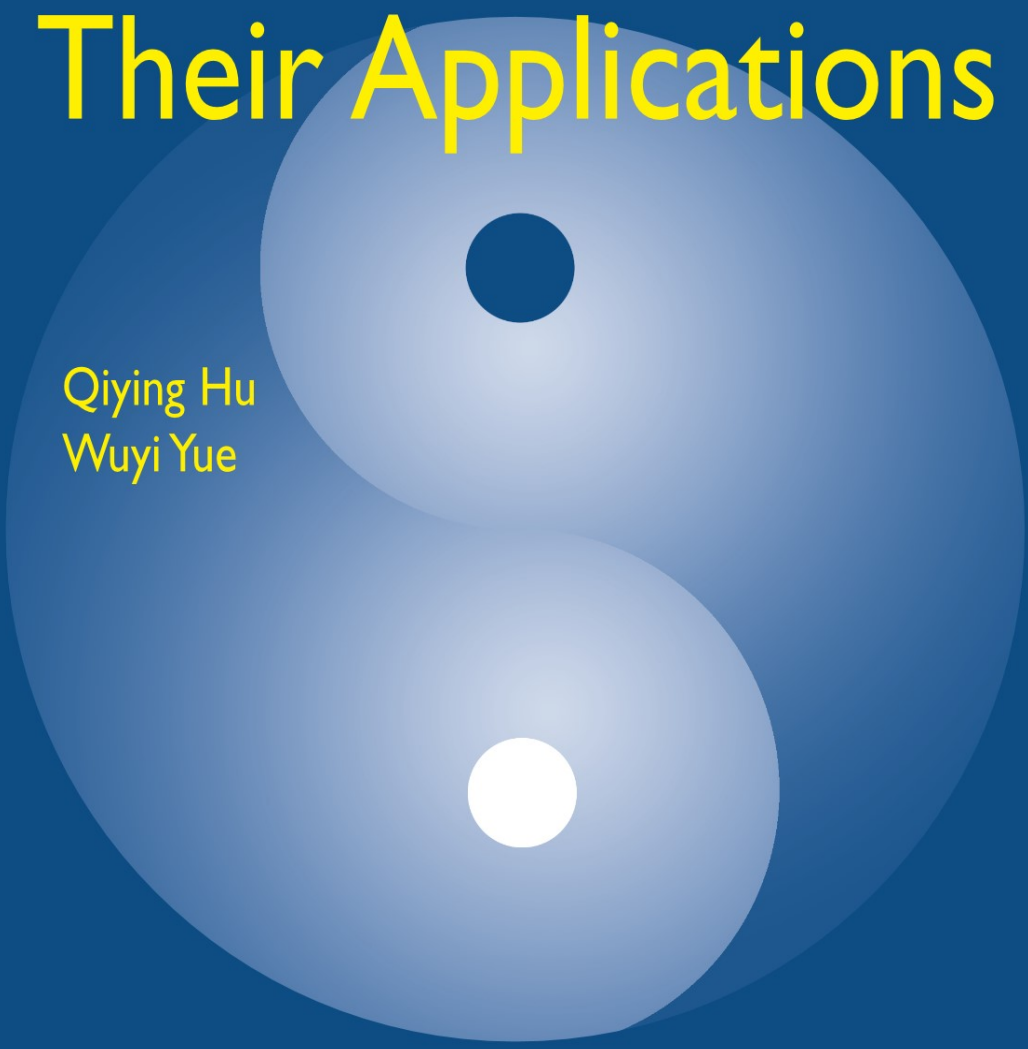

Markov Decision Processes with Their Applications

Qiying Hu
Wuyi Yue

Springer

MARKOV DECISION PROCESSES WITH THEIR APPLICATIONS

Advances in Mechanics and Mathematics

VOLUME 14

MARKOV DECISION PROCESSES WITH THEIR APPLICATIONS

By

Prof. Ph.D. Qiying Hu
Fudan University, China

Prof. Ph.D. Wuyi Yue
Konan University, Japan

Springer

Library of Congress Control Number: 2006930245

ISBN-13: 978-0-387-36950-1 e-ISBN-13: 978-0-387-36951-8

Printed on acid-free paper.

AMS Subject Classifications: 90C40, 90C39, 93C65, 91B26, 90B25

9 8 7 6 5 4 3 2 1

springer.com

Contents

List of Figures

List of Tables

Preface

Markov decision processes (MDPs), also called stochastic dynamic programming, were born in 1960s. MDPs model and solve dynamic decision-making problems with multi-periods under stochastic circumstances. There are three basic branches in MDPs: discrete time MDPs, continuous time MDPs, and semi-Markov decision processes. Based on these branches, many generalized MDP models were presented to model various practical problems, such as partially observable MDPs, adaptive MDPs, MDPs in stochastic environments, and MDPs with multiple objectives, constraints, or imprecise parameters. MDPs have been applied in many areas, such as communications, signal processing, artificial intelligence, stochastic scheduling and manufacturing systems, discrete event systems, management, and economics.

In this book, we mainly present three ideas for MDPs.

The first one is to present a new methodology for MDPs with a discounted total reward criterion. The usual methodology for MDPs is first to present a set of sufficient conditions, and then to show under the conditions the well definition of the model and the validity of the optimality equation together with its properties. Usually, different MDP models need different methods. This makes the research and the applications of MDPs more complex. Contrary to this, the methodology in this book is to show the validity of the optimality equation and its properties from the well definition of the model by reducing the scale of MDP models based on action reduction and state decomposition. The idea of the action reduction is that an action can be eliminated if any policy using it would not be optimal, whereas that of the state decomposition is to decompose the state space into several subspaces such that in each subspace an optimal policy can be obtained or the sub-MDP model can be easily solved. Thus, the original MDP model is decomposed into several smaller MDP models. The purpose of reducing the scale of the MDP model is mainly to separate the case with finite optimal value from the cases with positive or negative infinite optimal value, and then we can just study the case with finite optimal value.

It is difficult to deal with the optimality equation when the optimal value is infinite. The condition we need is that the model is well defined. Otherwise, we could not study MDP models. So, we call the condition a necessary one. Hence, when the model is well defined we can directly use the results, instead of proving them. By using the methodology above, we study a discrete time MDP model and a continuous time MDP model with the discounted total reward criterion under the necessary condition. Based on these, we present two new optimal control problems for discrete event systems and study them by using our methodology for MDPs.

The second idea of this book is the transformation for the continuous time MDPs and the semi-Markov decision processes. We transform them into equivalent discrete time MDPs both for the discounted total reward criterion and the average criterion. The equivalence is shown by basic algebraic computations. Then, we can directly use the results in the latter for the former two MDPs.

The systems modeled by the traditional MDPs are closed but many practical systems are not closed because they are influenced by their environments. Our third idea is MDPs in stochastic environments. This type of MDP can describe such a system that itself can be modeled by a Markov decision process, but the system is influenced by its environment. We study continuous time MDPs and semi-Markov decision processes in semi-Markov environments, and mixed MDPs in a semi-Markov environment. We use this type of MDP models to study two optimal replacement problems in stochastic environments.

Acknowledgments

The first author would like to thank Professors Tomoyasu Taguti, Hirotaka Nakayama, Hidetoshi Nakayasu, Masahiro Tanaka, Shigeki Matsumoto, and Atsushi Watanabe at Konan University, Kobe, Japan, Professor Shouyang Wang at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Bejing, China, and Professor Dinghua Shi at the Department of Mathematics, Shanghai University, China. The later author would like to thank Professor David Y. Gao at Virginia Polytechnic Institute and State University, USA. The authors would like to thank our coauthors, Professors Jianyong Liu, Chen Xu, Jinling Wang, and Dr. Li Du for several papers that are the basis of several chapters in this book. The authors are grateful for the editor arranging the publish of this book. The authors are also grateful for the support for the research which led to this book received from the Natural National Science Foundation, China, and GRANT-IN-AID FOR SCIENTIFIC RESEARCH, Japan for the Promotion of Science.

Chapter 1

INTRODUCTION

1. A Brief Description of Markov Decision Processes

Markov decision processes (MDPs), also called stochastic dynamic programming, have been studied extensively since they were first introduced in 1960 [55]. MDPs were mainly used to model and solve dynamic decision-making problems with multi-periods under stochastic circumstances.

The most basic type of MDPs are the discrete time Markov decision processes (DTMDPs for short). One of these is given as follows,

$$\{S, A(i), p_{ij}(a), r(i,a), V\}.$$

The system with state space S is observed at discrete time periods $n = 0, 1, \ldots$. When the system is observed to be at state $i \in S$, an action a from the action set $A(i)$ should be chosen. Then the following two things will happen: (a) the system will receive a reward $r(i,a)$, and (b) the system will transfer to state j at the next period with state transition probability $p_{ij}(a)$. V in the model is the criterion (or the objective), defined later. For simplicity, we suppose that the state space S and all the action sets $A(i)$ are countable here. We let $\Gamma = \{(i,a)|i \in S, a \in A(i)\}$ be the set of possible pairs of state and action at each period.

Let $A := \bigcup_{i \in S} A(i)$ be the union of all action sets. We define a decision function by a map $f : S \to A$ satisfying $f(i) \in A(i)$ for $i \in S$. It means that action $f(i)$ will be chosen whenever state i is observed. Let F be the set of all decision functions. We also write $F = \times_i A(i)$.

A policy for the system is a rule to determine actions that should be taken whatever the system's history is and whenever the observation period is. Formally, we let $H_n = \Gamma^{n-1} \times S$ be the set of history up to n for $n > 0$ and $H_0 = S$. We define a policy $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$ by: for any integer $n \geq 0$ and history $h_n = (i_0, a_0, \ldots, i_n) \in H_n, \pi_n(\cdot|h_n)$ is a probability distribution

on $A(i_n)$. Taking a policy π means that if a history h_n occurs at period n then the action is chosen according to the probability distribution $\pi_n(\cdot|h_n)$. If $\pi_n(\cdot \mid h_n) = \pi_n(\cdot \mid i_n)$ depends only on n and the last state i_n for each h_n, then we call π a Markov policy, the set of which is denoted by Π_m. Under Markov policies, the action is chosen irrespectively of the system's past. A special case of the type of Markov policies is that of stochastic stationary policies, which are defined as $\pi_n(\cdot|i_n) = \pi_0(\cdot|i_n)$ for each n. The set of all stochastic stationary policies is denoted by Π_s. On the other hand, if for each $n = 0, 1, \ldots$, there is $f_n \in F$ such that $\pi_n(f_n(i)|i) = 1$ for all $i \in S$ (in this case, we write $\pi = (f_0, f_1, \ldots)$), then such a policy π is called a deterministic Markov policy, the set of which is denoted by Π_m^d. A stationary deterministic Markov policy $\pi = (f, f, \ldots)$ for some $f \in F$ is simply called a stationary policy. We write it as f^∞ or f. Surely, a stationary policy corresponds to a decision function. So, we view F as the set of all stationary policies.

