Markov Decision Problem Based On A Two-Level System

by

Huang Jin

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Master of Philosophy
in the Department of Electrical and Electronic Engineering

Aug.2003, Hong Kong
Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Huang Jin

Huang Jin
Markov Decision Problem Based On A Two-Level System

by

Huang Jin

This is to certify that I have examined the above MPhil thesis and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the thesis examination committee have been made.

Prof. Xi-Ren Cao, Thesis Supervisor

Prof. Li-Xin Wang, Thesis Committee Chairman

Prof. K. B LETAIEF, Head of Department

Department of Electrical and Electronic Engineering

Aug, 2003,
Acknowledgements

First I would like to give my thanks to my supervisor Prof. Xi-Ren Cao for his guidance during my studies in Hong Kong University of Science and Technology. I am impressed by his serious attitude and academic enthusiasms in research. Next I would like to thank Dr. Yat-Wah Wan in the department of IEEM for his guidance and encouragement. The discussion with Prof. Cao and Dr. Wan give me much help in finishing this research. And I would like to thank my parents for their support. Also I thank my friends in HKUST.
# TABLE OF CONTENTS

Title Page .................................................................................................................. i
Authorization Page ..................................................................................................... ii
Signature Page ............................................................................................................. iii
Acknowledgements .................................................................................................... iv
Table of Contents ....................................................................................................... v
Abstract ....................................................................................................................... vi

1 Fundamentals............................................................................................................. 1
   1.1 Markov Processes ............................................................................................... 1
   1.2 Markov Decision Processes .............................................................................. 2
   1.3 Performance Potential ...................................................................................... 4
   1.4 Policy Iteration ................................................................................................. 6

2 Markov decision Problem Based On a Two-level System ..................... 10
   2.1 Introduction ...................................................................................................... 10
   2.2 The Two-level Model ...................................................................................... 10
   2.3 The Analysis of the Two-Level System ......................................................... 12
   2.4 The Two-Level Decision Problem ................................................................. 20
   2.5 Conclusion ....................................................................................................... 32
Markov Decision Problem Based On A Two-Level System

by Huang Jin

Department of Electrical and Electronic Engineering

The Hong Kong University of Science and Technology

Abstract

Many problems in discrete event dynamic systems (DEDS) can be modeled as Markov Decision Processes. However, the computation of a MDP problem increases rapidly with the increase of the state space. In this thesis, we consider the Markov decision problem based on a two-level system. Although the two-level system may have many states in its state space, we can give an efficient policy iteration algorithm to find the optimal policy.
Chapter 1

Fundamentals

1.1 Markov Processes

Markov process is a very important stochastic process in discrete event dynamic system (DEDS). DEDS is introduced in detail in ([1],[2]),

Markov model is highly useful in a wide variety of practical systems such as communication networks and manufacturing systems. A random process is called a Markov Process if the conditional probability of future values of given past values depends only upon the most recent value ([4]). Let $S$ be a countable state space of a stochastic process $X(t)$, $X(t)$ is a Markov process ([3]) if for any $t, s \geq 0$ and $j \in S$,

$$P[X(t + s) < j|X(u); u \leq t] = P[X(t + s) < j|X(t)]$$ \hspace{1cm} (1.1)

In this thesis we only discuss discrete time Markov chain. For a discrete time Markov chain, the state space is a discrete set and state transitions are constrained to occur at time instants $0, 1, \ldots, k, \ldots$. A discrete time Markov chain satisfy

$$P(X_{n+1} = j|X_m; m \leq n) = P(X_{n+1} = j|X_n)$$ \hspace{1cm} (1.2)

Markov chains with time-homogeneous transition probabilities means that for any $i, j \in S$

$$P(X_{n+1} = j|X_n = i) = p_{ij}$$
That is \( P(X_{n+1} = j | X_n = i) \) is independent of the time parameter \( n \). The probabilities \( p_{ij} \) is referred to as the one-step transition probabilities, and

\[
\sum_{j \in S} p_{ij} = 1
\]

The matrix \( P = \{ p_{ij} \} \) is called state transition probability matrix.

For some Markov chains, the state probability vector approaches a constant vector independent of the initial probability vector, i.e., as \( n \to \infty \)

\[
p_{ij}^{(n)} = P(X_n = j | X_0 = i) \to \pi_j
\]

The vector \( \pi = (\pi_1, \pi_2, \ldots, \) \( ) \) is called the vector of steady state probabilities. And we have

\[
\pi P = \pi, \quad \text{and} \quad \pi e = 1
\]

where \( e \) is a vector with all components are 1.

1.2 Markov Decision Processes

Markov decision processes ([5],[6],[7]) is a Markov processes whose state transitions can be controlled by taking a sequence of actions. MDPs have attracted the attention of many researchers because they are important both from the practical and the intellectual points of view. MDPs provide tools for the solution of important real-life problems. In particular, many business and engineering applications use MDP models. Analysis of various problems arising in MDPs leads to a large variety of interesting mathematical and computational problems.

We assume that the Markov chain, denoted as \( X = \{ X_n, n > 0 \} \) has a finite state space \( S = \{ 1, 2, \ldots, M \} \). At any state \( i \in S \), an action \( \alpha \) is taken from an action space \( A(i) \) and is applied to the Markov chain. As a result of taking action \( \alpha \) at state \( i \), a cost \( f(i, \alpha) \) is incurred and the system state at the next decision epoch is stochastically determined by the transition probability \( p^\alpha(i, j) \) for \( j = 1, 2, \ldots, M \). All the actions collectively compose a policy denoted as \( L \). A policy is called stationary if the action depends only on the state at which the system stays, that is, the action doesn’t depend on time.

Our objective is to determine a policy which optimizes a predetermined performance criterion \( \eta \). Generally, there are several frameworks of cost criteria considered in Markov decision problems.
In a finite-horizon Markov chain, decision-makers are only interested in the behavior of the controlled Markov chain over a finite decision epochs 0, 1, \ldots, N. Accordingly, the total expected cost over a finite horizon under policy $\mathcal{L}$ can be written as

$$\eta^\mathcal{L}(X_0) = E\left\{ \sum_{n=0}^{N-1} f[X_n, \mathcal{L}(X_n)] \right\}$$

where "$E$" denotes the expectation. Note that the performance defined in this way depends on the initial state of the chain $X_0$.

On the other hand, in an infinite-horizon Markov chain, a infinitely long system trajectory is observed and some different performance criteria are defined ([7]).

1. **Total expected discounted cost over an infinite horizon.** In this case, a concept "discount" is introduced based on the economic interpretation. We define a discount factor $\beta \in [0, 1]$, The cost under policy $\mathcal{L}$ with an initial state $X_0$ can be written as

$$\eta^\mathcal{L}(X_0) = E\left\{ \sum_{n=0}^{\infty} \beta^n f[X_n, \mathcal{L}(X_n)] \right\}$$

This criterion, however, only makes sense if it can yield finite values of at least some policies. As a special case, $\beta = 1$ implies a total expected undiscounted cost, where same concerns are put on the costs incurred in future and at present.

2. **Average expected cost over an infinite horizon.** Sometimes the decision-makers are interested in the average cost during system evolution instead. Specifically, the question addressed is "How much does it cost to operate the given system per unit of time?" Accordingly, the long run average cost is defined as

$$\eta^\mathcal{L}(X_0) = \lim_{N \to \infty} \frac{1}{N} E\left\{ \sum_{n=0}^{N-1} f[X_n, \mathcal{L}(X_n)] \right\}$$

Note that the above limit exists and does not depend on the initial state for ergodic markov chains.

To solve the optimal policy for a MDP problem, we can use linear programming([8],[9]), value iteration ([16]) or policy iteration ([14]). And we will introduce policy iteration in the following part.
1.3 Performance Potential

In ([13],[14]), performance potential theory is intensively explored and is applied to the infinite-horizon average-cost MDP. The term *performance potential* reflects the physical meaning of the concept, which leads to some important properties; that allow it be measured on a single sample path. The estimation of the potentials; can then be integrated into policy iteration to eventually achieve the optima policy, which yields the minimal average cost.

In this section, we will introduce the performance potential theory. We assume that the Markov chain, denoted as $X = \{X_n, n \geq 0\}$ be a ergodic time-homogenous Markov chain with a finite state space $S = \{1, 2, \ldots, M\}$. At any state $i \in S$, an action $\alpha$ is taken from an action space $A(i)$ and is applied to the Markov chain. We assume that the number of actions is finite, and we only consider stationary policies. A stationary policy is a mapping $\mathcal{L} : S \rightarrow A$, which determines the underlying transition probability matrix $P^\mathcal{L} = [P^{\mathcal{L}(i)}(i, j)]_{M \times M}$. Let $\mathcal{E}$ be the policy space, for simplicity, we assume the Markov chain under any policy $\mathcal{L} \in \mathcal{E}$ is ergodic.

