by inspection of the partial derivatives which take the following form for infinite horizon problem:

$$\frac{\partial \Phi(\alpha)}{\partial \alpha_{ka}} = \sum_{i \in S_k} w_i \left\{ c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a)\, v_j \right\} \tag{3.11}$$

The solution procedure is still Algorithm III. In that case, the line search procedure is time consuming because $f(\theta)$ is an implicit rational function of θ which is difficult to compute. The algorithm is not given here but can be found in Serin(1989).

Example III.3: The same as Example II.3,

Table 3.3 Example for Algorithm III

Iteration	Policy (α)	$\Phi(\alpha)$
1	(.5,.5,.5,.5)	14.29
2	(.5,.5,.2,.8)	12.62
3	(.5,.5,0,1)	11.54
4	(1,0,0,1)	11.04

Note that the final policy (1,0,0,1) is the stationary policy found for finite horizon case.

Example III.4: The same as Example II.2,

Table 3.4 Example for Algorithm III

Iteration	Policy (α)	$\Phi(\alpha)$
1	(0,1,1,0)	51.01
2	(0,1,0.7,0.3)	46.98
3	(1,0,0.7,0.3)	46.97
4	(1,0,0.68,0.32)	46.96
5	(1,0,0.6777,.3223)	46.956

The function $\Phi(\alpha)$ for this problem is plotted against α_{11} and α_{21} to demonstrate the nonlinearity. This plot is given in Figure 3.1.

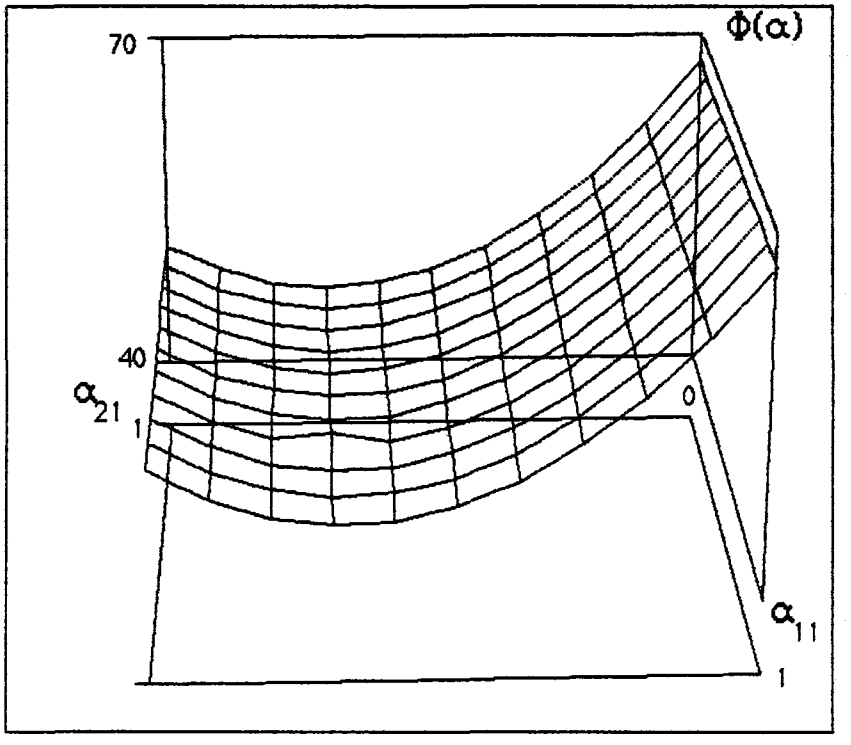

Figure 3.1 Cost Function for Infinite Horizon Case

3.3 Refinement of Partitions and Bounds on Cost Differences

The idea of grouping the states (partitioning) of a MDP can serve two purposes. If the information on the exact state can not be acquired, partitioning is forced by the system. Even if acquisition of the complete state information is physically feasible, partitioning may result in computational saving.

However, in any case the decision maker faces with the problem of how to partition the state set. Then, with respect to a given partition finding an upper bound on the additional cost incurred due to partitioning becomes useful. If, in addition, the decision maker has the information on the cost of detailed observation then combining these, he can make a choice for partitioning. On the other hand, methodology developed in the present study provides means for changing, specifically refining the partition, at any period if an improvement seems possible.

In this section, first we address the refinement concept for infinite horizon models. Then we develop bounds for the difference of the optimal objective values of restricted and unrestricted MDP's. Analogous results can be easily derived for finite horizon problems.