For $n \geq 0$, let X_n and Δ_n be the state and the action chosen at period n, respectively. Then it can be shown that the stochastic process $\{X_n, \Delta_n, n \geq 0\}$ is well defined under any policy $\pi \in \Pi$. Especially, under a Markov policy $\pi \in \Pi_m$, $\{X_n, \Delta_n, n \geq 0\}$ is a discrete time Markov chain [52]. For each $\pi \in \Pi$ and $i \in S$, let $P_{\pi,i}$ and $E_{\pi,i}$ be, respectively, the probability and the expectation corresponding to the stochastic process $\{X_n, \Delta_n, n \geq 0\}$ under policy π with the initial state i.

The system will receive a reward $r(X_n, \Delta_n)$ at period n. This reward is random under any policy. Then, how to compare different policies? This can be answered by the decision criteria. The basic criteria are the discounted expected total reward and the average criterion. The former applies to either finite horizons or infinite horizons whereas the latter only applies to infinite horizons.

Criterion 1. Discounted expected total reward in finite horizons.

This criterion is defined by

$$V_{\beta,N}(\pi, i) = \sum_{n=0}^{N-1} \beta^n E_{\pi,i} r(X_n, \Delta_n), \quad \pi \in \Pi, \quad i \in S.$$

Here N is a finite integer that represents the number of horizons and β is a positive constant that represents a discount factor. In general, $\beta \in (0, 1]$. If we write ρ as the interest rate in the market, then the discount factor and the interest rate have the following relationship,

$$\beta = \frac{1}{1+\rho}.$$

The meaning of β is that one unit reward at period n values β^n at period 0.

The optimal value function for this criterion is defined by

$$V_{\beta,N}(i) = \sup_{\pi\in\Pi} V_{\beta,N}(\pi, i), \ \ i \in S.$$

It is the best one can achieve from the initial state i with the discount factor β when there remain N horizons. We call a policy π^* N-optimal if $V_{\beta,N}(\pi^*, i) = V_{\beta,N}(i)$ for all $i \in S$.

Criterion 2. Discounted criterion/total reward criterion.

This criterion is defined by

$$V_\beta(\pi, i) = \sum_{n=0}^{\infty} \beta^n E_{\pi,i} r(X_n, \Delta_n), \ \ \pi \in \Pi, \ \ i \in S,$$

which is similar to the above criterion but in infinite horizons. So, it is really the discounted expected total reward in infinite horizons. Similarly, let the optimal value function be

$$V_\beta(i) = \sup_{\pi\in\Pi} V_\beta(\pi, i), \ \ i \in S$$

and a policy π^* is called discounted-optimal if $V_\beta(\pi^*, i) = V_\beta(i)$ for all $i \in S$.

In the literature, the criterion with the discount factor $\beta \in (0, 1)$ is often called the discounted criterion, and the criterion with $\beta = 1$ is called the total reward criterion. But in this book, we consider these two cases mainly in the same framework and so we call them uniformly the total reward critera.

Criterion 3. Average (reward) criterion.

The average criterion is for infinite horizons and is defined by

$$V(\pi, i) = \liminf_{N\to\infty} \frac{1}{N+1} V_{1,N}(\pi, i), \ \ \pi \in \Pi, \ \ i \in S,$$

which is the long-running average expected reward per period. In the above formula, $V_{1,N}(\pi, i)$ is exactly the $V_{\beta,N}(\pi, i)$ when the discount factor $\beta = 1$. Let the optimal value function be

$$V^*(i) = \sup_{\pi\in\Pi} V(\pi, i), \ \ i \in S$$

and a policy π^* be called average-optimal if $V(\pi^*, i) = V^*(i)$ for all $i \in S$.

The three criteria above are used very often in the literature of MDPs. Other criteria include the discounted moment criterion [65], Blackwell criterion [5], utility criterion by using utility functions [146], the mixed criterion of the discounted reward criterion, and the average reward criterion [25]. The criteria discussed in this book include the total reward criterion and the average criterion.

There are three basic types of MDPs: discrete time MDPs [5], continuous time MDPs (CTMDPs) [82], and semi-Markov decision processes (SMDPs) [56], and [81]. As presented in the discussion above, DTMDPs are based on discrete time Markov chains. Similarly, CTMDPs are based on continuous time Markov chains and SMDPs are based on semi-Markov processes. In addition, based on the basic types of MDPs, several generalized MDPs were presented in the literature, such as partially observable MDPs [18], adaptive MDPs [109], and constrained MDPs [50]. Interested readers may refer to a survey paper [148] or a handbook [39].

Optimal control of other types of Markov chains, such as diffusion processes, is mainly part of optimal control theory and is also studied extensively, see, for example, [24], [11], and [43]. Recently, a new model called the hybrid system has been presented. This model combines event-driven dynamics and time-driven dynamics, for example, see [26]. In Chapter 6 of this book, we study MDPs in stochastic environments, where the influence of environments on systems is considered.

MDPs has been applied to many areas that include communications (dynamic routing problems, multiple-access problems [96] and [33], flow control [13] and [112], artificial intelligence [16], stochastic scheduling and dynamic control of manufacturing systems [32] and [84], discrete event systems [17], management (such as optimal replacement, production/inventory [2], product pricing [102] and [34]), and finance (dynamic asset pricing [31]). One can see some applications in books [69] and [124] and a survey paper [147].

2. Overview of the Book

The standard results for MDP models include the following four aspects.

1. The model is well defined; that is, the stochastic process under consideration is well defined (or regular in some cases). Moreover, the criterion is well defined and often is finite.
2. The optimal value function satisfies the optimality equation.
3. A stationary policy achieving the supremum of the optimality equation will be optimal.
4. Algorithms to compute the optimal value function/approximating optimal policies are presented.

The main methodology for studying a MDP model to obtain the standard results is as follows.

- First, a set of conditions for the model is presented under which the model is well defined.
- Second, the optimal value function is shown to be a solution or the unique solution of the optimality equation in a certain region (then we say that the

optimality equation is true). For example, the optimality equation for the total reward criterion is as follows,

$$V_\beta(i) = \sup_{a \in A(i)} \{r(i,a) + \beta \sum_j p_{ij}(a) V_\beta(j)\}, \quad i \in S.$$

- Third, any policy that achieves the supremum in the optimality equation is shown to be optimal, or more generally, any stationary policy achieving the ε-supremum of the optimality equation will be ε'-optimal, where ε' is a function of ε and tends to zero when ε tends to zero.
- Finally, algorithms may be presented to obtain the optimal value and an (approximate) optimal policy. However, the basic algorithms in MDPs are just successive approximation, policy improvement, and linear programming.

In the literature, to ensure that the model is well defined, some conditions are often imposed. For the total reward criterion, there are three classical cases: (a) the discount factor belongs to (0, 1) and the rewards are uniformly bounded, (b) the discount factor is one and the rewards are nonnegative, and (c) the discount factor is one and the rewards are nonpositive. Usually, MDP models with cases (a), (b), and (c) are called discounted MDP models, positive MDP models, and negative MDP models, respectively. But these three cases are too strong. In order to weaken them, various conditions are presented in the literature to suit various practical problems, especially for discrete time MDP models and semi-Markov decision process models. For example, for the discounted criterion, Lippman [92] presented a set of conditions on the unbounded reward functions for a SMDP model, whereas Harrison [49] and Wessels [145] presented conditions for DTMDP models. Hu and Hu [80] combined the conditions presented by the above three authors and presented a weaker one for a DTMDP model. All these conditions are sufficient conditions for studying MDP models.

Conversely to the above methodology, in this book, we try to study MDPs under the condition that the model is well defined. More precisely, we try to show the standard results 2, 3, and 4 above by assuming the standard result 1. This is explored in Chapters 2 and 4 for DTMDPs and CTMDPs, respectively. Moreover, this methodology is applied to study optimal control problems in discrete event systems in Chapters 7 and 8. This is the first idea of this book.

The second idea of this book is to transform systematically the CTMDPs and SMDPs into DTMDPs for the discounted criterion, the total reward criterion, and the average criterion. In the literature, main studies on CTMDPs, SMDPs, and DTMDPs are done separately, though some transformation were presented. Schweitzer [120], Hordijk et al. [54], and Federgruen and Tijms [36] presented transformations from SMDPs into DTMDPs for the average criterion. For CTMDP models, Serfozo [126] presented a transformation for the

discounted criterion with bounded transition rates among stationary policies, based on probability properties of Markov chains. Hu [57] presented a transformation for the discounted criterion with unbounded transition rates. Under his transformation, the corresponding optimality equation and the discounted criterion among stationary policies in the CTMDP model and those in the DTMDP model are equivalent. So the results for CTMDPs can be obtained directly from those for DTMDPs. In this book, we focus on the method of transformation for the discounted criterion, the total reward criterion, and the average criterion. It seems that this is the first time CTMDPs and SMDPs have been studied systematically by transforming them into DTMDPs. The transformations are based only on basic algebraic computations. This idea is contained in Section 3 of Chapter 4 for CTMDPs, Chapter 5 for SMDPs, and also appears for MDPs in stochastic environments in Chapter 6.

The third idea is to consider the influence of environments on systems. Systems described by the traditional MDP models are all closed in the meaning that no influence of the environments to the systems is considered. But in practice, many systems are influenced by their environments. Some other areas had considered the influence of the environments, for example, Neuts [98] for queueing systems and Cao [15] for reliability systems. We present MDP models in stochastic environments in this book. These models can describe such a system that itself can be modeled by a Markov decision process, but the system is influenced by its environment which is modeled by a semi-Markov process. The influence includes changing the MDP model or its parameters, inducing an instantaneous state transition of the system, and letting the system incur a reward. This idea is explored in Chapter 6. We also apply this type of MDP model to describe and prove optimal replacement problems in Chapter 9.

3. Organization of the Book

The rest of the book is organized as follows.

In Chapter 2, we study a discrete time MDP model with the total reward criterion, where the state space is countable, the action set is arbitrary but nonempty and is endowed with a measurable structure, the reward function is extended real-valued, and the discount factor is any real number. The condition that the model is well defined here is just that the criterion, as a series, is well defined under each policy and each initial state.

We first show that the state space can be divided into two parts. In the first part the optimal value is positive infinite and there is an optimal stochastic stationary policy, and in the second part the reward function is finite and bounded above over the action set at each state. Hence, this normalizes the original model, where the reward function is extended real-valued, into a submodel, where the reward function is real-valued and bounded above over the action set at each state. Thus, it suffices to discuss the MDP model in the second part. In this

part, the optimality equation is shown to be true if its right-hand side is well defined. Otherwise, by eliminating some worst actions, the state space can further be decomposed into four subsets: in the first subset, the optimal value is negative infinity and so each policy is optimal; in the second subset, the optimal value is positive infinity and there is an optimal policy; in the third subset, the optimal value is positive infinity but there is no optimal policy; and in the final subset, the optimal value is finite and the right-hand side of the corresponding optimality equation is well defined and so the optimality equation in this subset is true.