Accordingly, the steady state probabilities corresponding to policy $\mathcal{L}$ is denoted as a vector $\pi^\mathcal{L} = [\pi^\mathcal{L}(1), \pi^\mathcal{L}(2), \ldots, \pi^\mathcal{L}(M)]$. Suppose that at each stage with state $i$ and control action $\alpha \in A(i)$, a cost $f(i, \alpha)$ is incurred, and we define vector

$$f^\mathcal{L} = [f^\mathcal{L}(1, \mathcal{L}(1)), f^\mathcal{L}(2, \mathcal{L}(2)), \ldots, f^\mathcal{L}(M, \mathcal{L}(M))]^T$$

The long term expected value of the average cost corresponding to policy $\mathcal{L}$ is then

$$\eta^\mathcal{L} = \lim_{N \to \infty} \frac{1}{N} E \left\{ \sum_{n=0}^{N-1} f[X_n, \mathcal{L}(X_n)] \right\}$$

For ergodic chains, it follows

$$\eta^\mathcal{L} = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} f[X_n, \mathcal{L}(X_n)]$$

The objective is to minimize this average cost per stage over the policy space $\mathcal{E}$.

Formally, the performance potential $g$ is defined by *Poisson equation*.

$$(I - P + e\pi)g = f$$

where $I$ is the identity matrix and the $M$-dimensional vector $e = (1, \ldots, 1)^T$. The matrix $(I - P + e\pi)$ is called fundamental matrix [15]. It is proved in [15] that, the fundamental
matrix is invertible. Hence we have

$$g = (I - P + e\pi)^{-1} f$$  \hspace{1cm} (1.3)$$

We will see that only the difference between two potentials is important, i.e., we can replace $g$ with $g + ce$, where $c$ is any constant, in all the analysis using potentials.

In principle, Eqn. (1.3) can be used to calculate the potential vector $g$. However in practical system, some elements in the transition probability matrix $P$ may be unknown or time-varying, thereby making the computation difficult to implemented, if not impossible. For example, in a simple Markov queueing system, the transition probabilities are related to the arrival rate of the job or the service rate of the server. These parameters, however, may be unknown. Hence, the transition probability matrix may be partially unavailable.

Performance potentials can be estimated in a sample path without the full knowledge of system parameters ([10]-[14]), thereby removing the need to identify the system. It can be shown that the eigenvalues of $(P - e\pi)$ lie within the unit circle. Therefore,

$$(I - P + e\pi)^{-1} = \sum_{k=0}^{\infty} (P - e\pi)^k = \sum_{k=0}^{\infty} (P^k - e\pi)$$  \hspace{1cm} (1.4)$$

Combining (1.3) and (1.4), we have [13]

$$g(i) = \lim_{N \to \infty} \{ E[\sum_{n=0}^{N-1} f(X_n)|X_0 = i] - N\eta \}$$  \hspace{1cm} (1.5)$$

Furthermore, let

$$d(i, j) = g(j) - g(i), i, j = 1, 2, \ldots, M$$  \hspace{1cm} (1.6)$$

and

$$D = eg^T - ge^T$$

In ([13]), $d(i, j), i, j \in S$ are called realization factors and $D$ is called realization matrix.

From (1.5) and (1.6), we have

$$d(i, j) = \lim_{N \to \infty} E\{\sum_{n=0}^{N-1} f(X_n)|X_0 = j\} - \lim_{N \to \infty} E\{\sum_{n=0}^{N-1} f(X_n)|X_0 = i\}$$  \hspace{1cm} (1.7)$$

Both (1.5) and (1.7) relate $d(i, j)$ to the sample path of a Markov chain, but they require infinite stages.
Consider a Markov chain \( X = \{X_n, n \geq 0\} \) starting with \( X_0 = i \). Let \( L_i(j) = \min\{n : n \geq 0, X_n = j\} \), i.e., at \( n = L_i(j) \), the Markov chain reaches state \( j \) for the first time. It is proved in ([14]) that

\[
d(j, i) = \lim_{N \to \infty} E\left\{ \sum_{n=0}^{L_i(j)-1} [f(X_n) - \eta] | X_0 = i \right\}
\]  

(1.8)

Eqn. (1.8) suggests another method to estimate the performance potential in an finite sample path version. It requires no knowledge about the system structure (such as transition probability matrix) as long as the sample path is observable. This characteristic essentially converts the computation-based methodology to a sample-path-based approach.

To obtain the physical interpretation of the concept of performance potentials and realization factors, let us consider two independent chains, \( X \) and \( X' \), with the same state space \( S \) and transition probability matrix \( P \), and different initial states \( X_0 = i, X'_0 = j \). Define \( N_{ij} = \min\{n : n \geq 0, X_n = X'_n\} \), i.e., at \( n = N_{ij} \), the two chains merge for the first time. Then [13]

\[
d(j, i) = \lim_{N \to \infty} E\left\{ \sum_{n=0}^{N_{ij}} [f(X_n) - f(X'_n)] | X_0 = i, X'_0 = j \right\}
\]

Therefore, \( d(i, j) \) measures the difference between the performance of a Markov chain starting from the state \( j \) and that starting from \( i \). That is, \( d(i, j) \) measures the average effect of a perturbation of the state from \( i \) to \( j \) on the long term performance.

1.4 Policy Iteration

The performance potential plays an important role in optimizing the infinite-horizon, average-cost MDP. After it is derived by computation (1.3) or online estimation (1.5,1.8), it can be applied in policy iteration to achieve the optimal policy. It is easy to verify that

\[
\eta' - \eta = \pi'[(P'g + f') - (Pg + f)]
\]

Then we have the following optimal theorem [14].

**Proposition 1.4.1** A policy \( \mathcal{L} \) is optimal if and only if

\[
p^{\mathcal{L}} g^{\mathcal{L}} + f^{\mathcal{L}} \leq p^{\mathcal{L}'} g^{\mathcal{L}'} + f^{\mathcal{L}'}
\]

(1.9)

for all \( \mathcal{L}' \in \mathcal{E}, \mathcal{L}' \neq \mathcal{L} \).
Note that the optimal policy can be determined by checking (1.9) for each component separately, i.e., it only requires to verify
\[ \sum_{j=1}^{M} P^L(i,j)g^L(j) + f(i,L(i)) \leq \sum_{j=1}^{M} P^{L'}(i,j)g^L(j) + f'(i,L(i)) \]
over \( L(i) \in \mathcal{A}(i) \) for \( i = 1, 2, \ldots, M \). That is, the search is over the action space \( \mathcal{A}(i) \) for every \( i = 1, 2, \ldots, M \) instead of over the much larger policy space \( \mathcal{E} \). It can be found in (1.9) that we can yield the same result in evaluating policies if we replace the potential \( g \) with \( g + ce \) where \( c \) is any constant.

Based on the above Proposition, a policy iteration algorithm is proposed in [14]. It is also guaranteed that the iteration will converge to the optimal policy with probability one. Interested readers can refer to [13] and [14] for details.

**Algorithm** (standard policy iteration algorithm)

1. Choose an initial policy \( L_0 \) set \( k := 0 \)
2. Compute \( \pi L_k \) and \( g L_k \)
3. Determine the next policy \( L_{k+1} \) by applying
   \[ \sum_{j=1}^{M} P^{L_{k+1}}(i,j)g^L_k(j) + f(i,L_{k+1}(i)) \leq \sum_{j=1}^{M} P^\alpha(i,j)g^L_k(j) + f'(i,\alpha) \]
   for all \( \alpha \in \mathcal{A}(i) \), to all \( i = 1, 2, \ldots, M \).
4. If \( L_{k+1} \neq L_k \), set \( k := k + 1 \) and go to step 3; otherwise, exit.
Bibliography


Chapter 2

Markov Decision Problem Based 
On A Two-level System

2.1 Introduction

Some systems may be viewed as consisting of a lower-level component that corresponds to a markov process, and a higher-level component which switch between different process operating modes and is also a markov process. We concentrate on this two-level model with upper-level controller and low-level controller. This kind of system have been discussed in the paper [19]. However in [19], the markov decision problem is only discussed in the case of that either of the upper controller or the low-level controller is uncontrollable. In this paper we'll discuss the markov decision problem when both of the upper controller and the low-level controller are controllable.

There are some papers involved in large markov chains ([10], [19], [1], [16]). Some papers ([20], [21], [22]) on hybrid system are also interesting.

2.2 The Two-level Model

The two-level model discussed in this paper is proposed in [19]. We describe the two-level model here again. And the description is similar.

Consider a machine that can be set to work in one of the $M$ modes. When the machine is in mode $m$, $m \in \{1, \cdots, M\}$, it can be at one of the $N_m$ mode-dependent settings,
and the machine changes settings according to an irreducible discrete-time Markov chain (DTMC) with the transition probability matrix \( S^{(m)} = [s_{j,n}^{(m)}] \). The mode of the machine also changes according to an irreducible DTMC, with the transition probability matrix \( R = [r_{im}], i, m = 1, ..., M \). The mode and setting changes are assumed to be independent, and in general, the machine setting changes in a pace several orders of magnitude quicker than the machine mode changes. (The difference in the pace of changes is to suit real-life applications, not for the solution procedure.) Whenever the machine mode changes from \( i \) to \( m \), the initial machine setting follows the distribution \( \theta^{(i,m)} = (\theta_1^{(i,m)}, \ldots, \theta_{N_m}^{(i,m)}) \), \( i \neq m; i, m = 1, ..., M \). For each period that the machine is in mode \( m \) and setting \( n \), a reward of \( f_{(m,n)} \) is generated.