A partition $\mathcal{S}'=\{Q_1, ..., Q_L\}$ of state space is a refinement of the partition $\mathcal{S}=\{S_1, ..., S_K\}$ if for all $l=1, ..., L$ $Q_l \subseteq S_k$ for some $k=1, ..., K$. Note that the partition $\{\{1\}, ..., \{N\}\}$ cannot be refined any further. The feasible policy space with respect to partition $\mathcal{S}$ is a subset of the feasible policy space with respect to its refinement $\mathcal{S}'$. Refining a partition corresponds to relaxing some of the restrictions that force to use the same decision rule over each subset. For that reason, the optimal policy with respect to partition $\mathcal{S}'$ is at least as good as the optimal policy with respect to partition $\mathcal{S}$. On the other hand, the observation process with respect to partition $\mathcal{S}'$ is more detailed, supposedly more costly, than partition $\mathcal{S}$.

At every iteration of the solution algorithm, we can search for refinements that imply improvement in the objective function. However, we should point out that due to the unobservability constraints of the system every refinement may not be feasible in terms of the physical system. Theorem III.1 shows how to search for refinement to obtain an improved policy at an optimal point to a MDP with respect to a given partition.

Theorem III.1: Suppose that $r(k^*) \geq 0$ at the current policy. Let S_k^- be the set of states such that $i \in S_k$ and satisfies the following inequality

$$w_i \left(c_{iu(k)} + \gamma \sum_{j=1}^{N} P_{ij}(u(k))\, v_j - c_{id(k)} - \gamma \sum_{j=1}^{N} P_{ij}(d(k))\, v_j \right) < 0$$

where

$$u(k) = \underset{a \in A \ni \alpha_{ka} < 1}{\operatorname{argmin}} \left\{ \frac{\partial \Phi(\alpha)}{\partial \alpha_{ka}} \right\} \text{ and } d(k) = \underset{a \in A \ni \alpha_{ka} > 0}{\operatorname{argmax}} \left\{ \frac{\partial \Phi(\alpha)}{\partial \alpha_{ka}} \right\}$$

If S_k^- is not empty for some $k \in O$, then there is an improved policy which is feasible with respect to partition $\mathbf{S}'=\{S_1, \ldots, S_{k-1}, S_k', S_k \backslash S_k', S_{k+1}, \ldots, S_K\}$ where S_k' is any subset of S_k^-.

Proof: Suppose S_k^- is not empty. Let S_k' be a subset of S_k^-. Since $r(k^*) \geq 0$, S_k' is a proper subset of S_k and $S_k \backslash S_k'$ is not empty. Then, the observation process Z_t' with respect to partition $\mathbf{S}'$ is defined by

$$
\begin{aligned}
Z_t' &= k' \quad \text{if } X_t \in S_k' \\
Z_t' &= k^- \quad \text{if } X_t \in S_k^- \backslash S_k' \\
Z_t' &= k^+ \quad \text{if } X_t \in S_k \backslash S_k^- \\
Z_t' &= Z_t \quad \text{otherwise}
\end{aligned}
$$

and $O'=\{1, 2, \ldots, k-1, k^-, k', k^+, k+1, \ldots, K\}$. By definition of S_k^-, a direction making changes in partial policy of k', gives a negative directional derivative at the current point.

Corollary III.2: If

$$
w_i \left(c_{i\alpha(k)} + \gamma \sum_{j=1}^{N} P_{ij}(u(k))\, v_j - c_{id(k)} - \gamma \sum_{j=1}^{N} P_{ij}(d(k))\, v_j \right) \geq 0
$$

for all $i \in S_k$, then no improvement in the objective value seems possible with $u(k)$ and $d(k)$, even refining $\mathbf{S}$ to $\{S_1, \ldots, S_{k-1}, \{i_1\}, \{i_2\}, \ldots, \{i_{|S_k|}\}, S_{k+1}, \ldots, S_K\}$, where $i_1, i_2, \ldots, i_{|S_k|} \in S_k$.

Once there is a possibility of improving the objective value by refinement, the next question is how far the current policy is from the unconstrained optimal.

First, we clearly have the following bound,

$$\Phi(\alpha^*) - \Phi(\tau^*) \le \Phi(\alpha) - \frac{1}{1-\gamma} \underset{i\varepsilon S,\, a\varepsilon A}{\text{minimum}} \{c_{ia}\} \tag{3.12}$$

where α^* and τ^* are the optimum policies for the restricted and unrestricted MDP, respectively, and α is a feasible policy for restricted MDP.

Note that if this bound is computed at every iteration of the solution algorithm, a decreasing sequence of bounds can be obtained because the algorithm moves to improved policies whereas second term does not change from one iteration to another.

Now, we present another upper bound on the difference of the optimum expected total discounted cost of MDP restricted with respect to a partition $\mathbf{S}$ and the unrestricted form of the same MDP. The bound is obtained by utilizing a feasible policy to the restricted MDP. For that reason, we can compute the bound at each iteration of Algorithm III. Note that if it is possible to obtain the optimum policy of unrestricted MDP problem with a reasonable effort, then the bound functions as an upper bound on the optimum objective function value of the restricted MDP.