Based on the above results, the remainder of Chapter 2 is discussed under the condition that the optimal value is finite and satisfies the optimality equation. We characterize the optimal value as a solution of the optimality equation and study the optimality of a policy that achieves the supremum of the optimality equation or that its criterion value satisfies the optimality equation. Also we give a structure of the set of all optimal policies. Moreover, we discuss successive approximation. Finally, we give some sufficient conditions for the necessary condition that the model be well defined.

In Chapter 3, we study the average criterion for DTMDPs. First, we introduce some lemmas from the theory of Markov chains and mathematical analysis. Then, we study the optimality equation together with its properties. New conditions are presented for them. Finally, we study optimality inequalities, which need weaker conditions than the optimality equation.

In Chapter 4, we first apply the ideas and method presented in Chapter 2 for DTMDP models to study a CTMDP model with the total reward criterion, where the state space and all the action sets are countable. We focus our attention on the expected total reward for a stationary model. Similar results to those in Chapter 2 are obtained, although more properties from continuous time Markov chains are needed in the proof. Then, we deal with a nonstationary model with the total reward criterion. By dividing the time axis into shorter intervals, we obtain standard results, such as the optimality equation and the relationship between the optimality of a policy with the optimality equation. Finally, we study the average criterion for a stationary CTMDP model by transforming it into a DTMDP model.

In Chapter 5, we study a semi-Markov decision process model. For a stationary SMDP model, we transform it into a stationary DTMDP model for the discounted criterion, the total reward criterion, or the average criterion. Hence, all results for DTMDPs (e.g., in Chapter 2 and Chapter 3) can be used directly for SMDPs.

In Chapter 6, we deal with MDPs in semi-Markov environments with the discounted criterion. The model can describe such a system that itself can be modeled by a Markov decision process, but the system is influenced by its environment which is modeled by a semi-Markov process. And according to each

change of the environment's states, three things occur: (1) an instantaneous state (of the system) transition, (2) an instantaneous reward, and (3) the parameters of the Markov decision process changes. We first study CTMDPs in a semi-Markov environment and then SMDPs in a semi-Markov environment. Based on them, we study a mixed Markov decision process model in a semi-Markov environment, where the underlying MDP model can be either CTMDP or SMDP according to which environment states are entered. The mixed MDPs generalize the CTMDPs and SMDPs in semi-Markov environments.

In Chapters 7 and 8, we present two new models for optimal control of discrete event systems (DESs) by combining models and ideas in MDPs and supervisory control of DESs. There is no formal model for optimal control of DESs. The ideas and methods presented in Chapter 2 are applied to study the models and similar results (such as the decomposition and the optimality equation) to those in Chapter 2 are obtained.

In Chapter 7, the reward is for occurrence of an event. Moreover, the basic supervisory control problem for DESs is fitted in the framework of the model. Based on it, we establish some links between the supervisory control of DESs and our model. Finally, we apply the model to the resource allocation of a system.

In Chapter 8, the reward is for choosing a control input. Moreover, we present and study supervisory control of the DESs with an arbitrary control pattern, and we obtain some new results for the supervisory control of DESs. Finally, we apply the model to a job-matching problem.

In Chapter 9, we study two optimal replacement problems under stochastic environments, as applications of MDPs in stochastic environments discussed in Chapter 6. The first one is for discrete time. Here, the system is modeled by a discrete time Markov chain and the influence of the environment on the system is modeled by a Poisson process. The second one is for continuous time, where both the system and the environment are modeled by semi-Markov processes and each state change of the environment will change parameters of the system's model. We study them by applying DTMDPs and SMDPs in stochastic environments, respectively, to them. Based on the optimality equations, we discuss monotone properties of the optimal values and show the existence of optimal extended control policies for both problems with the discounted criteria. We also show for both problems that under certain conditions, the optimal replacement problems with infinite system states can be reduced to those with finite system states. Finally, a numerical example is given to illustrate the problems.

In Chapter 10, we study optimal allocation in a sequential Internet auction system with a set reserve price. In the sequential Internet auction system, a seller wants to sell a given amount of items through sequential auctions on the Internet. The seller has a reserve price for each item. For each auction, the seller should allocate a quantity of items from the total available items to be

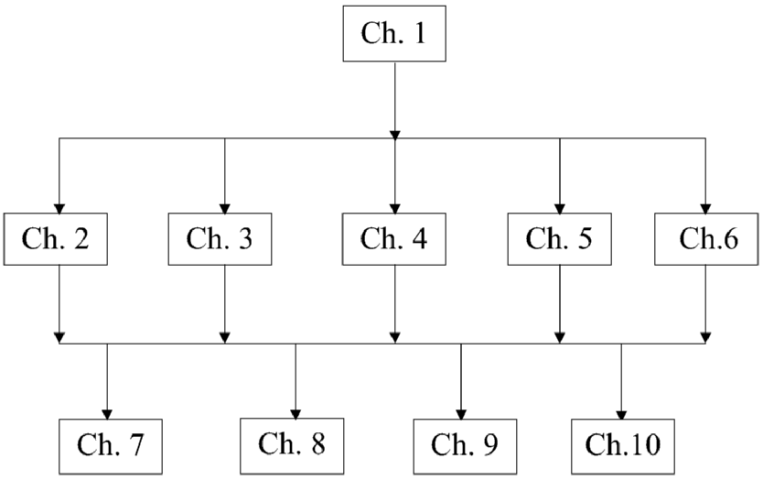

Figure 1.1. The flow chart of the chapters.

auctioned. The buyers arrive according to a Poisson process and bid honestly. We first consider the model to be a Markov decision process. We show that the result is not different whether the reserve price is private or public. Then we show monotonous properties of the optimal value and the optimal policy. Finally, numerical results are given.

The flow chart of the chapters in the book is given in Figure 1.1. But the contents of each chapter from Chapter 2 to Chapter 6 are self-closed. So, readers can read each chapter after having basic knowledge of MDPs, for example, from Chapter 1. Chapters 2 to 6 consist of the theoretical part of this book, and Chapters 7 to 10 consist of the second part: applications. In this part, Chapters 7 and 8 are based on Chapter 2, the problems discussed in Chapter 9 are applications of those in Chapter 6, and the optimal allocation problem discussed in Chapter 10 is an application of those in Chapter 2.

Chapter 2

DISCRETE TIME MARKOV DECISION PROCESSES: TOTAL REWARD

This chapter studies a discrete time Markov decision process with the total reward criterion, where the state space is countable, the action sets are measurable, the reward function is extended real-valued, and the discount factor $\beta \in (-\infty, +\infty)$ may be any real number although $\beta \in [0, 1]$ used to be required in the literature. Two conditions are presented, which are necessary for studying MDPs and are weaker than those presented in the literature. By eliminating some worst actions, the state space S can be partitioned into subsets $S_\infty, S_{-\infty}, S_0$, on which the optimal value function equals $+\infty, -\infty$, or is finite, respectively. Furthermore, the validity of the optimality equation is shown when its right-hand side is well defined, especially, when it is restricted to the subset S_0. The reward function $r(i, a)$ becomes finite and bounded above in a for each $i \in S_0$. Then, the optimal value function is characterized as a solution of the optimality equation in S_0 and the structure of optimal policies is studied. Moreover, successive approximation is studied. Finally, some sufficient conditions for the necessary conditions are presented. The method we use here is elementary. In fact, only some basic concepts from MDPs and discrete time Markov chains are used.

1. Model and Preliminaries

1.1 System Model

The model of the discrete time Markov decision processes discussed in this chapter is

$$\{S, (A(i), \mathcal{A}(i)), p_{ij}(a), r(i, a), V_\beta\},$$

where the state space S is countable, for $i \in S$, the action set $A(i)$, available at state i, is nonempty, $(A(i), \mathcal{A}(i))$ is a measurable space, and each single point set of $A(i)$ is measurable. When the system is in state i and an action $a \in A(i)$

is taken at some period, the system will transfer to state j at the next period with probability $p_{ij}(a)$ and incur an extended real-valued reward $r(i,a)$. We assume that both $p_{ij}(a)$ and $r(i,a)$ are measurable in a for any $i,j \in S$. The policies are defined in Chapter 1.

For any real x, put $x^{\pm} = \max\{0, \pm x\}$. We define the expectation of any random variable X by $E(X) = E(X^+) - E(X^-)$ if either $E(X^+)$ or $E(X^-)$ is finite. Also we say that a series $\sum_j c_j = \sum_j c_j^+ - \sum_j c_j^-$ is well defined if $\sum_j c_j^+$ or $\sum_j c_j^-$ is finite.

Let X_n, Δ_n denote the state and the action taken (by the system) at period n. The criterion discussed in this chapter is the total reward:

$$V_\beta(\pi, i) = \sum_{n=0}^{\infty} \beta^n E_{\pi,i} r(X_n, \Delta_n), \quad i \in S, \quad \pi \in \Pi, \tag{2.1}$$

where $\beta \in (-\infty, +\infty)$ is a given discount rate. In the literature, the discount rate $\beta \in [0,1]$ is often assumed. However, $\beta > 1$ means the situation of inflation with the interest rate $\rho = 1/\beta - 1$ is negative, and the negative discount rate is only mathematical because the method we used can also deal with it. But the negative discount rate will influence the satisfaction of Conditions 2.1 and 2.2 given below. For $V_\beta(\pi, i)$, a necessary condition should be the following one, which is the basis for discussing the MDP model.

Condition 2.1: *$V_\beta(\pi, i)$ is well defined for all $\pi \in \Pi$ and $i \in S$.*

It should be noted that the above condition implies that (1) for each policy π, state i and integer $n \geq 0$, $E_{\pi,i}\, r(X_n, \Delta_n)$ are well defined (may be infinite), and (2) as a series, $V_\beta(\pi, i)$ is well defined (also may be infinite). We say that the MDP model is well defined if Condition 2.1 is true.

Surely, we cannot discuss the MDP model if it is not well defined. Hence, Condition 2.1 is a necessary condition and is assumed throughout this chapter.

Let the optimal value function $V_\beta(i) = \sup\{V_\beta(\pi, i) | \pi \in \Pi\}$ for $i \in S$. For $\varepsilon \geq 0, \pi^* \in \Pi$ and $i \in S$, if $V_\beta(\pi^*, i) \geq V_\beta(i) - \varepsilon$ (when $V_\beta(i) < +\infty$) or $\geq 1/\varepsilon$ (when $V_\beta(i) = +\infty$), then π^* is called ε-optimal at state i. Here, $1/0 = +\infty$ is assumed. If π^* is ε-optimal at all $i \in S$ then π^* is called ε-optimal. An 0-optimal policy is simply called an optimal policy.