Define the system to be the evolution of the machine modes and settings, i.e., the state of the system is \((m, n)\) when the machine is in setting \( n \) of mode \( m \). Arrange the states in the lexicographical order with the mode as the primary and the setting as the secondary key, i.e., the states of the system in their ascending lexicographical order are \( = \{(1, 1), \ldots, (1, N_1), (2, 1), \ldots, (M, N_M)\}\). Suppose that \( R, \theta^{(i,m)} \) and \( S^{(m)} \) are all given. Let \( Q^{(i,m)} \) be the matrix whose rows are all equal to \( \theta^{(i,m)} \). Then the system can be modeled as a DTMC \( \{(X_t, Y_t)\} \), where \( X_t \) and \( Y_t \) are the machine mode and setting at period \( t \), respectively. The transition probability of \( \{(X_t, Y_t)\} \) is then

\[
P = \begin{bmatrix} r_{11}S^{(1)} & r_{12}Q^{(1,2)} & \cdots & r_{1M}Q^{(1,M)} \\ r_{21}Q^{(2,1)} & r_{22}S^{(2)} & \cdots & r_{2M}Q^{(2,M)} \\ \vdots & \vdots & \ddots & \vdots \\ r_{M1}Q^{(M,1)} & r_{M2}Q^{(M,2)} & \cdots & r_{MM}S^{(M)} \end{bmatrix} \tag{2.1}
\]

\( P \) also can be written as:

\[
P = \begin{bmatrix} r_{11}S^{(1)} & r_{12}e^{(1)}\theta^{(1,2)} & \cdots & r_{1M}e^{(1)}\theta^{(1,M)} \\ r_{21}e^{(2)}\theta^{(2,1)} & r_{22}S^{(2)} & \cdots & r_{2M}e^{(2)}\theta^{(2,M)} \\ \vdots & \vdots & \ddots & \vdots \\ r_{M1}e^{(M)}\theta^{(M,1)} & r_{M2}e^{(M)}\theta^{(M,2)} & \cdots & r_{MM}S^{(M)} \end{bmatrix} \tag{2.2}
\]

where \( e^{(m)} \) is a \( N_m \times 1 \) column vector with all its components are 1. By assumption, \( P \) is finite, irreducible and hence is ergodic.

Let \((\cdot)^T \) be the transpose of a vector or a matrix \((\cdot)\). The performance function of mode
\[ f_m = (f_{(m,1)}, \ldots, f_{(m,n_m)})^T, \quad (2.3) \]

is the reward per period in mode \( m \); the performance function of the system is then

\[ f = (f_{(1,1)}, \ldots, f_{(M,N_M)})^T. \quad (2.4) \]

Sometimes we use \( f_{mn} \) to denote \( f_{(m,n)} \). The long-run average reward of this ergodic chain is well-defined and is given by

\[ \eta = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t, Y_t). \]

Now suppose that the machine is controlled by an administrator. The administrator wants to maximize the long-run average reward of the machine, exerting his control over \( \theta^{(i,m)} \), \( R \) and \( S^{(m)} \). The administrator use only stationary policy, and for any policy, \( R \) and \( S^{(m)} \) are irreducible matrices. We would like to find a procedure to determine the optimal policy for them. We’ll introduce the model with an administrator in detail later. Now let’s analyze the system for a fixed policy.

### 2.3 The Analysis of the Two-Level System

In this subsection, we will analyze the system for a fixed policy. This analysis gives intermediate results that later ones are built on. Because the analysis is generic, we omit the dependence on policies.

Given a fixed policy, \( \{(X_t, Y_t)\} \) is an ergodic DTMC with the transition probability matrix \( P \) as given in (2.1), and the mode process \( \{X_t\} \) is another ergodic DTMC with the transition probability matrix \( R \). Let \( \rho \) be the stationary distribution of \( R \), i.e.,

\[ \sum_{m=1}^{M} \rho_m = 1, \quad \text{and} \quad \rho = \rho R. \quad (2.5) \]

Let \( I \) be an identity matrix; \( e \) be a column vector with all elements equal to 1; \( e_n \) be a column vector with all elements equal to zero and the \( n \)th equal to 1. The dimensions of \( I \), \( e \), and \( e_n \) are defined through the context.

The following proposition have been given and proofed in [19]. Here we stated it again and give a new proof to it.
Proposition 2.3.1 Let $\pi$ be the stationary distribution of the DTMC $P$. Then

$$\pi_{mn} = \sum_{i \neq m} \rho_i r_{im} \theta^{(i,m)}(I - r_{mm} S^{(m)})^{-1} e_n.$$  \hspace{1cm} (2.6)

The long-run average reward from mode $m$,

$$v_m = \sum_{i \neq m} \rho_i r_{im} \theta^{(i,m)}(I - r_{mm} S^{(m)})^{-1} f_m.$$  \hspace{1cm} (2.7)

and the long-run average reward of the system is

$$\eta = \sum_{m=1}^{M} v_m = \sum_{m=1}^{M} \sum_{i \neq m} \rho_i r_{im} \theta^{(i,m)}(I - r_{mm} S^{(m)})^{-1} f_m.$$ \hspace{1cm} (2.8)

We can prove 2.3.1 by mathematical method. And we need two lemma below.

Lemma 2.3.2 (Sherman-Morrison-Woodbury Formula) If $A, B$ are both invertible matrix and $B = A + UV^T$ where $U$ and $V$ are two matrices, then:

$$B^{-1} = (A + UV^T)^{-1} = A^{-1} - [A^{-1}U(I + V^T A^{-1}U)^{-1} V^T A^{-1}]$$ \hspace{1cm} (2.9)

Note that $B$ is invertible if and only if $(I + V^T A^{-1}U)$ is invertible.

Lemma 2.3.3 For an ergodic markov chain with probability transition matrix $P$ and steady state vector $\pi$, we have

$$\lim_{\lambda \to 1} (\lambda - 1)(\lambda I - P)^{-1} = e \pi$$ \hspace{1cm} (2.10)

where $e$ is a column vector that all components are one.

Proof of Proposition 2.3.1.

Let

$$A = \begin{bmatrix} \lambda I - r_{11} S^{(1)} & 0 & \cdots & 0 \\ 0 & \lambda I - r_{22} S^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda I - r_{MM} S^{(M)} \end{bmatrix}$$

$$U = \begin{bmatrix} e^{(1)} & 0 & \cdots & 0 \\ 0 & e^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{(M)} \end{bmatrix}$$

and

$$V^T = \begin{bmatrix} 0 & r_{12} \theta^{(1,2)} & \cdots & r_{1M} \theta^{(1,M)} \\ r_{21} \theta^{(2,1)} & 0 & \cdots & r_{2M} \theta^{(2,M)} \\ \vdots & \vdots & \ddots & \vdots \\ r_{M1} \theta^{(M,1)} & r_{M2} \theta^{(M,2)} & \cdots & 0 \end{bmatrix}$$
For the two level model (2.1), we have

\[ \lambda I - P = A + UV^T \]

Then by Sherman Formula (Lemma 2.3.2):

\[
(\lambda - 1)(\lambda I - P)^{-1} = (\lambda - 1)\left(A^{-1} - [A^{-1}U(I + V^TA^{-1}U)^{-1}V^TA^{-1}]\right)
\]

And we can compute \( I + V^TA^{-1}U \) as follows:

\[
I + V^TA^{-1}U = I - V^T \begin{bmatrix}
(\lambda - r_{11})^{-1}e^{(1)} & 0 & \cdots & 0 \\
0 & (\lambda - r_{22})^{-1}e^{(2)} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & (\lambda - r_{MM})^{-1}e^{(M)}
\end{bmatrix}
\]

\[
= I - \begin{bmatrix}
0 & r_{12}(\lambda - r_{22})^{-1} & \cdots & r_{1M}(\lambda - r_{MM})^{-1} \\
r_{21}(\lambda - r_{11})^{-1} & 0 & \cdots & r_{2M}(\lambda - r_{MM})^{-1} \\
\vdots & \vdots & \ddots & \vdots \\
r_{M1}(\lambda - r_{11})^{-1} & r_{M2}(\lambda - r_{22})^{-1} & \cdots & 0
\end{bmatrix}
\]

Note that \((I - r_{mm}S^{(m)})\) is invertible. Then \((\lambda I - r_{mm}S^{(m)})\) is also invertible for \(\lambda\) is near to 1.

\[
A^{-1} - [A^{-1}U(I + V^TA^{-1}U)^{-1}V^TA^{-1}]
\]

\[
= [I - A^{-1}U(I + V^TA^{-1}U)^{-1}V^T]A^{-1}
\]

\[
= (I + \begin{bmatrix}
(\lambda - r_{11})^{-1}e^{(1)} & 0 & \cdots & 0 \\
0 & (\lambda - r_{22})^{-1}e^{(2)} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & (\lambda - r_{MM})^{-1}e^{(M)}
\end{bmatrix}) (I + V^TA^{-1}U)^{-1}V^TA^{-1}
\]
\[
= (I - U)
\begin{bmatrix}
(\lambda - \tau_{11})^{-1} & 0 & \cdots & 0 \\
0 & (\lambda - \tau_{22})^{-1} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & (\lambda - \tau_{MM})^{-1}
\end{bmatrix}
(I + V^T A^{-1} U)^{-1} V^T A^{-1}
\]

\[
= (I - U(\lambda I - 
\begin{bmatrix}
\tau_{11} & \tau_{12} & \cdots & \tau_{1M} \\
\tau_{21} & \tau_{22} & \cdots & \tau_{2M} \\
\vdots & \vdots & \ddots & \vdots \\
\tau_{M1} & \tau_{M2} & \cdots & \tau_{MM}
\end{bmatrix})^{-1} V^T A^{-1}
\]

\[
= A^{-1} + 
\begin{bmatrix}
e^{(1)} & 0 & \cdots & 0 \\
0 & e^{(2)} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & e^{(M)}
\end{bmatrix}
(\lambda I - R)^{-1} V^T A^{-1}
\]