Let α and ρ be feasible policies for restricted MDP and τ and η be feasible policies for unrestricted MDP. In this section, for the conditional cost function $\mathbf{v}$ and the discounted probability $\mathbf{w}$, we will use the corresponding policy as an argument, as they were first introduced in Section 1.2.

Let α be a feasible policy for Problem D_3. α is also feasible for unrestricted form of the same MDP and the equivalent of α for unrestricted MDP is given by τ using the following relation:

$$\tau_{ia} = \alpha_{k(i)a} \quad \text{for all } i\varepsilon S,\ a\varepsilon A \tag{3.13}$$

Proposition III.3: Let α and ρ be two feasible policies for a restricted MDP. Suppose $\rho=\alpha+\theta\beta$. Then,

$$\Phi(\alpha)-\Phi(\rho) = \sum_{i=1}^{N} w_i(\rho)\left(-\theta\sum_{a=1}^{M} \beta_{k(i)a}\left(c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\alpha)\right)\right) \qquad (3.14)$$

Proof: Proof of Proposition III.3 is given in Appendix B.

In obtaining the bound, we need the steepest descent direction for the unrestricted MDP at the point τ. So the Problem SD2(τ), presented in Section 3.1, is solved for unrestricted form of the MDP over all feasible directions replacing (3.8c) and (3.8d) by

$$\beta_{ka} \le 1-\alpha_{ka} \text{ and } \beta_{ka} \ge -\alpha_{ka} \quad \text{for all } k\varepsilon O,\ a\varepsilon A$$

and taking the step size $\theta=1$. By Theorem III.4, we give the bound.

Theorem III.4: Let α^* and τ^* be the optimum policies for restricted and unrestricted problems of the same MDP, respectively. Let α be a feasible policy to restricted infinite horizon problem and τ be the equivalent of α for unrestricted MDP. Suppose Algorithm III moves from α to ρ along direction β and the Howard's algorithm from τ to η along direction β^*, which is obtained by Problem SD2(τ), i.e., $\rho=\alpha+\theta\beta$ and $\eta=\tau+\beta^*$. Then,

$$\Phi(\alpha^*)-\Phi(\tau^*) \le \Phi(\rho)-\Phi(\tau)$$

$$+\frac{1}{1-\gamma}\underset{i\varepsilon S}{\text{maximum}}\left\{-\sum_{a=1}^{M} \beta^*_{ia}\left(c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\tau)\right)\right\} \qquad (3.15)$$

Proof: From (3.14),

$$\Phi(\tau)-\Phi(\eta) = \sum_{i=1}^{N} w_i(\eta)\left(-\sum_{a=1}^{M} \beta^*_{ia}\left(c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\tau)\right)\right) \qquad (3.16a)$$

On the other hand, let $\tau^* = \tau + \varphi$. Then, again from (3.14),

$$\Phi(\tau)-\Phi(\tau^*) = \sum_{i=1}^{N} w_i(\tau^*)\left(-\sum_{a=1}^{M} \varphi_{ia}\left(c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\tau)\right)\right) \quad (3.16b)$$

Recall from Section 2.1 that

$$\sum_{i=1}^{N} w_i(\alpha) = \frac{1}{1-\gamma}$$

for all $\alpha \varepsilon \mathcal{A}_2$. Note that $\mathcal{A}_2 = \mathcal{A}$ when partition $\mathcal{S}$ is defined for unrestricted MDP, i.e., as $S_i = \{i\}$ for all $i \varepsilon S$. As a result, since

$$\sum_{i=1}^{N} (1-\gamma)\, w_i(\tau^*)\left(-\sum_{a=1}^{M} \varphi_{ia}\left(c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\tau)\right)\right)$$

is a convex combination of

$$-\sum_{a=1}^{M} \varphi_{ia}\left(c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)\, v_j(\tau)\right)$$

terms and β^* is the steepest descent direction, i.e.,

$$\sum_{a=1}^{M} \beta^*_{ia}\left(c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\tau)\right) \le \sum_{a=1}^{M} \varphi_{ia}\left(c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\tau)\right) \text{ for all } i \varepsilon S \quad (3.17)$$

$$\Phi(\tau)-\Phi(\tau^*) \le \frac{1}{1-\gamma} \underset{i \varepsilon S}{\text{maximum}}\left\{-\sum_{a=1}^{M} \beta^*_{ia}\left(c_{ia}+\gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\tau)\right)\right\} \quad (3.18)$$

Also, we can use the above reasoning to obtain

$$\Phi(\alpha)-\Phi(\rho) \geq \frac{1}{1-\gamma} \underset{i \in S}{\text{minimum}} \left\{ -\theta \sum_{a=1}^{M} \beta_{k(i)a} \left(c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_j(\alpha) \right) \right\} \quad (3.19)$$

Thus, from (3.18) and (3.19) and τ being equivalent to α,i.e., $\Phi(\alpha) = \Phi(\tau)$, (3.15) is obtained.