1.2 Some Concepts

We introduce some concepts in this subsection.

Definition 2.1: *State j can be reached from state i if there are a policy π and an integer $n \geq 0$ such that $P_{\pi,i}\{X_n = j\} > 0$, which is denoted by $i \overset{\pi,n}{\to} j$, or $i \to j$ for short. For a state subset $S_0 \subset S$, if there is $j \in S_0$ such that $i \to j$, then we say that S_0 can be reached from i and denote*

it by $i \to S_0$. Similarly, we define $S_0 \to i$ if there is state $j \in S_0$ such that $j \to i$.

For a subset $S_0 \subset S$, let $\overline{S_0} = \{j | S_0 \to j\}$ and $S_0^* = \{j | j \to S_0\}$. $\overline{S_0}$ is the set of states that can be reached from S_0 and S_0^* is the set of states that can reach S_0. Because $n = 0$ is permitted in the definition, we have $i \to i$ for each $i \in S$. So, $S_0 \subset S_0^*$ and $S_0 \subset \overline{S_0}$.

Definition 2.2: *(1) A state subset S_0 of S is called a closed set, if $p_{ij}(a) = 0$ for all $i \in S_0, j \notin S_0$, and $a \in A(i)$, equivalently, $\overline{S_0} = S_0$, or $(S - S_0)^* = S - S_0$.*

(2) For a closed set S_0, the restriction of the MDP model to S_0 is defined by

$$S_0\text{-}MDPs := \{S_0, (A(i), \mathcal{A}(i), i \in S_0), p_{ij}(a), r(i,a), V_\beta^{S_0}\}$$

which is called the sub-MDP model induced by S_0.

For a closed set S_0, let $H_n(S_0)$ be the history set up to n for S_0-MDPs. Then $H_n(S_0) \subset H_n$ for $n \geq 0$. For any policy $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$, restricting $\pi_n(\cdot|h_n)$ to $h_n \in H_n(S_0)$ will result in a policy of S_0-MDPs, which will also be denoted by π. We denote the criterion of S_0-MDPs by $V_\beta^{S_0}(\pi, i)$. It is obvious that for any closed set $S_0 \subset S, V_\beta(\pi, i) = V_\beta^{S_0}(\pi, i)$ for all $\pi \in \Pi$ and $i \in S_0$; that is, the MDP model is equivalent to the induced S_0-MDPs in the closed subset S_0. So, if both S_0 and $S - S_0$ are closed, then the MDP model can be partitioned into two parts: S_0-MDP and $(S - S_0)$-MDP. On the other hand, if S_0 is closed and $V_\beta(i)$ for $i \in S - S_0$ is known, or an $(\varepsilon$-$)$optimal policy can be obtained in $S - S_0$, then one can discuss only S_0-MDPs. Thus the state space is decomposed.

Corresponding to decomposing the state space defined in Definition 2.2, there is the possibility of eliminating actions.

Definition 2.3: *Suppose that $A_1(i) \in \mathcal{A}(i)$ for all $i \in S$. We denote by $V'_\beta(\pi, i)$ the total reward criterion for the new Markov decision process model with $A(i)$ being replaced by $A_1(i)$ (where $p_{ij}(a)$ and $r(i,a)$ are restricted to $A_1(i)$). If for any policy π of the (original) MDP model there is a policy π' of the new MDP model such that*

$$V_\beta(\pi, i) \leq V'_\beta(\pi', i), \quad i \in S, \tag{2.2}$$

then we say that $A(i)$ can be sized down to $A_1(i)$ for $i \in S$ (i.e., all actions belonging to $A(i) - A_1(i)$ can be eliminated).

Certainly, the history set of the new MDP model H'_n is included in H_n for each n. Denote the restriction of $\pi \in \Pi$ to H'_n by $l(\pi) \in \Pi'$. If $P_\pi(\bigcup_n H'_n) = 1$, then the equality in (2.2) holds for $\pi' = l(\pi)$. So, the optimal value function of

the original MDP model equals that of the new MDP model and the (ε-)optimal policies can be taken to be the same. Thus we can solve the original MDP model by solving the new MDP model, which is simpler because its action sets are smaller than those of the original MDP model.

Decomposing the state space and eliminating worst actions are the main methods to prove the validity of the optimality equation in the following.

1.3 Finiteness of the Reward

In this subsection, we show that the reward function $r(i,a)$ can be taken to be finite and bounded above in $a \in A(i)$ for each $i \in S$. First, we prove the following lemma.

Lemma 2.1: *For any series $\{r_n, n \geq 1\}$ with $r_n \geq n$, there exists a series $\{c_n \geq 0\}$ such that $\sum_n c_n = 1$ and $\sum_n c_n r_n = +\infty$.*

Proof: Let $\delta \in (0,1)$ and $c = \sum_n n^{-(1+\delta)} < \infty$. Then $c_n = cn^{-(1+\delta)}$ satisfies the lemma. □

For two policies $\pi = (\pi_0, \pi_1, \ldots)$ and $\sigma = (\sigma_0, \sigma_1, \ldots)$, their linear combination $d_1\pi + d_2\sigma$, for $d_1, d_2 \in [0,1]$ with $d_1 + d_2 = 1$, is a policy $(d_1\pi_0 + d_2\sigma_0, d_1\pi_1 + d_2\sigma_1, \ldots)$ defined by

$$(d_1\pi_n + d_2\sigma_n)(\cdot|h_n) = d_1\pi_n(\cdot|h_n) + d_2\sigma_n(\cdot|h_n), \quad n \geq 0.$$

For $i \in S$, let

$$U(i) = \sup\{r(i,a)|a \in A(i)\}, \quad L(i) = \inf\{r(i,a)|a \in A(i)\}$$

be the supremum and the infimum of the reward function $r(i,a)$ in the action set $a \in A(i)$, respectively. For the infinity of $U(i)$ and $L(i)$, we have the following lemma.

Lemma 2.2:

1. *For $i \in S$ with $U(i) = +\infty$, there is a policy $\pi_0 \in \Pi_s$ such that $E_{\pi_0,i} r(X_0, \Delta_0) = +\infty$.*
2. *For $i \in S$ with $L(i) = -\infty$, there is a policy $\pi_0 \in \Pi_s$ such that $E_{\pi_0,i} r(X_0, \Delta_0) = -\infty$.*
3. *For $i \in S$, $L(i) = -\infty$ and $U(i) = +\infty$ cannot be true simultaneously.*

Proof: 1. If there is $a \in A(i)$ such that $r(i,a) = +\infty$, then the result is obvious. Otherwise, there are actions $a_n \in A(i)$ for $n \geq 1$, which are different from each other, such that $r(i,a_n) \geq n$ for $n \geq 1$. Let c_n be as in Lemma 2.1 and thus the following policy π_0 is required,

$$\pi_0(a|i) = \begin{cases} c_n, & \text{if } a = a_n, n \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

2. can be proved similarly.

3. If $L(i) = -\infty$ and $U(i) = +\infty$ for some $i \in S$, let $\pi_0^{(1)}$ and $\pi_0^{(2)}$ be, respectively, the policies in 1 and 2. Then for the policy $\pi_0 = 0.5\pi_0^{(1)} + 0.5\pi_0^{(2)}$, $E_{\pi_0,i} r(X_0, \Delta_0) = 0.5 E_{\pi_0^{(1)},i} r(X_0, \Delta_0) + 0.5 E_{\pi_0^{(2)},i} r(X_0, \Delta_0)$ is undefined, which contradicts Condition 2.1. □

Lemma 2.2 implies that for each $i \in S$, there are no actions a_1 and a_2 such that $r(i, a_1) = +\infty$ and $r(i, a_2) = -\infty$. That is, $r(i, a) > -\infty$ for all $a \in A(i)$ or $r(i, a) < +\infty$ for all $a \in A(i)$.

In the following, we prove that only those states i with $U(i) < +\infty$ need to be considered. Let

$$S_U = \{i | U(i) = +\infty\}, \;\; S_L = \{i | L(i) = -\infty\},$$

$$W = \{i \mid \text{there are } \pi \in \Pi \text{ and } n \geq 0 \text{ such that } E_{\pi,i} r(X_n, \Delta_n) = +\infty\}.$$

It is apparent that

$$W = \{i \mid \text{there is } \pi_0 \in \Pi_s \text{ such that } E_{\pi_0,i} r(X_n, \Delta_n) = +\infty\}.$$

From Lemma 2.2, the sets S_U and S_L are disjoint and $S_U \subset W$. It is easy to see that for $i \in W$, there is $\pi \in \Pi$ such that $V_\beta(i) = V_\beta(\pi, i) = +\infty$.

The following two lemmas discuss properties of the state subsets W, S_U, and S_L.

Lemma 2.3:

1. *$W^* = W$, so $S - W$ is closed.*

2. *If $j \to S_L$, then there is a policy π such that $V_\beta(\pi, j) = -\infty$.*

Proof: 1. If $j \to i \in W$, then there is a policy σ and an integer $m \geq 0$ such that $j \overset{\sigma,m}{\to} i$. Let $\pi \in \Pi$ and $n \geq 0$ such that $E_{\pi,i} r(X_n, \Delta_n) = +\infty$. We define a policy by $\pi^* = (\sigma_0, \sigma_1, \ldots, \sigma_{m-1}, \pi)$. Then from Condition 2.1, $j \in W^+$ due to

$$E_{\pi^*,j} r(X_{m+n}, \Delta_{m+n}) = \sum_k P_{\sigma,j}\{X_m = k\} E_{\pi,k} r(X_n, \Delta_n) = +\infty.$$

2. can be proved similarly. □

Lemma 2.4: $S_L \bigcap \overline{S_U} = \overline{S_L} \bigcap S_U = S_L^* \bigcap S_U^* = \emptyset$.