Then by Lemma 2.3.3, we have

\[
e\pi = \lim_{\lambda \to 1} (\lambda - 1)(\lambda I - P)^{-1}
= \lim_{\lambda \to 1} (\lambda - 1)(A^{-1} - [A^{-1} U (I + V^T A^{-1} U)^{-1} V^T A^{-1}] )
= \lim_{\lambda \to 1} (\lambda - 1)(A^{-1} - U (\lambda I - R)^{-1} V^T A^{-1})
= \lim_{\lambda \to 1} -U(\lambda - 1) (\lambda I - R)^{-1} V^T A^{-1}
= \lim_{\lambda \to 1} \rho V^T A^{-1}
\]

where \(\rho\) is the steady state vector of \(R\). Then it's clear that

\[
\pi_{mn} = \sum_{i \neq m} \rho_i r_{im} \theta^{(i,m)} (I - \tau_{mm} S^{(m)})^{-1} e_n
\]

\[
\diamondsuit \diamondsuit
\]

We know the performance potential \([5]\) is important in Markov Decision Processes. The performance potential \(g\) satisfy the following Poisson equation:

\[
(I - P + e\pi)g = f \tag{2.11}
\]
where $g$ is a column vector, $e$ is a $\sum_{m=1}^{M} N_m$-dim column vector with all its components are 1. If the dimensions of $P$ is large, the computation of $g$ is difficult. We’ll give an effective method to compute the potential $g$ for our two-level system.

From (2.11), we know

$$g = (I - P + e\pi)^{-1} f$$

where $(I - P + e\pi)^{-1}$ is called fundamental matrix (Kemeny & Snell, 1990).

We denote

$$G_P = (I - P + e\pi)^{-1}$$

$$G_R = (I - R + e_M\rho)^{-1}$$

where $e$ is a $M$-dim column vector with all its components are 1.

Then let’s see how to solve $g$ from the Poisson equation (2.11) by the proposition below.

**Proposition 2.3.4** The fundamental matrix $G_P$ in (2.11) can be obtained by the following equation:

$$G_P = K + E^{(M)} G_R V^T K - e\rho V^T K^2$$

Then

$$g = \left( K + E^{(M)} G_R V^T K - e\rho V^T K^2 \right) f$$

where $K = diag[K_m] = diag[(I - r_{mm} S^{(m)})^{-1}], E^{(M)} = diag[e^{(m)}]$,

$$V^T = \begin{bmatrix}
0 & r_{12} \theta^{(1,2)} & \cdots & r_{1M} \theta^{(1,M)} \\
r_{21} \theta^{(2,1)} & 0 & \cdots & r_{2M} \theta^{(2,M)} \\
\vdots & \vdots & \ddots & \vdots \\
r_{M1} \theta^{(M,1)} & r_{M2} \theta^{(M,2)} & \cdots & 0
\end{bmatrix}$$

**Proof**: $(I - P + e\pi)$ is invertible for an ergodic Markov chain [5]. So from 2.11 we know

$$g = (I - P + e\pi)^{-1} f$$

Now we use Lemma 2.3.2 to compute $(I - P + e\pi)^{-1}$.  

16
Denote

\[ A_1 = \begin{bmatrix}
I - r_{11}S^{(1)} & 0 & \cdots & 0 \\
0 & I - r_{22}S^{(2)} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & I - r_{MM}S^{(M)}
\end{bmatrix} \]

\[ U = \begin{bmatrix}
e^{(1)} & 0 & \cdots & 0 \\
0 & e^{(2)} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & e^{(M)}
\end{bmatrix} \quad \text{and} \quad V^T = \begin{bmatrix}
0 & r_{12}\theta^{(1,2)} & \cdots & r_{1M}\theta^{(1,M)} \\
r_{21}\theta^{(2,1)} & 0 & \cdots & r_{2M}\theta^{(2,M)} \\
\vdots & \vdots & \ddots & \vdots \\
r_{M1}\theta^{(M,1)} & r_{M2}\theta^{(M,2)} & \cdots & 0
\end{bmatrix} \]

Letting \( B = I - P + e\pi \) and \( A = A_1 + e\pi \). We have

\[ I - P + e\pi = B = A + UV^T = (A_1 + e\pi) + UV^T \quad (2.17) \]

By Lemma 2.3.2, we obtain

\[ A^{-1} = A_1^{-1} - A_1^{-1}e\pi A_1^{-1} \frac{1}{1 + \pi A_1^{-1}e} \quad (2.18) \]

Note that \( 1 + \pi A_1^{-1}e > 0 \), then \( A^{-1} \) is invertible. Also by Lemma 2.3.2, we obtain

\[ G_P = (I - P + e\pi)^{-1} = B^{-1} = (A + UV^T)^{-1} = A^{-1} - [A^{-1}U(I + V^TA^{-1}U)^{-1}V^TA^{-1}] \]

Denote

\[ K_m = (I - r_{mm}S^{(m)})^{-1} \quad \tau = 1 + \pi A_1^{-1}e = 1 + \sum_{m=1}^{M} \frac{\rho_m}{1 - r_{mm}} \quad (2.20) \]

and

\[ e_\tau = \frac{1}{\tau} \text{diag}(\frac{1}{1 - r_{mm}})e_M \]

Some matrices above can be simplified as follows:

\[ A^{-1} = A_1^{-1} - A_1^{-1}e\pi A_1^{-1} \frac{1}{1 + \pi A_1^{-1}e} \]

\[ = (I - \frac{1}{\tau} A_1^{-1}e\pi)A_1^{-1} \]

\[ = (I - \frac{1}{\tau} \text{diag}(e^{(m)}) \text{diag}(\frac{1}{1 - r_{mm}})e_M\pi)A_1^{-1} \]

\[ = (I - \frac{1}{\tau} \text{diag}(e^{(m)}) \text{diag}(\frac{1}{1 - r_{mm}})e_M\pi) \text{diag}(K_m) \]

\[ = (I - \text{diag}(e^{(m)})e_\tau\pi) \text{diag}(K_m) \]
So

\[
A^{-1}U = -(I - \text{diag}(e^{(m)}\epsilon_\tau \pi)\text{diag}(K_m)\text{diag}(e^{(m)})) = -\text{diag}(e^{(m)})(I - \epsilon_\tau \rho)\text{diag}(\frac{1}{1 - r_{mm}})
\]

Then

\[
G_R = A^{-1} - [A^{-1}U(I + V^T A^{-1}U)^{-1}V^T A^{-1}] = A^{-1} - [A^{-1}U(I - V^T \text{diag}(e^{(m)})(I - \epsilon_\tau \rho)\text{diag}(\frac{1}{1 - r_{mm}}))^{-1}V^T A^{-1}] = A^{-1} - [A^{-1}U(I - R^0(I - \epsilon_\tau \rho)\text{diag}(\frac{1}{1 - r_{mm}}))^{-1}V^T A^{-1}] = A^{-1} - [A^{-1}U((I - R + R^0 \epsilon_\tau \rho)\text{diag}(\frac{1}{1 - r_{mm}}))^{-1}V^T A^{-1}] = A^{-1} - [A^{-1}U\text{diag}(1 - r_{mm})(I - R + R^0 \epsilon_\tau \rho)^{-1}V^T A^{-1}] = \left(I + \text{diag}(e^{(m)})(I - \epsilon_\tau \rho)(I - R + R^0 \epsilon_\tau \rho)^{-1}V^T\right)A^{-1}
\]

where

\[
R^0 = \begin{bmatrix}
0 & r_{12} & \cdots & r_{1M} \\
r_{21} & 0 & \cdots & r_{2M} \\
\vdots & \vdots & \ddots & \vdots \\
r_{M1} & r_{M2} & \cdots & 0
\end{bmatrix}
\]

By Sherman formula, we can know that

\[
(I - R + R^0 \epsilon_\tau \rho)^{-1} = (I - R + \epsilon_M \rho + (R^0 \epsilon_\tau - \epsilon_M)\rho)^{-1} = G_R - \frac{G_R(R^0 \epsilon_\tau - \epsilon_M)\rho}{1 + \rho G_R(R^0 \epsilon_\tau - \epsilon_M)} = G_R - \frac{G_R(R^0 \epsilon_\tau - \epsilon_M)\rho}{1 + \rho(R^0 \epsilon_\tau - \epsilon_M)} = G_R - \frac{G_R(R^0 \epsilon_\tau - \epsilon_M)\rho}{\rho R^0 \epsilon_\tau} = G_R - \tau G_R(R^0 \epsilon_\tau - \epsilon_M)\rho
\]

Note that we make use of the following properties of $G_R$ in above equation:

$\rho G_R = \rho$ and $G_R \epsilon_M = \epsilon_M$
Then

\[(I - e_r \rho) (I - R + R^0 e_r \rho)^{-1}\]
\[= (I - e_r \rho)(G_R - \tau G_R(R^0 e_r - e_M)\rho)\]
\[= G_R - \tau G_R(R^0 e_r - e_M)\rho - e_r \rho(G_R - \tau G_R(R^0 e_r - e_M)\rho)\]
\[= G_R - \tau G_R(R^0 e_r - e_M)\rho - e_r(\rho - \tau(\frac{1}{\tau} - 1))\rho\]
\[= G_R - G_R R^0 diag(\frac{1}{1 - r_{mm}}) e_M \rho + \tau G_R e_M \rho - diag(\frac{1}{1 - r_{mm}}) e_M \rho\]
\[= G_R + \tau G_R e_M \rho - G_R(diag(1 - r_{mm}) + e_M \rho)diag(\frac{1}{1 - r_{mm}}) e_M \rho\]
\[= G_R + \tau e_M \rho - G_R diag(1 - r_{mm}) + e_M \rho diag(\frac{1}{1 - r_{mm}}) e_M \rho\]
\[= G_R + \tau e_M \rho - G_R e_M \rho - G_R e_M \rho diag(\frac{1}{1 - r_{mm}}) e_M \rho\]
\[= G_R + \tau e_M \rho - e_M \rho - e_M(\tau - 1)\rho\]
\[= G_R\]

So

\[G_P\]
\[= \left( I + E^{(M)} G_R V^T \right)(I - E^{(M)} e_r \pi) K\]
\[= (I - E^{(M)} e_r \pi) K + E^{(M)}(G_R V^T - G_R R^0 e_r \rho V^T K) K\]
\[= (I - E^{(M)} e_r \pi) K + E^{(M)}(G_R V^T - ((G_R(diag(1 - r_{mm}) + e_M \rho) - I) e_r \rho V^T K) K\]
\[= (I - E^{(M)} e_r \pi) K + E^{(M)}(G_R V^T - G_R(\frac{1}{\tau} e_M + e_M \rho e_r) \rho V^T K + e_r \rho V^T K) K\]
\[= (I - E^{(M)} e_r \pi) K + E^{(M)}(G_R V^T - (\frac{1}{\tau} e_M + e_M \rho e_r) \rho V^T K + e_r \rho V^T K) K\]
\[= (I - E^{(M)} e_r \pi) K + E^{(M)}(G_R V^T - e_M \rho V^T K - e_r \rho V^T K) K\]
\[= K + E^{(M)} G_R V^T K - e_\rho V^T K^2\]

\diamondsuit \diamondsuit

The major computation involved in the above Proposition consists of three parts: (i) Calculating \((I - r_{mm} S^{(m)})^{-1}\) for \(m = 1, 2, \cdots, M\); (ii) Solving for the steady probability \(\rho\) of the upper-level; (iii) Calculating \((I + V^T A^{-1} U)^{-1}\) in (2.19). The total computation is roughly of the order \((\sum_{m=1}^{M} N_m^3 + 2M^3)\).
2.4 The Two-Level Decision Problem

In this subsection, we’ll describe the decision problem based on our two-level system in detail. Then we propose a policy algorithm. We know the transition probability matrix for our system have the form \((2.2), r_{ij}, S^{(m)}\) and \(\theta^{(i,m)}\) is important to \(P\). We’ll control them separately by the actions in their action set.

We consider the upper-level of the system. There is an action set \(A_u(i)\) for the mode \(i\). The row vector \((r_{i1}, \ldots, r_{iM})\) is changed by different actions in \(A_u(i)\). A policy \(\mathcal{L}^u\) for the upper-level specifies the action for each mode \(i\). When system is in mode \(i\), the administrator take the action \(\mathcal{L}^u(i)\). So under the policy \(\mathcal{L}^u\) the transition probability matrix of the system is:

\[
P_{\mathcal{L}^u} = \begin{bmatrix}
    r_{11}^{u(1)} S_{(1)} & r_{12}^{u(1)} e_{(1)} & \cdots & r_{1M}^{u(1)} e_{(1)} \\
    r_{21}^{u(2)} e_{(2)} & r_{22}^{u(2)} S_{(2)} & \cdots & r_{2M}^{u(2)} e_{(2)} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{M1}^{u(M)} e_{(M)} & r_{M2}^{u(M)} e_{(M)} & \cdots & r_{MM}^{u(M)} S_{(M)}
\end{bmatrix}
\] (2.22)

We consider the initial distribution \(\theta^{(i,m)}\) when system changes from mode \(i\) to mode \(m(i \neq m)\). There is an action set \(A_i(i, m)\) for \(\theta^{(i,m)}\). The row vector \(\theta^{(i,m)}\) is changed by different actions in \(A_i(i, m)\). A policy \(\mathcal{L}^i\) specifies the action for each \(\theta^{(i,m)}\). When system changes from mode \(i\) to mode \(m\), the administrator take the action \(\mathcal{L}^i(i, m)\). So under the policy \(\mathcal{L}^i\) the transition probability matrix of the system is:

\[
P_{\mathcal{L}^i} = \begin{bmatrix}
    r_{11}^{i(1)} S_{(1)} & r_{12} e_{(1)} & \cdots & r_{1M} e_{(1)} \\
    r_{21} e_{(2)} & r_{22} S_{(2)} & \cdots & r_{2M} e_{(2)} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{M1} e_{(M)} & r_{M2} S_{(M)} & \cdots & r_{MM} S_{(M)}
\end{bmatrix}
\] (2.23)

Now we consider the low-level of the system. There is an action set \(A_l(m, j)\) for the setting \((m, j)\). The row vector \((s_{j1}^{(m)}, \ldots, s_{jN_m}^{(m)})\) is changed by different actions in \(A_l(m, j)\). A policy \(\mathcal{L}^l(m)\) specifies the action for each setting \((m, j)\) in mode \(m\). When system is in the setting \((m, j)\), the administrator take the action \(\mathcal{L}^l(m, j)\). So under the policy \(\mathcal{L}^l\) the
transition probability matrix of the system is:

$$
P^{L} = \begin{bmatrix}
    r_{11} S^{(1)} L_{1}^{(1)} & r_{12} e^{(1)}(1,2) & \cdots & r_{1M} e^{(1)}(1,M) \\
    r_{21} e^{(2)}(2,1) & r_{22} S^{(2)} L_{2}^{(2)} & \cdots & r_{2M} e^{(2)}(2,M) \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{M1} e^{(M)}(M,1) & r_{M2} e^{(M)}(M,2) & \cdots & r_{MM} S^{(M)} L_{M}^{(M)}
\end{bmatrix}$$

(2.24)

where

$$
S(m) L^{(m)} = \begin{bmatrix}
    s_{11}^{(m)} L_{1}^{(m,1)} & s_{12}^{(m)} L_{1}^{(m,1)} & \cdots & s_{1N_{m}}^{(m)} L_{1}^{(m,1)} \\
    s_{21}^{(m)} L_{2}^{(m,2)} & s_{22}^{(m)} L_{2}^{(m,2)} & \cdots & s_{2N_{m}}^{(m)} L_{2}^{(m,2)} \\
    \vdots & \vdots & \ddots & \vdots \\
    s_{N_{m1}}^{(m)} L_{N_{m}}^{(m,N_{m})} & s_{N_{m2}}^{(m)} L_{N_{m}}^{(m,N_{m})} & \cdots & s_{N_{mN_{m}}}^{(m)} L_{N_{m}}^{(m,N_{m})}
\end{bmatrix}
$$

(2.25)

Since the system can be control by the three kind of policies $L^{u}, L^{l}$ and $L^{i}$, we can define a new policy $L$ which specifies $L^{u}, L^{i}$ and $L^{l}$. So under the policy $L$ the transition probability matrix of the system is:

$$
P^{L} = \begin{bmatrix}
    r_{11}^{L_{1}} S^{(1)} L_{1}^{(1)} & r_{12}^{L_{1}} Q^{(1,2)} L_{2}^{(2,1)} & \cdots & r_{1M}^{L_{1}} Q^{(1,M)} L_{M}^{(M,1)} \\
    r_{21}^{L_{2}} Q^{(2,1)} L_{2}^{(2,1)} & r_{22}^{L_{2}} S^{(2)} L_{2}^{(2)} & \cdots & r_{2M}^{L_{2}} Q^{(2,M)} L_{M}^{(M,2)} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{M1}^{L_{M}} Q^{(M,1)} L_{M}^{(M,1)} & r_{M2}^{L_{M}} Q^{(M,2)} L_{M}^{(M,2)} & \cdots & r_{MM}^{L_{M}} S^{(M)} L_{M}^{(M)}
\end{bmatrix}
$$

(2.26)

where $S^{(m)} L^{(m)}$ has the same form as (2.25) and $Q^{(i,m)} L^{(i,m)} = e^{(i)}(i,m) L^{(i,m)}$.

The performance function $f = \begin{bmatrix} f_1 & f_2 & \cdots & f_M \end{bmatrix}^T$ also is affected by the policy $L$. In general $f$ can be written as:

$$
f^{L} = \begin{bmatrix}
    f_1^{L_{1}} \\
    f_2^{L_{2}} \\
    \vdots \\
    f_M^{L_{M}}
\end{bmatrix} = \begin{bmatrix}
    f_1^{L_{1}^{u}, L_{1}^{(1)}, L_{1}^{(1)}} \\
    f_2^{L_{2}^{u}, L_{2}^{(2)}, L_{2}^{(2)}} \\
    \vdots \\
    f_M^{L_{M}^{u}, L_{M}^{(M)}, L_{M}^{(M)}}
\end{bmatrix}
$$

(2.27)

where $L^{i}(m)$ represent $L^{i}(m,1), L^{i}(m,2), \cdots, L^{i}(m,M)$.