Example III.5: The same as Example II.3.

Table 3.5 Example for Bounds

Iteration	$\Phi(\alpha)$	Bounds corresponding to (3.12)	(3.15)
1	16.720	6.72	8.945
4	11.650	1.65	0.488
5	11.04	1.04	0

Example III.6: The same as Example II.3 for infinite planning horizon. Starting with partition $\mathcal{S}$ ={ {1,2,3} } ,

Table 3.6 Example for Refinement

Iteration	$(\alpha_{11}, \alpha_{12})$	$\Phi(\alpha)$
1	(0.5, 0.5)	14.288
2	(0.2, 0.8)	12.964
3	(0 , 1)	12.086

At the last iteration, $S_1^- = \{1\}$ and the upper bound on $\Phi(\alpha^*)-\Phi(\tau^*)$ is zero. Then, continuing with partition $\mathcal{S}$ ={ {1} , {2,3} }

Table 3.7 Example for Refinement

Iteration	$(\alpha_{11}, \alpha_{12}; \alpha_{21}, \alpha_{22})$	$\Phi(\alpha)$
1	(0, 5; 0, 1)	12.086
2	(1, 0; 0, 1)	11.04

CHAPTER IV

CONCLUSION

In this thesis, we considered MDP under unobservability constraints. These constraints state to use the same decision rule for a class of states. So, not the state of the process but only the class it belongs to should be known to take an action. Our objective is to minimize the expected total discounted cost over a finite planning horizon. We consider (i) nonstationary and (ii) stationary policies separately.

We formulate the problems by introducing unobservability constraints to the linear model of a MDP. The models can be constructed to have a nonlinear objective function to be minimized over a set of linear constraints defining the policy space. The algorithm that we propose to solve these problems, is in fact the method of feasible directions described in Bazaraa and Shetty (1979). It is also similar to the Howard's policy iteration algorithm developed for unrestricted MDP's. The algorithm is initialized by a feasible policy, the corresponding discounted cost is computed at the policy evaluation step, and it iterates to an improved policy if a descent direction can be found. At every iteration, improvement is guaranteed but the global optimality of the termination point is not due to the nonlinearity of the model. Algorithm terminates at a point satisfying necessary Kuhn-Tucker conditions, i.e., at a local optimum or a saddle point.

We check the validity of this solution method by solving some problems using the nonlinear programming software MINOS and using proposed algorithm (Example II.2). In addition, we solved

several textbook problems using the algorithm we propose with $\mathbb{S} = \{S_1, ..., S_N\}$ where $S_i = \{i\}$ for all $i \in S$. We end up with the same results.

For restricted MDP problems with nonstationary policies, we observed that there exists a deterministic global optimal policy, whereas the optimal stationary policy could be randomized. Due to the nature of the MDP's, randomized policies create implementation problems. The same problem is encountered by Serin (1989) for infinite horizon model of restricted MDP. The models developed in this study are all multilinear programming problems as explained in Section 2.1. We can deepen our studies on multilinear programming to improve the results about the structure of the problems. On the basis of problem (i) our study can be extended to rolling horizon procedures, where a T-period problem is solved at every period and the decision for the first period is employed. Shapiro(1968) and Hopp et al. (1987,1989) studied on the asymptotic behavior of the optimal policy for the discounted case of homogeneous and nonhomogeneous MDP, respectively. In both of the studies, observation is that convergence of optimal policy results from convergence of the expected total discounted cost to the optimal infinite horizon cost for every initial state, i.e., $v_{iT}(\alpha^*(T))$ converges to $v_i(\alpha^*)$, where α^* is the optimal stationary policy for infinite horizon problem, which could not be shown in the present case.

Ross (1989a) studied MDP models under k linear cost constraints. The optimal policy makes randomization in at most k states. It is not possible to obtain similar structure in the present case. The effects of different partitioning may be significant also in terms of identifying the structure, like the number of randomizations, of the optimal solution. Since the partition { {1}, {2}, ..., {N} } leads to a deterministic optimal solution, grouping the states gradually, e.g., starting with { {1,2}, {3}, ..., {N} } partitioning effect can be analyzed. There is the problem of implementation of a randomized policy. We are aware of that there is lack of comparison of numerical results of the current approach and POMDP approach. The main reason is the insufficient number of numerical results in POMDP literature.