Proof: Suppose that $i \in S_L \bigcap \overline{S_U}$. Then, from the definition there are $j \in S_U$, a policy $\sigma = (\sigma_0, \sigma_1, \ldots)$, and $n \geq 0$ such that $j \overset{\sigma,n}{\to} i$. But from Lemma 2.2 there are policies π_0 and π_0' such that

$$E_{\pi_0',i} r(X_0, \Delta_0) = -\infty, \;\; E_{\pi_0,j} r(X_0, \Delta_0) = +\infty$$

and $\pi_0(a_0|j) > 0$. By taking $\pi = (\pi_0^*, \sigma_1, \ldots, \sigma_{n-1}, \pi_0', \pi_{n+1}, \ldots)$ with $\pi_0^* = 0.5\pi_0 + 0.5\sigma_0$, we have that

$$E_{\pi,j}r(X_0, \Delta_0) = 0.5E_{\pi_0,j}r(X_0, \Delta_0) + 0.5E_{\sigma_0,j}r(X_0, \Delta_0) = +\infty,$$
$$E_{\pi,j}r(X_n, \Delta_n) = \sum_l P_{\pi,j}\{X_n = l\}E_{\pi_0',l}r(X_0, \Delta_0) = -\infty,$$

where the last equality holds for $P_{\pi,j}\{X_n = i\} \geq P_{\sigma,j}\{X_n = i\} > 0$. So $V_\beta(\pi, j)$ is undefined, which results in a contradiction. Thus, $S_L \bigcap \overline{S_U} = \emptyset$.

$\overline{S_L} \bigcap S_U = \emptyset$ can be proved similarly.

Now we suppose that $j \in S_L^* \bigcap S_U^*$, that is, there are $i \in S_L, i' \in S_U$, policies π, σ, and integers $n, m \geq 0$ such that $j \overset{\pi,n}{\to} i$ and $j \overset{\sigma,m}{\to} i'$. It follows from Lemma 2.2 that there are π_0 and π_0' such that

$$E_{\pi_0,i}r(X_0, \Delta_0) = -\infty, \;\; E_{\pi_0',i'}r(X_0, \Delta_0) = +\infty.$$

(a) If $m \neq n$, then we construct a policy $\pi^* = (\pi_0^*, \pi_1^*, \ldots)$ by

$$\pi_k^* = \begin{cases} d\pi_k + (1-d)\sigma_k, & k \neq m, n, \\ (d/2)\pi_k + ((1-d)/2)\sigma_k + (1/2)\pi_0, & k = n, \\ (d/2)\pi_k + ((1-d)/2)\sigma_k + (1/2)\pi_0', & k = m \end{cases}$$

for some constant $d \in (0, 1)$. Thus, $j \overset{\pi^*,n}{\to} i$, $j \overset{\pi^*,m}{\to} i'$, and $E_{\pi^*,j}r(X_m, \Delta_m) = +\infty$, $E_{\pi^*,j}r(X_n, \Delta_n) = -\infty$. So, $V_\beta(\pi^*, j)$ is undefined.

(b) If $m = n$, we construct a policy π^* such that $\pi_k^* = d\pi_k + (1-d)\sigma_k, k = 0, 1, \ldots, m-1$ for some constant $d \in (0, 1)$ and $\pi_m^* = c\pi_0 + (1-c)\pi_0'$ for some constant $c \in (0, 1)$. Then, $j \overset{\pi^*,n}{\to} i$ and $j \overset{\pi^*,n}{\to} i'$ and so $E_{\pi^*,j}r(X_n, \Delta_n)$ is undefined. Hence, $V_\beta(\pi^*, j)$ is undefined.

Overall, $S_L^* \bigcap S_U^* = \emptyset$. □

Having the above lemmas, we now show the following main theorem of this section.

Theorem 2.1:

1. *There is a policy π that is optimal in W; that is, $V_\beta(\pi, i) = V_\beta(i) = +\infty$ for $i \in W$.*

2. *$A(i)$ can be sized down to*

$$A_1(i) = \{a \in A(i) \mid r(i, a) > -\infty\}, \;\; i \in S.$$

So, about $(S - W)$-MDPs, one has that

$$-\infty < r(i, a) \leq U(i) < +\infty, \;\; \forall (i, a). \tag{2.3}$$

Proof: 1. This can be obtained immediately from Lemma 2.3.

2. For any policy π, states i, j, and integer $n \geq 0$, if $i \overset{\pi,n}{\to} j$ and $\pi_n(A(j) - A_1(j)|h_n) > 0$ for all $h_n = (i, \ldots, j) \in H_n^*$ with $P_\pi(H_n^*) > 0$ for some subset H_n^* of H_n, then $E_{\pi,i} r(X_n, \Delta_n) = -\infty$. So, $V_\beta(\pi, i) = -\infty$; that is, such a policy π will never be optimal unless $V_\beta(i) = -\infty$. So, $a \in A(i) - A_1(i)$ can be eliminated from $A(i)$. About $(S - W)$-MDPs, it is apparent that the reward $r(i, a)$ satisfies the condition given in Eq. (2.3). □

The above theorem says that the whole state space S can be partitioned into two parts W and $S - W$ and there is an optimal policy in W, whereas $S - W$ is closed (Lemma 2.3) and the reward function $r(i, a)$ is finite and bounded above in a for each $i \in S - W$ after eliminating some worst actions. Hence, the reward is modified from extended real-valued into finite and bounded above.

From Theorem 2.1, it is assumed in the following sections that the reward function is finite and bounded above over actions.

2. Optimality Equation

In this section, we discuss the optimality equation, including its validity and properties.

2.1 Validity of the Optimality Equation

In order to get the optimality equation, we now give the second condition.

Condition 2.2: *For any policy* $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$ *and state* $i \in S$,

$$V_\beta(\pi, i) = \int_{A(i)} \pi_0(da|i)\{r(i, a) + \beta \sum_j p_{ij}(a) V_\beta(\pi^{i,a}, j)\}, \tag{2.4}$$

where $\pi^{i,a} = (\sigma_0, \sigma_1, \ldots) \in \Pi$ *with* $\sigma_n(\cdot|h_n) = \pi_{n+1}(\cdot|i, a, h_n)$ *for* $n \geq 0$.

The above condition states that the total process under any policy π can be partitioned into two subprocesses: the first period and the remaining periods. This partition condition is the essence of the optimality equation. Hence, it is necessary to show the validity of the optimality equation. Condition 2.2 implies that the series $\sum_j$ and integration $\int_{A(i)}$ in Eq. (2.4) are well defined. Condition 2.2 is shown in the literature usually under conditions that $r(i, a)$ is nonnegative, nonpositive, or satisfies some bounded conditions, and so on (e.g., see [80], [58], and [92]). In Section 5 we give some sufficient conditions for the above condition. We assume that Condition 2.2 is true throughout this section.

Let

$$S_\infty := \{i|V_\beta(i) = +\infty\}, \quad S_{-\infty} := \{i|V_\beta(i) = -\infty\}$$

and

$$S_0 := S - S_\infty - S_{-\infty}$$

be state subsets of positive infinite, negative infinite, and finite optimal values, respectively. Moreover, let $S_{=\infty} := \{i|$ there is $\pi \in \Pi$ such that $V_\beta(\pi, i) = +\infty\}$. Obviously, $S_{=\infty} \subset S_\infty$.

Lemma 2.5: *Under Conditions 2.1 and 2.2,* $\sum_{j\in S_0} p_{ij}(a)V_\beta(j)$ *is well defined for any* $(i, a) \in \Gamma$.

Proof: First, it should be noted that $\sum_j p_{ij}(a)V_\beta(\pi, j)$ is well defined for any $(i, a) \in \Gamma$ and $\pi \in \Pi$ from Condition 2.2 for policy (f, π) with $f(i) = a$. Here, (f, π) is a policy defined by using f in the first period and then π in the remaining periods. For any positive constant ε and state $j \in S_0$, let $\pi(\varepsilon, j)$ be a policy such that $V_\beta(\pi(\varepsilon, j), j) \geq V_\beta(j) - \varepsilon$ and $\pi(\varepsilon)$ be a policy choosing $\pi(\varepsilon, j)$ when the initial state is $j \in S_0$. Then

$$\sum_{j\in S_0} p_{ij}(a)V_\beta(\pi(\varepsilon, j), j) = \sum_{j\in S_0} p_{ij}(a)V_\beta(\pi(\varepsilon), j)$$

is well defined and for any subset $S'' \subset S_0$,

$$\sum_{j\in S''} p_{ij}(a)V_\beta(\pi(\varepsilon), j) \leq \sum_{j\in S''} p_{ij}(a)V_\beta(j) \leq \sum_{j\in S''} p_{ij}(a)[V_\beta(\pi(\varepsilon), j) + \varepsilon].$$

So, $\sum_{j\in S_0} p_{ij}(a)V_\beta(j)$ is well defined by the above formula and the definition of series. □

Now, we show the validity of the optimality equation if its right-hand side (see Eq. (2.5) below) is well defined.

Theorem 2.2: *Provided that Condition 2.1 and Condition 2.2 are true and that* $\sum_j p_{ij}(a)V_\beta(j)$ *is well defined for any* $(i, a) \in \Gamma$, *then* V_β *satisfies the following optimality equation:*

$$V_\beta(i) = \sup_{a\in A(i)} \{r(i, a) + \beta \sum_j p_{ij}(a)V_\beta(j)\}, \quad i \in S. \tag{2.5}$$

Proof: Following Condition 2.2, we have that

$$V_\beta(i) \leq \sup_{a\in A(i)} \{r(i, a) + \beta \sum_j p_{ij}(a)V_\beta(j)\}, \quad i \in S. \tag{2.6}$$

For any $\varepsilon > 0$, let $\pi(\varepsilon, i)$ be a policy such that $V_\beta(\pi(\varepsilon, i), i) \geq V_\beta(i) - \varepsilon$ for $i \in S_0$ and $V_\beta(\pi(\varepsilon, i), i) \geq 1/\varepsilon$ for $i \in S_\infty$, and for $i \in S_{-\infty}$, $V_\beta(\pi, i) = -\infty$ for any policy π. Let $\pi(\varepsilon)$ be a policy choosing $\pi(\varepsilon, j)$ when the initial state is $j \in S_0 \bigcup S_\infty$. Now we prove that for any $(i, a) \in \Gamma$,

$$V_\beta(i) \geq r(i, a) + \beta \sum_j p_{ij}(a)V_\beta(j). \tag{2.7}$$

Equation (2.7) is trivial for $i \in S_\infty$. For $i \in S - S_\infty$, let f with $f(i) = a$. Then, due to Condition 2.2,

$$V_\beta(i) \geq V_\beta((f, \pi(\varepsilon)), i) = r(i,a) + \beta \sum_j p_{ij}(a) V_\beta(\pi(\varepsilon), j). \tag{2.8}$$

Consider the following three cases.

1. $p_{iS_{-\infty}}(a) := \sum_{j \in S_{-\infty}} p_{ij}(a) > 0$ or $\sum_{j \in S_0} p_{ij}(a) V_\beta(j) = -\infty$. Because $\sum_{j \in S_0} p_{ij}(a) V_\beta(j)$ is well defined, it equals the negative infinite, which implies Eq. (2.7).