Now we simplify the performance function $f$. We assume that $f$ does not depend on $L^{i}(m,j)$ ($j = 1, 2, \cdots, M$). That is:
Assumption 2.4.1 The performance function $f^L$ is independent of the policy $L_i(m, j)$ ($j = 1, 2, \cdots, M$), where the policy $L_i(m, j)$ controls the initial distribution $\theta^{(m,j)}$ when the two-level system is changed from mode $m$ to mode $j$.

Under the above assumption, we have

$$f^L = \begin{bmatrix}
    f^L_1 \\
    f^L_2 \\
    \vdots \\
    f^L_M
\end{bmatrix} = \begin{bmatrix}
    f^L_{1(1)}, L^L(1) \\
    f^L_{1(2)}, L^L(2) \\
    \vdots \\
    f^L_{M(1)}, L^L(1)
\end{bmatrix}$$ \hspace{1cm} (2.28)

Given Assumption 2.4.1, a special property for the optimal policy is induced.

Proposition 2.4.1 For any mode $m$, it suffices for the administrator to consider $\theta^{(i,m)}$ of the form $\theta^{(i,m)} = \theta^{(m)} = \epsilon^n_T$ for some $n \in \{1, \cdots, N_m\}$.

Proof. For any mode $m$ with a given $R$ and $S^{(m)}$, Lemma ?? shows that the expected total reward of the sojourn time starting at state $(m, n)$ is $\epsilon_n^T(I - r_{mm}S^{(m)})^{-1}f_m$. Any randomization of the initial setting of mode $m$ cannot give a higher expected value from the setting that maximizes $\epsilon_n^T(I - r_{mm}S^{(m)})^{-1}f_m$. \hfill $\diamondsuit$

Because of Proposition 2.4.1, it suffices for the administrator to search among policies such that $\theta^{(i,m)}$ is of a point mass dependent only on $m$. Let $Q^{(m)}$ be an $N_m \times N_m$ matrix of each row identically equal to $\theta^{(m)}$. The transition probability matrix $P$ is simplified into

$$P = \begin{bmatrix}
    r_{11}S^{(1)} & r_{12}e^{(1)} & \cdots & r_{1M}e^{(1)} \theta^{(M)} \\
    r_{21}e^{(2)} & r_{22}S^{(2)} & \cdots & r_{2M}e^{(2)} \theta^{(M)} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{M1}e^{(M)} \theta^{(1)} & r_{M2}e^{(M)} \theta^{(2)} & \cdots & r_{MM}S^{(M)}
\end{bmatrix} \hspace{1cm} (2.29)$$

Now we introduce the concept of the stochastic complement in the paper [16]. There is some connection with this concept and Schur complementation in matrix theory. The probabilistic interpretation of stochastic complementation is also given in paper [16]. The definition of the stochastic complementation is:

22
Definition 2.4.2 Let $P$ be an $N \times N$ irreducible stochastic matrix with a $M \times M$ partition

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1M} \\ P_{21} & P_{22} & \cdots & P_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ P_{M1} & P_{M2} & \cdots & P_{MM} \end{bmatrix}$$

(2.30)

in which all diagonal blocks are square. For a given index $i$, let $P_i$ denotes the block submatrix of $P$ obtained by deleting the $i$th row and $i$th column of blocks from $P$, and let $P_{is}$ and $P_{si}$ designate

$$P_{is} = \begin{bmatrix} P_{i1} & P_{i2} & \cdots & P_{i,i-1} & P_{i,i+1} & \cdots & P_{iM} \end{bmatrix}$$

(2.31)

and

$$P_{si} = \begin{bmatrix} P_{s1} & P_{s2} & \cdots & P_{s-1,i} & P_{s+1,i} & \cdots & P_{sM} \end{bmatrix}^T$$

(2.32)

The stochastic complement of $P_{ii}$ in $P$ is defined to be the matrix

$$S_{ii} = P_{ii} + P_{is} (I - P_i)^{-1} P_{si}$$

(2.33)

Note that it is well known that $I - P$ is a singular $M$-matrix of rank $N - 1$ and that every principal submatrix of $I - P$ of order $N - 1$ or smaller is a nonsingular $M$-matrix (see [17], p.156). So $(I - P_i)$ in (2.33) is invertible.

The coupling theorem (Theorem 4.1 in [16]) is important to us. We stated it as below. Some notations are different from Theorem 4.1 in [16] for convenience.

Lemma 2.4.3 (Coupling Theorem in paper [16]) If $P$ is an $N \times N$ irreducible stochastic matrix partitioned as

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1M} \\ P_{21} & P_{22} & \cdots & P_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ P_{M1} & P_{M2} & \cdots & P_{MM} \end{bmatrix}$$
with square diagonal blocks. Then the stochastic complementation \( S_{ii} \) also is an irreducible stochastic matrix (Theorem 2.3 in paper [16]). Now we partition the stationary distribution vector for \( P \) into

\[
\pi = \begin{bmatrix} \pi^{(1)} & \pi^{(2)} & \ldots & \pi^{(M)} \end{bmatrix}
\]

(2.34)

then the stationary distribution vector for \( P \) is given by

\[
\pi = \begin{bmatrix} \xi_1 s_1 & \xi_2 s_2 & \ldots & \xi_M s_M \end{bmatrix}
\]

(2.35)

where \( s_i \) is the unique stationary distribution vector for the stochastic complement

\[
S_{ii} = P_{ii} + P_{i*} (I - P_i)^{-1} P_{si}
\]

(2.36)

and where

\[
\xi = \begin{bmatrix} \xi_1 & \xi_2 & \ldots & \xi_M \end{bmatrix}
\]

(2.37)

is the unique stationary distribution vector for the \( M \times M \) irreducible stochastic matrix \( D \) whose entries are defined by

\[
d_{ij} = s_i P_{ij} e^{(j)}
\]

(2.38)

The matrix \( D \) is hereafter referred to as the **coupling matrix**, and the scalars \( \xi_i \) are called the **coupling factors**.

For the two-level model, we denote:

\[
S^{(m)*} = \text{diag} \left[ \begin{array}{cccc} r_{11} S^{(1)} & r_{22} S^{(2)} & \ldots & r_{i-1,i-1} S^{(i-1)} & r_{i+1,i+1} S^{(i+1)} & \ldots & r_{MM} S^{(M)} \end{array} \right]
\]

(2.39)

\[
\theta^{(m)*} = \text{diag} \left[ \begin{array}{cccc} \theta^{(1)} & \theta^{(2)} & \ldots & \theta^{(i-1)} & \theta^{(i+1)} & \ldots & \theta^{(M)} \end{array} \right]
\]

(2.40)

and

\[
e^{(m)*} = \text{diag} \left[ \begin{array}{cccc} e^{(1)} & e^{(2)} & \ldots & e^{(m-1)} & e^{(m+1)} & \ldots & e^{(M)} \end{array} \right]
\]

(2.41)

For the upper level, the we denote

\[
r_{m*} = \begin{bmatrix} r_{m1} & r_{m2} & \cdots & r_{m,m-1} & r_{m,m+1} & \cdots & r_{m,M} \end{bmatrix}
\]

(2.42)
\[ r_{*m} = \begin{bmatrix} r_{1m} & r_{2m} & \cdots & r_{m-1,m} & r_{m+1,m} & \cdots & r_{M,m} \end{bmatrix}^T \] (2.43)

and

\[ R^0 = \begin{bmatrix} 0 & r_{12} & \cdots & r_{1M} \\ r_{21} & 0 & \cdots & r_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ r_{M1} & r_{M2} & \cdots & 0 \end{bmatrix} \] (2.44)

If we delete the \( m \)th row and \( m \)th column, we obtain the matrix \( R^0_m \); the matrix \( R^{(m)*} \) is defined as

\[ R^{(m)*} = R^0_m \theta^{(m)*} \] (2.45)

**Proposition 2.4.4** If we use the stochastic complementation method in Lemma 2.4.3 to solve the steady state vector \( \pi \) for the two-level model, we have the following results under Assumption 2.4.1:

(i) For the two-level model, the stochastic complementation \( S_{mm} \) defined in (2.33) is:

\[ S_{mm} = r_{mm}S^{(m)} + \epsilon^{(m)}\theta^{(m)}(1 - r_{mm}) \]

where \( R_m \) is obtained by deleting the \( m \)th row and the \( m \)th column of the matrix \( R \).

(ii) For the two-level model, we have

\[ D = R \text{ and } \xi = \rho \] (2.46)

where the matrix \( D \) is defined in (2.38) and \( R = [r_{ij}]_{M \times M} \cdot \xi \) and \( \rho \) is defined in (2.37) and (2.5).

(iii) For the two-level model, we have

\[ s_m = (1 - r_{mm})\theta^{(m)}(I - r_{mm}S^{(m)})^{-1} \] (2.47)

where \( s_m \) is the steady state vector of the stochastic complement \( S_{mm} \).