The proposed solution procedure seems to be an extension of Howard's policy iteration algorithm to partitioned MDP. The linearity of the expected discounted cost function guarantees convergence to the global optimal solution for unrestricted MDP. In policy improvement step, we check the existence of a better policy by computing the steepest descent direction. This check is performed on the basis of comparing the partial derivatives of the expected discounted cost function which is a weighted sum

$$\sum_{i \in S_k} w_i \left(c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a)\, v_j \right),$$

of Howard's test quantities

$$c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a)\, v_j .$$

The weight w_i is the discounted probability of being in state i and computed using the initial probability distribution. Considering independence of the unrestricted optimal policy from initial distribution, this observation may seem to be counter intuitive, but it is in fact directly related with not satisfying the Bellman's optimality equation. For finite horizon case specifically, it is the reason for not being able to use dynamic programming as a solution procedure. Moreover, the policy obtained by using POMDP method also depend on the initial distribution.

We introduce the concept of refinement and develop bounds on the optimal cost value that could be evaluated at every iteration of the algorithm. For our sample problems, the bounds do not perform well, implying a further study on obtaining stronger bounds. As in all aggregation/disaggregation procedures of Mendelssohn(1983), Birge (1985a, 1985b), there is the problem of answering the question "how to partition?", if a natural one is not imposed by the system or when an

aggregation that would dominate the natural one is preferred. In this respect, refinement discussion and strong bounds may be very useful. The following rough bound, for instance, gives the principal that the closer the cost figures of states in a subset are the closer the optimal cost value, $\Phi(\alpha^*)$, to the unconstrained cost, $\Phi(\tau^*)$,

$$0 \le \Phi(\alpha^*) - \Phi(\tau^*) \le \frac{1}{1-\gamma}\left(\sum_{k=1}^{K}\sum_{i\in S_k} \underset{i\in S_k,\, a\in A}{\text{maximum}}\{c_{ia}\} - \sum_{i\in S} \underset{i\in S,\, a\in A}{\text{minimum}}\{c_{ia}\}\right)$$

Note that the upper bound decreases as the starting partition is refined more.

It is possible to bound the number of changes from a policy to the next by changing the normalization constraint on the feasible directions, e.g., allowing changes in one period only. As we allow less number of changes, the algorithm performs slower in terms of the number of iterations. On the other hand, as the number of changes increase, the line search step becomes harder if an exact minimization is carried out because the degree of the polynomial to be minimized is the number of changes. Study on a large scale example may provide opportunity for better comparison of these directions.

Our primary concern is on obtaining an optimal policy not on the performance of algorithm in terms of CPU time or number of iterations. In these respects, study on a large scale MDP may be useful, and can help to improve the algorithm.

REFERENCES

Agafanov, G. and Makarova, A., 1976. "An Algorithm for Iterative Aggregation of Economic Hierarchy Systems as an Instrument for Consistency of Solutions of Multilevel Systems", Optimization Methods and Operations Research, pp. 132-148.

Albright, S.C.,1979. "Structural Results for Partially Observable Markov Decision Processes", Operations Research, Vol.27, pp. 1041-1053.

Bazaraa, M.S., and Shetty C.M.,1979. Nonlinear Programming, John Wiley and Sons, Inc., U.S.A..

Bean, J.C. et al, 1987. "Aggregation in Dynamic Programming", Operations Research, Vol.35, No. 2, pp. 215-220.

Bertsekas, D.P., 1976. Dynamic Programming and Stochastic Control, Academic Press, Inc., New York.

Birge, J. R., 1985a. "Aggregation Bounds in Stochastic Linear Programming", Mathematical Programming, Vol.31, pp. 25-41.

Birge, J. R., 1985b. "Decomposition and Partitioning Methods for Multistage Stochastic Linear Programs", Operations Research, Vol.33, No. 5, pp. 989-1007.

Blackwell, D., 1965. "Discounted Dynamic Programming", Annals of Mathematical Statistics, Vol.36, pp. 226-235.

Borkar, V.S., 1991. "A Remark on Control of Partially Observed Markov Chains", Annals of Operations Research, Vol.29, pp. 429-438.

Chelsea, C.W. et al.,1985. "Reward Revision for Discounted Markov Decision Problems", Operations Research, Vol.33, pp. 875-883.

Denardo, E.V. et al.,1970. "On Linear Programming in a Markov Decision Problem", Management Science, Vol.16, No. 5, pp. 282-288.

Denardo, E.V., 1982. Dynamic Programming Models and Applications, Prentice Hall, Inc., New Jersey.

Derman, C., 1970. Finite State Markov Decision Processes, Academic Press, Inc., New York.

Drenick, R.F., 1992. "Multilinear Programming: Duality Theories", Journal of Optimization Theory and Applications, Vol.72, No. 3,pp. 459-486.

Eckles, J.E., 1968. "Optimum Maintenance with Incomplete Information", Operations Research, Vol.16, pp. 1058-1067.

Fernandez-Gaucherand, E. et al.,1991. "On the Average Cost Optimality Equation and the Structure of Optimal Policies for Partially Observable Markov Decision Processes", Annals of Operations Research, Vol.29, pp. 439-470.