2. $p_{iS_{-\infty}}(a) = 0$ and $\sum_{j \in S_0} p_{ij}(a) V_\beta(j) > -\infty$ but $p_{iS_\infty}(a) > 0$. Then

$$\sum_j p_{ij}(a) V_\beta(j) = \sum_{j \in S_0} p_{ij}(a) V_\beta(j) + \sum_{j \in S_\infty} p_{ij}(a) V_\beta(j) = +\infty.$$

But from Eq. (2.8), we have

$$V_\beta(i) \geq r(i,a) + \beta \sum_{j \in S_0} p_{ij}(a)[V_\beta(j) - \varepsilon] + \beta p_{iS_\infty}(a)(1/\varepsilon)$$

which implies $V_\beta(i) = +\infty$ by letting $\varepsilon \to 0^+$. So Eq. (2.7) holds.

3. When the conditions in the above cases 1 and 2 do not hold, due to Eq. (2.8),

$$V_\beta(i) \geq r(i,a) + \beta \sum_{j \in S_0} p_{ij}(a) V_\beta(j) - \varepsilon.$$

This also implies Eq. (2.7) due to the arbitrariness of ε.

So Eq. (2.7) holds, which implies Eq. (2.5) from Eq. (2.6) and the arbitrariness of i and a. □

In Theorem 2.2, it is assumed that $\sum_j p_{ij}(a) V_\beta(j)$ is well defined for all (i,a), which will be true after eliminating some worst actions. This is shown in the following theorem.

Theorem 2.3: *Provided that Condition 2.1 and Condition 2.2 hold, then* $S^*_{=\infty} = S_{=\infty}$ *and* $A_1(i)$ *can be sized down to*

$$A_2(i) = \{a \in A_1(i) | p_{iS_{-\infty}}(a) = 0\}, \quad i \in S - S_{-\infty}.$$

After this sizing down, (1) $S^*_{-\infty} = S_{-\infty}$ *and so* $S' := S - S_{-\infty}$ *is closed, and (2) about* S'*-MDPs,* $A_2(i)$ *can further be sized down to*

$$A_3(i) = \{a \in A_2(i) | \sum_{j \in S' - S_\infty} p_{ij}(a) V_\beta(j) > -\infty\}, \quad i \in S'.$$

Then $S^*_\infty = S_\infty$ *and so* $S_0 = S - S_{-\infty} - S_\infty$ *is closed.*

Proof: From Condition 2.2, one can get $S^*_{=\infty} = S_{=\infty}$ as in Lemma 2.3. For $i \notin S^*_{-\infty}$, $A_2(i) = A_1(i)$. Now for $i \in S^*_{-\infty}$, suppose that there are $a \in A_1(i)$ and $j \in S_{-\infty}$ such that $p_{ij}(a) > 0$. Observing that $V_\beta(\pi, j) = -\infty$ for each π, we know that if $\pi_0(a|i) > 0$ then $V_\beta(\pi, i) = -\infty$ from Condition 2.2. So, $A_1(i)$ can be sized down to $A_2(i)$, after which it is obvious that $S^*_{-\infty} = S_{-\infty}$ (noting that if $A_1(i)$ is empty then $i \in S_{-\infty}$).

In order to prove (2), suppose that $i \in S'$ and $a \in A_2(i) - A_3(i)$. Then $\sum_{j \in S'-S_\infty} p_{ij}(a)V_\beta(\pi, j) = -\infty$ for any policy π, which implies that $V_\beta(\pi, i) = -\infty$ for any policy π satisfying $\pi_0(a|i) > 0$. So, $A_2(i)$ can be sized down to $A_3(i)$. After this, for $i \in S^*_\infty$, suppose that there are $a \in A_3(i)$ and $j_0 \in S_\infty$ such that $p_{ij_0}(a) > 0$. We say that there is a policy π^* such that

$$\sum_j p_{ij}(a)V_\beta(\pi^*, j) > -\infty. \tag{2.9}$$

Otherwise, for any π,

$$\sum_{j \in S_\infty} p_{ij}(a)V_\beta(\pi, j) + \sum_{j \in S'-S_\infty} p_{ij}(a)V_\beta(\pi, j) = -\infty.$$

But for $a \in A_3(i)$ there is $\{\pi^*(j), j \in S' - S_\infty\}$ with

$$\sum_{j \in S'-S_\infty} p_{ij}(a)V_\beta(\pi^*(j), j) > -\infty.$$

This implies together with the above equation and the definition of policies that

$$\sup_{\pi \in \Pi} \sum_{j \in S_\infty} p_{ij}(a)V_\beta(\pi, j) = -\infty.$$

On the other hand, for each constant $M > 0$ and state $j \in S_\infty$, there is a policy $\pi(M, j)$ with $V_\beta(\pi(M, j), j) \geq M$. Define a policy $\pi(M)$ by choosing $\pi(M, j)$ when the initial state is j. Then

$$\sup_{\pi \in \Pi} \sum_{j \in S_\infty} p_{ij}(a)V_\beta(\pi, j) \geq \sum_{j \in S_\infty} p_{ij}(a)V_\beta(\pi(M), j) \geq p_{ij_0}(a)M.$$

Letting $M \to \infty$ results in a contradiction. So there is a policy π^* satisfying Eq. (2.9).

Due to $j_0 \in S_\infty$, we have that for any $M > 0$, there is a policy $\pi^*(M)$ such that $V_\beta(\pi^*(M), j_0) \geq M$. For any f with $f(i) = a$, we define a policy $\pi(M) = (\pi_0, \pi_1, \ldots)$ by $\pi_0 = f$ and for $n > 0$,

$$\pi_n(a|h_n) = \begin{cases} \pi^*(M)_{n-1}(a|h_n), & \text{if } h_n = (i_0, a_0, j_0, a_1, i_2, \ldots, i_n) \\ \pi^*_{n-1}(a|h_n), & \text{otherwise.} \end{cases}$$

Then from Condition 2.2,

$$\begin{aligned} V_\beta(i) &\geq V_\beta(\pi(M), i) = r(i,a) + \beta \sum_j p_{ij}(a) V_\beta(\pi(M)^{i,a}, j) \\ &\geq r(i,a) + \beta \sum_{j \neq j_0} p_{ij}(a) V_\beta(\pi^*, j) + \beta p_{ij_0}(a) M. \end{aligned}$$

This implies together with the arbitrariness of M and Eq. (2.9) that $i \in S_\infty$. So $S^*_\infty = S_\infty$.

Obviously, $S_0 = S - S_{-\infty} - S_\infty$ is closed. □

The above theorem says that an action a at state i can be deleted from the action set $A(i)$ if $\sum_j p_{ij}(a)V_\beta(j)$ is not well defined. After deleting all such bad actions, $\sum_j p_{ij}(a)V_\beta(j)$ is certainly well defined and thus the optimality equation is true due to Theorem 2.2.

Corollary 2.1: *Suppose that Conditions 2.1 and 2.2 hold. Then the action set $A(i)$ can be sized down to $A_3(i)$ for $i \in S - S_{-\infty}$ and $\sum_j p_{ij}(a)V_\beta(j)$ is well defined for $i \in S_0$ and $a \in A_3(i)$.*

Proof: The former conclusion follows Theorem 2.3. About the latter one, it is obvious that $S^*_\infty = S_\infty, S^*_{-\infty} = S_{-\infty}, S_0 := S - S_\infty - S_{-\infty}$ is closed and $\sum_{j \in S_0} p_{ij}(a)V_\beta(j) > -\infty$ for $i \in S_0$ and $a \in A_3(i)$. So, $\sum_j p_{ij}(a)V_\beta(j)$ is well defined for each $i \in S_0$ and $a \in A_3(i)$. □

From Corollary 2.1, the original MDP model now is partitioned into four sub-MDP models with state spaces beginning, respectively, $S_{=\infty}, S_\infty - S_{=\infty}, S_{-\infty}, S - S_\infty - S_{-\infty}$. In $S_{=\infty}$, the optimal value function equals $+\infty$ and there is an optimal policy. In $S_\infty - S_{=\infty}$, the optimal value function equals $+\infty$ but there is no optimal policy. In $S_{-\infty}$, the optimal value function equals $-\infty$ and each policy is optimal. And in $S - S_\infty - S_{-\infty}$, the optimal value function is finite.

Without Condition 2.2, one can obtain similar results as Theorem 2.3 if the optimality equation (2.5) is true.

Corollary 2.2: *Suppose that Conditions 2.1 and 2.2 hold and that V_β satisfies Eq. (2.5). Then $S^*_\infty = S_\infty$ and all the results in Theorem 2.3 are true. Moreover, $\sum_{j \in S_0} p_{ij}(a)V_\beta(j)$ is well defined and finite for each $i \in S_0$ and $a \in A_3(i)$.* □

From the above results, we discuss only the S_0-MDP model in the rest of this chapter. That is, we assume that the reward function satisfies Eq. (2.3), and the optimal value function is a finite solution of the optimality equation (2.5).

2.2 Properties of the Optimality Equation

Although the optimal value function is a finite solution of the optimality equation (2.5), the solutions of Eq. (2.5) may be not unique in general. An example can

be seen in [115]. In this section, we discuss questions of which one among the solutions of the optimality equation is the optimal value function, and under which condition the solution of Eq. (2.5) is unique.

We define

$$V_{\beta,N}(\pi,i) = \sum_{n=0}^{N-1} \beta^n E_{\pi,i}\{r(X_n,\Delta_n)\}, \;\; \pi \in \Pi, \;\; i \in S \tag{2.10}$$

as the expected discounted total reward through period 0 to period $N-1$ from the initial state i under policy π.

First, we prove the following lemma.

Lemma 2.6: *For each policy π and state $i \in S$ such that $V_\beta(\pi,i)$ is finite,*

$$\liminf_{n\to\infty} \beta^n E_{\pi,i} V_\beta(X_n) \geq 0. \tag{2.11}$$

Moreover, if $\pi = (f_0, f_1, \ldots) \in \Pi_m^d$ is a deterministic Markov policy then the following term is finite,

$$\limsup_{n\to\infty} \beta^n E_{\pi,i} V_\beta(X_n) = \limsup_{n\to\infty} \beta^n [P(f_0)P(f_1)\cdots P(f_{n-1})V_\beta]_i,$$

where $[u]_i$ denotes the ith coordinate of a countable vector u and $P(f) = (p_{ij}(f(i)))$ is a matrix for $f \in F$.