(iv) We have

\[ G_S = (I - S_{mm} + \epsilon^{(m)}s_m)^{-1} \]
\[
G_S = -(1 - r_{mm})O^{(m)}(K_m)^2 + (1 - r_{mm})O^{(m)}K_m + K_m
\]  
(2.48)

where

\[
O^{(m)} = e^{(m)}g^{(m)} \text{ and } K_m = (I - r_{mm}S^{(m)})^{-1}
\]

Proof:

(i) We know the stochastic complementation \( S_{mm} \) is

\[
S_{mm} = P_{mm} + P_{m*}(I - P_m)^{-1}P_{*m}
\]  
(2.49)

where

\[
I - P_m = I - S^{(m)*} - e^{(m)*}R^{(m)*}
\]  
(2.50)

Letting

\[
Z^{(m)} = [z^{(m)}_{ij}]_{(M-1) \times (M-1)} = \left( I - S^{(m)*} - e^{(m)*}R^{(m)*} \right)^{-1}
\]

By Sherman formula (Lemma 2.3.2), we have:

\[
Z^{(m)} = (I - P_m)^{-1} = \left( I - S^{(m)*} - e^{(m)*}R^{(m)*} \right)^{-1}
\]

\[
= \left( I - S^{(m)*} \right)^{-1} + \left( I - S^{(m)*} \right)^{-1}e^{(m)*}
\]
\[
\cdot \left( I - R^{(m)*} \right)^{-1}e^{(m)*} \left( I - S^{(m)*} \right)^{-1}R^{(m)*} \left( I - S^{(m)*} \right)^{-1}
\]

\[
= \left( I - S^{(m)*} \right)^{-1} + e^{(m)*}(I - R_m)^{-1}R^{(m)*} \left( I - S^{(m)*} \right)^{-1}
\]

where \( R_m \) is obtained by deleting the \( m \)th row and the \( m \)th column of the matrix \( R = [r_{ij}]_{M \times M} \).

Then

\[
S_{mm} = P_{mm} + P_{m*}(I - P_m)^{-1}P_{*m}
\]

\[
= r_{mm}S^{(m)} + e^{(m)}g^{(m)}(r_{m*}g^{(m)}Z^{(m)}e^{(m)*}r_{*m})
\]

\[
= r_{mm}S^{(m)} + e^{(m)}g^{(m)}(r_{m*}g^{(m)}(I + e^{(m)*}(I - R_m)^{-1}R^{(m)*}) \left( I - S^{(m)*} \right)^{-1}e^{(m)*}r_{*m})
\]

\[
= r_{mm}S^{(m)} + e^{(m)}g^{(m)}(r_{m*}g^{(m)}(I + e^{(m)*}(I - R_m)^{-1}R^{(m)*})e^{(m)*}diag[1 - \frac{1}{1 - r_{ii}}r_{*m})]
\]

26
\[ r_{mm}S^{(m)} + e^{(m)}\theta^{(m)}(r_{mm} + e^{(m)}\theta^{(m)}e^{(m)}\theta^{(m)})(1 - r_{mm}) = r_{mm}S^{(m)} + e^{(m)}\theta^{(m)}(r_{mm}(I + (I - R_m)^{-1}R_m^0)\text{diag} \left[ \frac{1}{1 - r_{ii}} \right] r_{mm}) \]

\[ = r_{mm}S^{(m)} + e^{(m)}\theta^{(m)}(r_{mm}(I + r_{mm} - I + R_m + R_m^0)\text{diag} \left[ \frac{1}{1 - r_{ii}} \right] r_{mm}) \]

\[ = r_{mm}S^{(m)} + e^{(m)}\theta^{(m)}(r_{mm}(I - R_m)^{-1}r_{mm}) \]

\[ = r_{mm}S^{(m)} + e^{(m)}\theta^{(m)}(1 - r_{mm}) \]

(ii) It's easy to obtain this by the definition of the two-level model.

(iii) Since \( \theta^{(i,m)} = \theta^{(m)} \) and \( \sum_{i \neq m} \rho_i r_{im} = (1 - r_{mm}) \rho_m \), we know (2.47) holds by (2.6) and Proposition 2.4.1.

(iv) We know

\[ (I - S_{mm} + e^{(m)}s_m) = I - r_{mm}e^{(m)} - e^{(m)}\theta^{(m)}(I - R_m)^{-1}r_{mm}) + e^{(m)}s_m \]

\[ = I - r_{mm}e^{(m)} + e^{(m)}s_m - e^{(m)}\theta^{(m)}(I - R_m)^{-1}r_{mm}) \]

First we compute the inverse of \((I - r_{mm}S^{(m)} + e^{(m)}s_m)\) by Sherman formula

\[ (I - r_{mm}S^{(m)} + e^{(m)}s_m)^{-1} \]

\[ = (I - r_{mm}S^{(m)})^{-1} - \frac{(I - r_{mm}S^{(m)})^{-1}e^{(m)}s_m(I - r_{mm}S^{(m)})^{-1}}{1 + s_m(I - r_{mm}S^{(m)})^{-1}e^{(m)}} \]

\[ = (I - r_{mm}S^{(m)})^{-1} - \frac{(I - r_{mm}S^{(m)})^{-1}e^{(m)}s_m(I - r_{mm}S^{(m)})^{-1}}{1 + \frac{1}{1 - r_{mm}}} \]

\[ = (I - r_{mm}S^{(m)})^{-1} - \frac{1 - r_{mm}}{2 - r_{mm}}(I - r_{mm}S^{(m)})^{-1}e^{(m)}s_m(I - r_{mm}S^{(m)})^{-1} \]

\[ = (I - r_{mm}S^{(m)})^{-1} - \frac{1}{2 - r_{mm}}e^{(m)}s_m(I - r_{mm}S^{(m)})^{-1} \]

\[ = \left( I - \frac{1}{2 - r_{mm}}e^{(m)}s_m \right) (I - r_{mm}S^{(m)})^{-1} \]

Also by Sherman formula, we have

\[ \langle I - S_{mm} + e^{(m)}s_m \rangle = \left( I - \frac{1}{2 - r_{mm}}e^{(m)}s_m \right) K_m - \frac{1}{1 - (1 - r_{mm})\theta^{(m)}} \left( I - \frac{1}{2 - r_{mm}}e^{(m)}s_m \right) K_m e^{(m)} \]

\[ \left( I - \frac{1}{2 - r_{mm}}e^{(m)}s_m \right) K_m \left( -e^{(m)}\theta^{(m)}(1 - r_{mm}) \right) \left( I - \frac{1}{2 - r_{mm}}e^{(m)}s_m \right) K_m \]

27
\[\begin{align*}
(I + (1 - r_{mm})e^{(m)}g^{(m)}) & \left( I - \frac{1}{2 - r_{mm}}e^{(m)}s_m \right) K_m \\
(I + (1 - r_{mm})e^{(m)}g^{(m)}) & \left( I - \frac{1}{2 - r_{mm}}e^{(m)}g^{(m)}K_m \right) K_m \\
(I + (1 - r_{mm})O^{(m)}) & \left( I - \frac{1}{2 - r_{mm}}O^{(m)}K_m \right) K_m \\
-(1 - r_{mm})O^{(m)}(K_m)^2 + (1 - r_{mm})O^{(m)}K_m + K_m
\end{align*}\]

We know the standard policy iteration algorithm is based on the Lemma below.

**Lemma 2.4.5** (see [5]) For a markov chain with two different policies \( \mathcal{C} \) and \( \mathcal{C}' \), we have

\[\eta^{\mathcal{C}'} - \eta^{\mathcal{C}} = \pi^{\mathcal{C}'}[(P^{\mathcal{C}'} - P^{\mathcal{C}})g^{\mathcal{C}} + (f^{\mathcal{C}'} - f^{\mathcal{C}})] \tag{2.51}\]

Then we can maximize \((P^{\mathcal{C}'} - P^{\mathcal{C}})g^{\mathcal{C}} + (f^{\mathcal{C}'} - f^{\mathcal{C}})\) componentwise at each iteration. For our two-level model, (2.51) can be written as:

\[\eta^{\mathcal{C}'} - \eta^{\mathcal{C}} = \sum_{m=1}^{M} \pi^{\mathcal{C}'} \pi^{\mathcal{C}} F_m^{\mathcal{C}' \mathcal{C}} \tag{2.52}\]

where

\[F_m^{\mathcal{C}' \mathcal{C}} = (P^{\mathcal{C}'} - P^{\mathcal{C}})g^{\mathcal{C}} + (f^{\mathcal{C}'} - f^{\mathcal{C}})\]

\[= \sum_{n \neq m} \left( r_{mn}^{(m)}e^{(m)}g^{(n)},\mathcal{C}^{(m)} - r_{mn}^{(m)}e^{(m)}g^{(n)},\mathcal{C}^{(n)} \right) g^{\mathcal{C}} + \left( f^{\mathcal{C}'}(m) - f^{\mathcal{C}}(m) \right) \]

From above, we can see that the values of \(r_{mn}\) and \(f^{\mathcal{C}'}(m)\) affect all settings in mode \(m\) for different upper-level policies. Then it is impossible to maximize \((P^{\mathcal{C}'} - P^{\mathcal{C}})g^{\mathcal{C}} + (f^{\mathcal{C}'} - f^{\mathcal{C}})\) componentwise at each iteration. A new policy iteration algorithm is needed to be developed in order to find the optimal policy for the two-level system.