Hastings, N. A. J. , 1970. "Bounds on the Gain of a Markov Decision Process", Technical Notes, University of Birmingham, Birmingham, pp. 240-244.

Hillier, F.S., and Lieberman G.J., 1974. Operations Research, Holden-Day, Inc., San Francisco.

Hopp, W.J. et al.,1987. "A New Optimality Criterion for Nonhomogeneous Markov Decision Processes", Operations Research, Vol.35, No. 6, pp. 875-883.

Hopp, W.J., 1989. "Identifying Forecast Horizons in Nonhomogeneous Markov Decision Processes", Operations Research, Vol.37, No. 2, pp. 339-343.

Hordijk, A. and Kallenberg L.C.M.,1979. "Linear Programming and Markov Decision Chains", Management Science, Vol.25, No. 4, pp. 352-362.

Howard, R.A., 1966. Dynamic Programming and Markov Processes, The Massachusetts Institute of Technology, U.S.A..

Howard, R.A., 1971a. Dynamic Probabilistic Systems : Markov Models, Academic Press, U.S.A., Inc., Vol. I.

Howard, R.A., 1971b. Dynamic Probabilistic Systems : Semimarkov and Decision Processes, Academic Press, U.S.A., Inc., Vol. II.

Hughes, J., 1977. "Optimal Internal Audit Timing", Accounting Rev., Vol.LII, pp. 56-58.

Kaplan, R., 1969. "Optimal Investigation Strategies with Imperfect Information", J. Accounting Res., Vol.7, pp. 32-43.

Karush, W. and Dear, R., 1967. "Optimal Strategy for Item Presentation in Learning Models", Management Science, Vol.13, pp. 773-785.

Lane, E.D., 1989. "A Partially Observable Model of Decision Making by Fishermen", Operations Research, Vol.37, No. 2, pp. 240-254.

Lovejoy, W.S., 1986. "Policy Bounds for Markov Decision Processes", Operations Research, Vol.34, No. 4, pp. 630-637.

Lovejoy, W.S., 1987a. "Some Monotonicity Results for Partially Observed Markov Decision Processes", Operations Research, Vol.35, No. 5, pp. 736-743.

Lovejoy, W.S., 1987b. "On the Convexity of Policy Regions in Partially Observed Systems", Operations Research, Vol.35, No. 4, pp. 619-621.

Lovejoy, W.S., 1991. "Computationally Feasible Bounds for Partially Observed Markov Decision Processes", Operations Research, Vol.39, No. 1, pp. 162-175.

Mamer, J.W., 1986. "Successive Approximations for Finite Horizon, Semi-Markov Decision Processes With Application to Asset Liquidation", Operations Research, Vol.34, No. 4, pp. 638-644.

Mendelssohn, R., 1980. "Improved Bounds for Aggregated Linear Programs", Operations Research, Vol.28, No. 6, pp. 1450-1453.

Mendelssohn, R., 1983. "An Iterative Aggregation Procedure for Markov Decision Processes", Operations Research, Vol.31, No. 1, pp. 62-73.

Mine, H., and Osaki S., 1970. Markovian Decision Processes, American Elsevier Publishing Company, Inc., New York.

Monahan, G.E., 1980. "Optimal Stopping in a Partialy Observable Markov Process with Costly Information", Operations Research, Vol.28, No. 6, pp. 1319-1333.

Monahan, G.E., 1982. "A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms", Naval Research Logistics Quarterly, Vol.29, No. 1, pp. 1-16.

Morton, T.E. and Wecker W.E., 1977. "Discounting, Ergodicity and Convergence for Markov Decision Processes, Management Science, Vol.23, No. 8, pp. 890-900.

Morton, T.E., 1978. "The Nonstationary Infinite Horizon Inventory Problem", Management Science, Vol.24, No. 14, pp. 1474-1482.

Morton, T.E., 1979. "Infinite-Horizon Dynamic Programming Models: A Planning-Horizon Formulation", Operations Research, Vol.27, No. 4, pp. 730-742.

Popyack, J.L., and White, C.C., 1985. "Suboptimal Policy Determination for Large-Scale Markov Decision Processes, Part 2: Implementation and Numerical Evaluation", Journal of Optimization Theory and Applications, Vol.46, No. 3, pp. 343-358.

Putterman, M.L., and Brumelle, S.L., 1979. "On the Convergence of Policy Iteration in Stationary Dynamic Programming", Mathematics of Operations Research, Vol.4, No. 1, pp. 60-69.

Rosenfield, D., 1976a. "Markovian Deterioration with Uncertain Information", Operations Research, Vol.24, pp. 141-155.

Rosenfield, D., 1976b. "Markovian Deterioration with Uncertain Information-A More General Model", Naval Research Logistics Quarterly, Vol.23, pp. 389-406.