Proof: First, due to Conditions 2.1 and 2.2, we have

$$\begin{aligned} V_\beta(\pi,i) &= E_{\pi,i} r(X_0,\Delta_0) + \beta \int_{A(i)} \pi_0(da|i) \sum_j p_{ij}(a) V_\beta(\pi^{i,a}, j) \\ &= E_{\pi,i} r(X_0,\Delta_0) + \sum_{t=1}^{\infty} \beta^t E_{\pi,i} r(X_t,\Delta_t). \end{aligned}$$

Substituting the similar expression of $V_\beta(\pi^{i,a}, j)$ into the above equations will result in

$$\begin{aligned} V_\beta(\pi,i) &= \sum_{t=0}^{1} E_{\pi,i} r(X_t,\Delta_t) + \beta^2 \int_{A(i)} \pi_0(da|i) \sum_j p_{ij}(a) \\ &\quad \cdot \int_{A(j)} \pi_1(da_1|i,a,j) \sum_{j_2} p_{jj_2}(a_1) V_\beta(\pi^{i,a,j,a_1}, j_2) \\ &= V_{\beta,2}(\pi,i) + \sum_{t=2}^{\infty} \beta^t E_{\pi,i} r(X_t,\Delta_t). \end{aligned}$$

Moreover, by using the induction method, we can prove that

$$V_\beta(\pi,i) = V_{\beta,n}(\pi,i) + \beta^n \int_{A(i)} \pi_0(da|i) \sum_{j_1} p_{ij_1}(a) \cdots$$

$$\cdot \int_{A(j_{n-1})} \pi_{n-1}(da_{n-1}|i,a,\ldots,j_{n-1})$$
$$\cdot \sum_{j_n} p_{j_{n-1}j_n}(a_{n-1}) V_\beta(\pi^{i,a,j_1,\ldots,a_{n-1}}, j_n)$$
$$= \quad V_{\beta,n}(\pi,i) + \sum_{t=n}^{\infty} \beta^t E_{\pi,i} r(X_t, \Delta_t), \quad n \geq 1.$$

As the tail term of the series $\sum_{t=0}^{\infty} \beta^t E_{\pi,i} r(X_t, \Delta_t)$, the second term in the above equation, $\sum_{t=n}^{\infty} \beta^t E_{\pi,i} r(X_t, \Delta_t)$, tends to zero when n tends to infinity because $V_\beta(\pi, i)$ is finite. Then

$$\begin{aligned}
& \beta^n E_{\pi,i} V_\beta(X_n) \\
= \quad & \beta^n \int_{A(i)} \pi_0(da|i) \sum_{j_1} p_{ij_1}(a) \cdots \\
& \cdot \int_{A(j_{n-1})} \pi_{n-1}(da_{n-1}|i,a,\ldots,j_{n-1}) \sum_{j_n} p_{j_{n-1}j_n}(a_{n-1}) V_\beta(j_n) \\
\geq \quad & \sum_{t=n}^{\infty} \beta^t E_{\pi,i} r(X_t, \Delta_t), \quad n \geq 1,
\end{aligned}$$

where the inequality is due to $V_\beta(j_n) \geq V_\beta(\pi^{i,a,j_1,\ldots,a_{n-1}}, j_n)$. By letting $\liminf_n$ above, we obtain Eq. (2.11).

Second, for each $f \in F$, we write a vector $r(f)$ with its ith coordinate $r(i, f(i))$, a vector V_β with $V_\beta(i)$ and a vector $V_\beta(\pi)$ with $V_\beta(\pi, i)$. With the optimality equation (2.5), we have that

$$V_\beta \geq r(f) + \beta P(f) V_\beta, \quad \forall f \in F.$$

From this we can prove by the induction method that for any deterministic Markov policy $\pi = (f_0, f_1, \ldots) \in \Pi_m^d$,

$$V_\beta \geq V_{\beta,n}(\pi) + \beta^n P(f_0) P(f_1) \cdots P(f_n) V_\beta$$

which together with the finiteness of V_β result in that $\limsup_n \beta^n E_{\pi,i} V_\beta(X_n)$ is finite. □

The conclusion expressed in Eq. (2.11) is important for us to get result 1 in Theorem 2.4 below.

The following lemma is well known [132] and can be proved easily.

Lemma 2.7: *For any policy $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$ and state $i, j \in S$, we define a Markov policy $\sigma = (\sigma_0, \sigma_1, \ldots) \in \Pi_m$ by*

$$\sigma_n(a \mid j) = P_\pi\{\Delta_n = a \mid X_0 = i, X_n = j\}, \quad (j,a) \in \Gamma, \quad n \geq 0.$$

Then for any $n \geq 0$ and $(j, a) \in \Gamma$ we have

$$P_\sigma\{X_n = j, \Delta_n = a \mid X_0 = i\} = P_\pi\{X_n = j, \Delta_n = a \mid X_0 = i\}.$$

□

The above lemma says that we can consider only Markov policies, and

$$V_\beta(i) = \sup_{\pi \in \Pi_m} V_\beta(\pi, i), \quad i \in S. \tag{2.12}$$

For any vector V, state i, and action $a \in A(i)$, we define $T_a V(i)$ by

$$T_a V(i) = r(i, a) + \beta \sum_j p_{ij}(a) V(j), \quad i \in S.$$

Certainly, $T_a V(i)$ may be not well defined for some V. But $T_a V_\beta(i)$ is well defined from the optimality equation (2.5). Moreover, let $TV(i) = \sup_{a \in A(i)} T_a V(i)$ and $T_f V(i) = T_{f(i)} V(i)$ for any $f \in F$. Then the optimality equation can be rewritten by $V_\beta = TV_\beta$. It should be noted that introducing $T_f V, T_a V, TV$ here is only for notational simplicity although they played important roles in the literature.

Lemma 2.8: *If V is a finite solution of the optimality equation (2.5), then*

1. *$V \leq V_\beta$ if for each policy $\pi \in \Pi_m^d$ with that $V_\beta(\pi, i)$ is finite,*

$$\liminf_{n \to \infty} \beta^n E_{\pi,i}\{V(X_n)\} \leq 0, \quad \forall i \in S. \tag{2.13}$$

2. *$V \geq V_\beta$ if and only if for each policy $\pi \in \Pi_m^d$ with that $V_\beta(\pi, i)$ is finite,*

$$\limsup_{n \to \infty} \beta^n E_{\pi,i}\{V(X_n)\} \geq 0, \quad \forall i \in S. \tag{2.14}$$

3. *$V = V_\beta$ if for each policy $\pi \in \Pi_m^d$ with that $V_\beta(\pi, i)$ is finite,*

$$\liminf_{n \to \infty} \beta^n E_{\pi,i}\{V(X_n)\} = 0, \quad \forall i \in S. \tag{2.15}$$

So, if the optimality equation has a solution V satisfying Eq. (2.15), then the optimal value function V_β also satisfies Eq. (2.15).

Proof: 1. Suppose that $\{\varepsilon_n \geq 0\}$ is a sequence of constants with $\sum_{n=0}^\infty \beta^n \varepsilon_n < \infty$ and $\{f_n \in F\}$ satisfy that $T_{f_n} V \geq TV - \varepsilon_n$ for $n \geq 0$. Let a policy be $\pi = (f_0, f_1, \ldots)$. Then, we can prove by the induction method that for $N \geq 0$,

$$\begin{aligned} & \sum_{n=0}^{N} \beta^n E_{\pi,i} r(X_n, \Delta_n) + \beta^{N+1} E_{\pi,i}\{V(X_{N+1})\} \\ = \quad & T_{f_0} T_{f_1} \cdots T_{f_N} V(i) \geq V(i) - \sum_{n=0}^{N} \beta^n \varepsilon_n, \quad i \in S. \end{aligned}$$

By taking $\liminf_{N\to\infty}$ we get that $V_\beta \geq V_\beta(\pi) \geq V - \sum_{n=0}^{\infty} \beta^n \varepsilon_n$. Due to the arbitrariness of ε_n, $V_\beta \geq V$.

2. Suppose that Eq. (2.14) is true. Due to $V = TV$,

$$V(i) \geq r(i,a) + \beta \sum_j p_{ij}(a) V(j), \ \ \forall (i,a) \in \Gamma.$$

Then for each policy π we have that

$$V(i) \geq \sum_{n=0}^{N} \beta^n E_{\pi,i} r(X_n, \Delta_n) + \beta^{N+1} E_{\pi,i} V(X_{N+1}), \ \ i \in S.$$

Again by taking $\limsup_{N\to\infty}$, we know that $V \geq V_\beta(\pi)$ for each policy π. So $V \geq V_\beta$.

On the other hand, Eq. (2.14) can be obtained by $V \geq V_\beta$ and Lemma 2.6.

3. It follows from 1 and 2 above. □

From Lemma 2.6 we know that the fact of V_β satisfying Eq. (2.13) is equivalent to that of V_β satisfying Eq. (2.15). By this and Lemma 2.8, we have the following theorem characterizing the optimal value as a solution of the optimality equation.

Theorem 2.4:

1. *V_β is the smallest solution of the optimality equation satisfying Eq. (2.14).*

2. *V_β is the unique solution of the optimality equation satisfying Eq. (2.15) (or Eq. (2.13) equivalently) if and only if the optimality equation has a solution V that satisfies Eq. (2.15).* □

In the literature (see, e.g., [7] and [132]), it was only claimed that V_β is the largest nonpositive solution of the optimality equation for the negative MDP models (certainly satisfying Eq. (2.13)), or is the smallest nonnegative solution of the optimality equation for the positive MDP models (certainly satisfying Eq. (2.14)). Here V_β is the smallest solution satisfying Eq. (2.14) without conditions, and is the unique one if and only if there is a solution of the optimality equation satisfying Eq. (2.15) (for instance, when $V_\beta \leq 0$ or $r \leq 0$).

3. Properties of Optimal Policies

In this section, we discuss some properties of optimal policies. First we show the following theorem, which is about the existence of ε-optimal policies.

Theorem 2.5:

1. *Suppose that $\{\varepsilon_n \geq 0\}$ satisfies $\sum_{n=0}^{\infty} \beta^n \varepsilon_n < \infty$ and for each n, $f_n \in F$ attains the ε_n-supremum in the optimality Eq. (2.5); that is,*

$T_{f_n} V_\beta \geq TV_\beta - \varepsilon_n$. Then, the policy $\pi = (f_0, f_1, \ldots)$ is $\sum_{n=0}^{\infty} \beta^n \varepsilon_n$-optimal if π satisfies Eq. (2.15) with $V = V_\beta$.

2. *For any policy π, π is optimal if and only if $V_\beta(\pi)$ is a solution of Eq. (2.5) and satisfies Eq. (2.14) with $V = V_\beta(\pi)$.*

Proof: 1. For the policy $\pi = (f_0, f_1, \ldots)$, it can be shown similarly to that in 1 of Lemma 2.8 that $V_\beta(\pi) \geq V_\beta - \sum_{n=0}^{\infty} \beta^n \varepsilon_n$.