By Lemma 2.4.3, we know that

\[\pi = \begin{bmatrix} \xi_1 s_1 & \xi_2 s_2 & \cdots & \xi_M s_M \end{bmatrix}\]

Then (2.52) also can be written as

\[\eta^{\mathcal{C}'} - \eta^{\mathcal{C}} = \sum_{m=1}^{M} \xi_m^{\mathcal{C}'} \pi^m F_m^{\mathcal{C}' \mathcal{C}} \tag{2.53}\]
Our idea is maximizing $s_m^{\mathcal{L}_m'} F_m^{\mathcal{L}_m'} \mathcal{L}_k$ by the standard policy iteration algorithm. We note that $F_m^{\mathcal{L}_m'} \mathcal{L}_k$ can be viewed as a new performance function. And $s_m^{\mathcal{L}_m'}$ is the steady state vector of the stochastic complement $S_m^{\mathcal{L}_m'}$ which can be viewed as a probability transition matrix for a Markov chain. But there is still a problem. That is $S_m^{\mathcal{L}_m'}$ doesn’t only depend on $\mathcal{L}_k'(m)$. The matrix $S_m^{\mathcal{L}_m}$ also depends on $\mathcal{L}_k'(n)(n \neq m)$. It seems that we can’t decouple this. Let’s recall that $\theta^{(i,m)} = \theta^{(m)}$ from Assumption 2.4.1. Then we have the following Lemma which is helpful to decouple the case.

**Lemma 2.4.6** For the two-level system with three different policies $\mathcal{L}_k$, $\mathcal{L}$, and $\mathcal{L}'$, we have

$$s_m^{\mathcal{L}_m} F_m^{\mathcal{L}_m} \mathcal{L}_k = s_m^{\mathcal{L}_m} F_m^{\mathcal{L}_m'} \mathcal{L}_k$$

(2.54)

where $\mathcal{L}$ and $\mathcal{L}'$ only have the same $\mathcal{L}_k^u(m)$, $\mathcal{L}_k^l(m)$ and $\mathcal{L}_k^l(m)$ for the mode $m$.

**Proof:** By Proposition 2.4.4 (iii) and (2.53), we have

$$\eta^\mathcal{L} - \eta^\mathcal{L}_k = \sum_{m=1}^{M} \rho_m s_m^{\mathcal{L}_m} F_m^{\mathcal{L}_m} \mathcal{L}_k$$

where

$$s_m^{\mathcal{L}_m} F_m^{\mathcal{L}_m} \mathcal{L}_k = (1 - s_m^{\mathcal{L}_m^u(m)\theta^{(m)} F_m^{\mathcal{L}_m^u(m) S_m^{\mathcal{L}_m^l(m)}} (1 - s_m^{\mathcal{L}_m^u(m) S_m^{\mathcal{L}_m^l(m)}})^{-1} F_m^{\mathcal{L}_m^l(m)} \mathcal{L}_k$$

Since $\mathcal{L}$ and $\mathcal{L}'$ have the same $\mathcal{L}_k^u(m)$, $\mathcal{L}_k^l(m)$ and $\mathcal{L}_k^l(m)$ for the mode $m$, we have

$$s_m^{\mathcal{L}_m'} F_m^{\mathcal{L}_m'} \mathcal{L}_k = s_m^{\mathcal{L}_m} F_m^{\mathcal{L}_m} \mathcal{L}_k$$

Then under the Assumption 2.4.1, we have the following policy iteration algorithm for the two-level system:

**Algorithm 1** (Policy Iterations for the Two-Level System)

1. Set $k = 1$. Choose an initial policy $\mathcal{L}_1$.

2. At the $k$th iteration:

   (a) Compute $\pi^{\mathcal{L}_k}$ by (2.6).
(b) By the method suggested in Proposition 2.15, solve the Poisson equation

\[(I - pL_k + c\pi L_k)gL_k = fL_k\]

for the potential vector \(gL_m\) of states in mode \(m\).

(c) Find the improved policy \(L_{k+1}\) for the \((k+1)\)th iteration:

i. If \(L_{k+1}(n)\) haven’t been determined, take the maximum of \(\theta(n, n) L_{k+1}(n)\) \((n = 1, 2, \ldots, M\) and \(n \neq m)\) to determine an improved policy \(L_{k+1}(n)\) whenever applicable, the action of a mode in \(L_{k+1}(n)\) should remain the same in \(L_{k+1}(n)\) if the action is maximum for the mode in both \(L_{k+1}(n)\) and \(L_{k+1}(n)\).

ii. Find the improved policy \(L_{k+1}^u\) and \(L_{k+1}^l\) for the \((k+1)\)th iteration. (This step will be written below in detail.)

3. Stop if \(L_k = L_{k+1}\); otherwise set \(k = k + 1\) and return to step 2(a).

Before we describe the step 2(c)(ii) in Algorithm 1 in detail, we note that:

1. \(|A_u(i)|\) is the size of the action set \(A_u(i)\) for mode \(i\) in upper-level. The actions in \(A_u(i)\) are denoted as \(a_{1,1}, a_{2,1}, \ldots, a_{|A_u(i)|,1}\).

2. \(S_{mm}^{u}(m), L_{m}^{u}(m)\) is the stochastic complement of the block \(r_{mm}S^{(m)}\) under the policy \(L_{m}^{u}(m)\), \(L_{m}^{u}(i)\) \((i \neq m)\) and \(L_{m}^{u}(i)\) \((i \neq m)\).

3. When \(L_{m}^{u}(m)\) is fixed as the action \(\alpha\) in \(A_u(i)\), \(S_{mm, mk}\) is an irreducible stochastic matrix which can be controlled by the low-level policy \(L_{i}^{l}(m)\).

4. \(F_{m}^{L_{i}^{l}(m), L_{m}^{u}(m)}\) is the performance function of the matrix \(S_{mm}^{u}(m), L_{i}^{l}(m)\).

Step 2c(ii) in Algorithm 1 is stated in detail below:

**Algorithm 1.1** (Step 2(c)(ii) in Algorithm 1)

1. For each mode \(m\), set \(k_{A,m} = 1\) and choose \(a_{1}^{u,m}\) as the initial action for mode \(m\).

(a) Doing the standard policy iteration for \(S_{mm}^{u}(m), L_{i}^{l}(m)\) with performance function \(F_m^{L_{i}^{l}(m), L_{m}^{u}(m)}\) where \(L_{m}^{u}(m) = \alpha_{k_{A,m}}^{u,m}\). The potential for \(S_{mm}^{u}(m), L_{i}^{l}(m)\) can be computed by (2.48). Then we obtain the improved policy \(L_{k_{A,m}}^{u}(m)\) and the long-run average performance \(\eta_{k_{A,m}}^{u,m}\).
(b) If \( \eta^{u}_{k;A} > 0 \), then choose \( L_{k+1}^{u}(m) = \alpha_{k_{A,m}}^{u,k_{s}} L_{k+1}^{l}(m) = L_{k}^{l}(m) \); otherwise set \( k_{A,m} = k_{A,m} + 1 \) and return to step 2 in Algorithm 1.1.

2. Stop if \( L_k = L_{k+1} \); otherwise set \( k = k + 1 \) and return to step 2(a) in Algorithm 1.

The proof of Algorithm 1 is stated below.

**Proposition 2.4.7** Algorithm 1 terminates at an optimal policy for the two-level system in a finite number of iterations.

**Proof:** We know that at the \( k \)th iteration of Algorithm 1:

\[
\eta^{C} - \eta^{L} = \sum_{m=1}^{M} r_{m} s_{m} F_{m} L_{k}
\]

If there is a policy \( L' \) that is better than \( L_k \), then for some \( m \)

\[
s_{m} F_{m} L'_{l} L_{k} > 0
\]

We consider a new policy \( L'' \) where \( L'' \) and \( L' \) only have the same \( L''(m), L'(m) \) and \( L'(m) \) for the mode \( m \). And \( L'' \) and \( L_k \) only have the different \( L''(m), L'(m) \) and \( L'(m) \) for the mode \( m \). By Lemma 2.54, we also have

\[
s_{m} F_{m} L''_{l} L_{k} > 0
\]

So we can find the improved policy from Algorithm 1. Because the policy space is finite, the algorithm must terminate in a finite number of iterations.

The major computation involved in this algorithm consists of three parts: (i) Calculating the steady probability \( \pi^{L_{k}} \); (ii) Calculating the performance potential \( g^{L_{k}} \); (iii) The computation of Algorithm 1.1. The total computation is roughly of the order \( L_0(2M^3 + \sum_{m=1}^{M} N_m^3 + \sum_{m=1}^{M} L_m N_m^3) \), where \( L_0 \) denotes the number of iterations required for convergence \( L_m \) denotes the number of iterations needed to find a improved policy for the stochastic complement.
2.5 Conclusion

In this paper we give some results on the fundamental matrix and the stochastic complementation of a two-level markov process. By these results, it's possible to decouple the two-level system and obtain the global optimal policy for this system. An effective policy iteration algorithm is proposed. There are still some work to do in this direction. If this two-level system is a continuous markov process in high-level and low level, the markov decision problem is needed to be discussed.
Bibliography