Ross, K.W., 1989a. "Randomized and Past-Dependent Policies for Markov Decision Processes with Multiple Constraints", Operations Research, Vol.37, No. 3, pp. 474-477.

Ross, K.W., and Varadarajan, R., 1989b. "Markov Decision Processes with Sample Path Constraints: The Communicating Case", Operations Research, Vol.37, No. 5, pp. 780-790.

Ross, S., 1970. Applied Probability Models with Optimization Applications, Holden-Day, San Francisco. Calif.

Ross, S., 1971. "Quality Control Under Markovian Deterioration", Management Science, Vol.17, pp. 587-596.

Ross, S., 1983. Introduction to Stochastic Dynamic Programming, Academic Press, Inc., New York.

Runggaldier, W.J., 1991. "On the Construction of ε-Optimal Strategies in Partially Observed MDPs", Annals of Operations Research, Vol.28, pp. 81-96.

Sawaki, K., 1983. "Transformation of Partially Observable Markov Decision Processes into Piecewise Linear Ones", Journal of Mathematical Analysis and Applications, Vol.91, pp. 112-118.

Serin, Y.Y., 1989. "Implementable Policies for Markov Decision Processes", Ph.D dissertation, The University of North Carolina at Chapel Hill.

Sernik, E.L., and Marcus, S.I., 1991. "On the Computation of the Optimal Cost Function for Discrete Time Markov Models with Partial Observations", Annals of Operations Research, Vol.29, pp. 471-512.

Shapiro, J.F., 1968. "Turnpike Planning Horizons for a Markovian Decision Model", Management Science, Vol.14, No. 5, pp. 292-300.

Sheskin, T.J., 1987. "Successive Approximations in Value Determination for a Markov Decision Process", Operations Research, Vol.35, No. 5, pp. 784-786.

Smallwood, R.D., 1971a. "The Analysis of Economic Teaching Strategies for a Simple Learning Model", J. Math. Psych, Vol.8, pp.285-301.

Smallwood, R.D., et al, 1971b. "Toward an Integrated Methodology for the Analysis of Health-Care Systems", Operations Research, Vol.19, pp. 1300-1322.

Smallwood, R.D., and Sondik, E.J., 1973. "The Optimal Control of Partially Observable Markov Decision Processes over a Finite Horizon", Operations Research, Vol.21, pp. 1071-1088.

Smith, J. L., 1967. "Markov Decisions on a Partitioned State Space and the Control of Multiprogramming", Unpublished Ph. D. Dissertation, University of Michigan.

Smith, J. L., 1971. "Markov Decisions on a Partitioned State Space", IEEE Transactions On Systems, Man and Cybernetics, SMC, pp. 55-60.

Sondik, E.J., 1978. "The Optimal Control of Partially Observable Markov Processes over the Infinite Horizon: Discounted Costs", Operations Research, Vol.26, No. 2, pp. 282-304.

Vakhutinskii, I. and Dudkin, L., 1973. "Algorithm of Iterative Aggregation for the Solution of the Problem of Linear Programming of a General Nature", Izvestiia of the Siberian Section of the Academy of Sciences of the USSR Social Science Series, pp. 67-71.

Veugen, L.M.M. et al.,1985. "Aggregation and Disaggregation in Markov Decision Models for Inventory Control", European Journal of Operational Research, Vol.20, pp. 248-254.

Wang, R.C., 1977. "Optimal Replacement Policy with Unobservable States", Journal of Applied Probability, Vol.14, pp. 340-348.

White, D.J.,1978. Finite Dynamic Programming, John Wiley and Sons, Inc., Malta.

Zipkin, P., 1977. "Aggregation in Linear Programming", Ph. D. dissertation, Yale University, New Haven, Conn., pp.189.

Zipkin, P., 1980a. "Bounds on the Effect of Aggregating Variables in Linear Programs", Operations Research, Vol.28, pp. 403-418.

Zipkin, P., 1980b. "Bounds for Row-Aggregation in Linear Programming", Operations Research, Vol.28, pp. 903-916.

APPENDICES

APPENDIX A

PROOF OF LEMMA II.1

a) If the matrix $B(\alpha)$ is partitioned as

$$\begin{bmatrix} I & -\gamma P(\alpha, T) & 0 & \cdots & 0 & 0 \\ 0 & I & -\gamma P(\alpha, T-1) & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & I & -\gamma P(\alpha, 2) \\ 0 & 0 & 0 & \cdots & 0 & I \end{bmatrix} \tag{A.1}$$

I is the identity matrix of dimension N. It can be easily seen that the rows of $B(\alpha)$ are independent, meaning that $B(\alpha)$ is nonsingular.

b) Recall the partitioned form of $B(\alpha)$ given by (A.1), noting that for square matrices A, C, D where D is invertible, A is also invertible.

$$A=\begin{bmatrix} I & C \\ 0 & D \end{bmatrix} \qquad A^{-1}=\begin{bmatrix} I & -CD^{-1} \\ 0 & D^{-1} \end{bmatrix} \tag{A.2}$$

Then, if we partition $B(\alpha)$ as matrix A, we obtain square matrices, which are known to be nonsingular by the reasoning stated in part a of Lemma II.1, in the south-east corner of $B(\alpha)$ as matrix D; and the result follows. The partitioned form of $B(\alpha)^{-1}$ is given below:

$$\begin{bmatrix} I & \gamma P(\alpha, T) & \dots & \dots & \dots & \gamma^{(T-1)} P(\alpha, T)..P(\alpha, 2) \\ 0 & I & \gamma P(\alpha, T-1) & \dots & \dots & \gamma^{(T-2)} P(\alpha, T-1)..P(\alpha, 2) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \gamma P(\alpha, 3) & \gamma^2 P(\alpha, 3)P(\alpha, 2) \\ 0 & 0 & \dots & \dots & I & \gamma P(\alpha, 2) \\ 0 & 0 & \dots & \dots & 0 & I \end{bmatrix} \tag{A.3}$$

Note that all entries of $B(\alpha)^{-1}$ are nonnegative for all i, $j \in S$ and all t, n=1, ..., T, since $0 < \gamma < 1$ and P's are transition matrices and I's are identity matrices of dimension N.

APPENDIX B

PROOF OF PROPOSITION III.3

We will obtain (3.14) by analyzing the difference between $v_i(\alpha)$ and $v_i(\rho)$'s. In general,

$$v_i(\alpha) = \sum_{a=1}^{M} \alpha_{k(i)a}\left(c_{ia} + \gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\alpha)\right) \tag{B.1a}$$

$$v_i(\rho) = \sum_{a=1}^{M} \rho_{k(i)a}\left(c_{ia} + \gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\rho)\right) \tag{B.1b}$$

for all $i \varepsilon S$. Since $\rho_{ka} = \alpha_{ka} + \theta\beta_{ka}$,

$$v_i(\alpha) - v_i(\rho) = \gamma\sum_{j=1}^{N} P_{ij}(\alpha)\left(v_j(\alpha) - v_j(\rho)\right)$$

$$- \theta\sum_{a=1}^{M} \beta_{k(i)a}\left(c_{ia} + \gamma\sum_{j=1}^{N} P_{ij}(a)v_j(\rho)\right)$$

$$= \gamma\sum_{j=1}^{N} P_{ij}(\alpha)\left(v_j(\alpha) - v_j(\rho)\right)$$

$$+ \theta\sum_{a=1}^{M} \beta_{k(i)a}\,\gamma\sum_{j=1}^{N} P_{ij}(a)\left(v_j(\alpha) - v_j(\rho)\right)$$

$$- \theta \sum_{a=1}^{M} \beta_{k(i)a} \left(c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_j(\alpha) \right) \qquad (B.2)$$

Using

$$\gamma \sum_{j=1}^{N} P_{ij}(\alpha) + \theta \sum_{a=1}^{M} \beta_{k(i)a} \gamma \sum_{j=1}^{N} P_{ij}(a) = \gamma \sum_{j=1}^{N} P_{ij}(\rho),$$

$$v_i(\alpha) - v_i(\rho) = \gamma \sum_{j=1}^{N} P_{ij}(\rho) \left(v_j(\alpha) - v_j(\rho) \right)$$

$$- \theta \sum_{a=1}^{M} \beta_{k(i)a} \left(c_{ia} + \gamma \sum_{j=1}^{N} P_{ij}(a) v_j(\alpha) \right) \qquad (B.3)$$

for all $i \in S$. Then,

$$v_i(\alpha) - v_i(\rho) = \sum_{j=1}^{N} \left(I - \gamma P(\rho) \right)^{-1}_{ij} \left(-\theta \sum_{a=1}^{M} \beta_{k(j)a} \left(c_{ja} + \gamma \sum_{m=1}^{N} P_{jm}(a)\, v_m(\alpha) \right) \right) \qquad (B.4)$$

Multiplying both sides by the probability of being in state i initially and taking summation over all states, we obtain the difference of objective function values for the two policies.

$$\Phi(\alpha) - \Phi(\rho) = \sum_{i=1}^{N} p_i \sum_{j=1}^{N} B(\rho)^{-1}_{ij} \left(-\theta \sum_{a=1}^{M} \beta_{k(j)a} \left(c_{ia} + \gamma \sum_{m=1}^{N} P_{jm}(a) v_m(\alpha) \right) \right) \qquad (B.5)$$

where

$$\sum_{i=1}^{N} p_i\, B(\rho)^{-1}_{ij} = w_j(\rho) \qquad (B.6)$$

and the proof is complete.