2. If policy π is optimal (i.e., $V_\beta(\pi) = V_\beta$), then $V_\beta(\pi)$ is obviously a solution of Eq. (2.5) and satisfies Eq. (2.14) due to Lemma 2.6. On the other hand, if $V_\beta(\pi)$ is a solution of Eq. (2.5) and satisfies Eq. (2.14), then $V_\beta(\pi) \geq V_\beta$ due to Lemma 2.8. So, π is optimal. □

The first conclusion of Theorem 2.5 addresses the optimality of policies attaining the ε-supremum of the optimality Eq. (2.5), whereas the second conclusion characterizes the optimality of a policy π by its criterion value $V_\beta(\pi)$. A policy π satisfies Eq. (2.14) with $V = V_\beta(\pi)$ if $V_\beta(\pi) \geq 0$, or $r \geq 0$ especially.

The following corollary on ε-optimal stationary policies follows immediately 1 of Theorem 2.5.

Corollary 2.3: *If a stationary policy f attains the $\varepsilon(\geq 0)$-supremum in the optimality equation and $\liminf_{n\to\infty} \beta^n E_{f,i} V_\beta(X_n) \leq 0$, then f is $(1-\beta)^{-1}\varepsilon$-optimal when $\beta \in (0,1)$. Moreover, when $\varepsilon = 0$, f is optimal irrespectively of the value of β.* □

The following theorem shows the dominance of stationary policies and deterministic Markov policies.

Theorem 2.6:

1. *$V_\beta = \sup_{\pi \in \Pi_m^d} V_\beta(\pi)$ if V_β satisfies Eq. (2.15) for each $\pi \in \Pi_m^d$ ($V_\beta \leq 0$ or $r \leq 0$ especially).*

2. *$V_\beta = \sup_f V_\beta(f)$ if either (a) V_β satisfies Eq. (2.15) for each $\pi \in \Pi_m^d$, when the discount factor $\beta \in (0,1)$ or there is $f \in F$ attaining the supremum of the optimality Eq. (2.5), or (b) there is a policy π such that $V_\beta(\pi) \geq 0$ ($r \geq 0$ especially).*

Proof: 1 and 2 (a) follow from 1 of Theorem 2.5 and Corollary 2.3. About 2 (b), because $V_\beta \geq V_\beta(\pi) \geq 0$, for any $i \in S$ with $V_\beta(i) = 0$, $V_\beta(\pi, i) = V_\beta(i) = 0$. Then, due to Theorem 1.1 in [143], the result follows. □

In the following, we discuss the structure of optimal policies under the restriction that all the action sets are countable. This restriction is not essential and is just to avoid measure theory. First we give the following two concepts.

Definition 2.4: *A history $h_n = (i_0, a_0, \ldots, i_{n-1}, a_{n-1}, i) \in H_n$ is called a realized history under policy π if the probability of h_n under policy π is*

positive; that is

$$\pi_0(a_0|i_0)p_{i_0,i_1}(a_0)\cdots\pi_{n-1}(a_{n-1}|i_0,a_0,i_1,\ldots,i_{n-1})p_{i_{n-1},i}(a_{n-1})>0.$$

And $k_{n-1}=(i_0,a_0,\ldots,i_{n-1},a_{n-1})$ *in* h_n *is also called a realized history if its probability under policy* π *is positive. For* $n=0$*, we always say* $h_0=(i)$ *to be realized.*

Definition 2.5: *For any policy* $\pi=(\pi_0,\pi_1,\ldots)$ *and a history* k_n*, we define a new policy* $\pi^{k_n}=(\pi'_0,\pi'_1,\ldots)$ *by*

$$\pi'_m(\cdot\mid h_m)=\pi_{m+n+1}(\cdot\mid k_n,h_m),\quad m\geq 0, h_m\in H_m.$$

Obviously, π^{k_n} generalizes $\pi^{i,a}$ introduced in Condition 2.2, and is exactly the remaining part of policy π when the history k_n has happened. For convenience we denote for any policy $\pi=(\pi_0,\pi_1,\ldots)$, state i, a function U on Γ, and a function V on S that

$$\begin{aligned}
\pi_0U(i) &= \sum_{a_0\in A(i)}\pi_0(a_0|i)U(i,a_0)\\
\pi_0p\pi_1U(i) &= \sum_{a_0\in A(i)}\pi_0(a_0|i)\sum_{i_1}p_{ii_1}(a_0)\\
&\quad\cdot\sum_{a_1\in A(i_1)}\pi_1(a_1|i,a_0,i_1)U(i_1,a_1)\\
\pi_0p\pi_1p\cdots\pi_nU(i) &= \sum_{a_0\in A(i)}\pi_0(a_0|i)\sum_{i_1}p_{ii_1}(a_0)\cdots\\
&\quad\cdot\sum_{a_n\in A(i_n)}\pi_n(a_n|i,a_0,\ldots,i_n)U(i_n,a_n),\ \ n\geq 0
\end{aligned}$$

and

$$\begin{aligned}
\pi_0pV(i) &= \sum_{a_0\in A(i)}\pi_0(a_0|i)\sum_{i_1}p_{ii_1}(a_0)V(i_1)\\
\pi_0p\pi_1pV(i) &= \sum_{a_0\in A(i)}\pi_0(a_0|i)\sum_{i_1}p_{ii_1}(a_0)\\
&\quad\sum_{a_1\in A(i_1)}\pi_1(a_1|i,a_0,i_1)\sum_{i_2}p_{i_1i_2}(a_1)V(i_2)\\
\pi_0p\pi_1p\cdots\pi_npV(i) &= \sum_{a_0\in A(i)}\pi_0(a_0|i)\sum_{i_1}p_{ii_1}(a_0)\cdots\\
&\quad\cdot\sum_{a_n\in A(i_n)}\pi_n(a_n|i,a_0,i_1,\ldots,i_n)
\end{aligned}$$

$$\cdot \sum_{i_{n+1}} p_{i_n i_{n+1}}(a_n) V(i_{n+1}), \quad n \geq 0.$$

The following lemma says that any remaining part of an optimal policy is still optimal for the remaining subprocess.

Lemma 2.9: *$\pi = (\pi_0, \pi_1, \ldots)$ is an optimal policy if and only if $V_\beta(i) = V_\beta(\pi^{k_{n-1}}, i)$ for any realized history $h_n = (k_{n-1}, i) = (i_0, a_0, \ldots, i_{n-1}, a_{n-1}, i)$ under policy π.*

Proof: The sufficiency is obvious by only taking $n = 0$ and $h_0 = (i)$. We now prove the necessity. Suppose that π is optimal; that is, $V_\beta(i) = V_\beta(\pi, i) = V_\beta(\pi^{k_{-1}}, i)$. So the result for $n = 0$ is true.

For $n \geq 1$, we have

$$\begin{aligned}
V_\beta &= V_\beta(\pi) = \sum_{t=0}^{\infty} \beta^t \pi_0 p \pi_1 p \cdots \pi_{t-1} p \pi_t r \\
&= \sum_{t=0}^{n-1} \beta^t \pi_0 p \cdots \pi_t r + \beta^n \sum_{t=n}^{\infty} \beta^{t-n} \pi_0 p \cdots \pi_t r \\
&= \sum_{t=0}^{n-1} \beta^t \pi_0 p \cdots \pi_t r + \beta^n \pi_0 p \cdots \pi_{n-1} p \sum_{t=n}^{\infty} \beta^{t-n} \pi_n p \cdots \pi_t r \\
&= \sum_{t=0}^{n-1} \beta^t \pi_0 p \cdots \pi_t r + \beta^n \pi_0 p \cdots \pi_{n-1} p V_\beta(\pi^{k_{n-1}}) \\
&\leq \sum_{t=0}^{n-1} \beta^t \pi_0 p \cdots \pi_t r + \beta^n \pi_0 p \cdots \pi_{n-1} p V_\beta(\pi) \\
&= V_\beta(\pi') \leq V_\beta, \qquad (2.16)
\end{aligned}$$

where the first inequality follows the fact that π is optimal, and the policy $\pi' = (\pi_0, \pi_1, \ldots, \pi_{n-1}, \pi_0, \pi_1, \pi_2, \ldots)$. Hence, the above inequalities should be equalities and so

$$\pi_0 p \cdots \pi_{n-1} p V_\beta(\pi^{k_{n-1}}) = \pi_0 p \cdots \pi_{n-1} p V_\beta(\pi).$$

Because $V_\beta(\pi) = V_\beta$, we have that

$$\pi_0 p \cdots \pi_{n-1} p [V_\beta - V_\beta(\pi^{k_{n-1}})] = 0.$$

But $V_\beta \geq V_\beta(\pi^{k_{n-1}})$, so we know that $V_\beta(i) = V_\beta(\pi^{k_{n-1}}, i)$ if (k_{n-1}, i) is realized under π. □

We define for $(i, a) \in \Gamma$,

$$G(i, a) := V_\beta(i) - \{r(i, a) + \beta \sum_j p_{ij}(a) V_\beta(j)\} \geq 0.$$

It is nonnegative and represents the deviation from the optimal value in state i if action $a \in A(i)$ is chosen. The optimal action set $A^*(i)$ at state i is now defined by

$$A^*(i) = \{a \mid a \in A(i), G(i,a) = 0\}, \;\; i \in S.$$

The optimality of an action $a \in A^*(i)$ at state i implies that no deviation from the optimal value occurs when choosing a, or equivalently, optimal actions at state i are actions that attain the supremum in the optimality equation. In the literature (e.g., [38]), $a \in A^*(i)$ is said to be a conserving action at state i, and $f \in F$ a conserving stationary policy if $f(i)$ is conserving at i for each $i \in S$.

Lemma 2.10: *If there is an optimal policy $\pi^* = (\pi_0^*, \pi_1^*, \ldots)$, then all the optimal action sets $A^*(i)$ are nonempty and $\pi_0^*(A^*(i) \mid i) = 1$ for each $i \in S$.*

Proof: In the proof of Lemma 2.9, it has been shown already that all the inequalities in Eq. (2.16) are equalities. Taking $n = 1$ results in $V_\beta = \pi_0^* r + \beta \pi_0^* p V_\beta$. That is, for $i \in S$,

$$\sum_{a \in A(i)} \pi_0^*(a \mid i) G(i,a) = \pi_0^* \{V_\beta - (r + \beta p V_\beta)\}(i) = 0,$$

which implies that $\pi_0^*(A^*(i) \mid i) = 1$ for each $i \in S$ because $G(i,a) \geq 0$ and $\sum_{a \in A(i)} \pi_0^*(a \mid i) = 1$. Certainly, all the optimal action sets $A^*(i)$ are nonempty. □

Based on the above Lemmas 2.9 and 2.10, we can now characterize the structure of optimal policies.

Theorem 2.7: *$\pi = (\pi_0, \pi_1, \ldots)$ is optimal if and only if the following three statements are true.*

1. *The optimal action set $A^*(i)$ is nonempty for each $i \in S$.*
2. *$\pi_n(A^*(i) \mid h_n) = 1$ for any $n \geq 0$ and realized history $h_n = (i_0, a_0, \ldots, a_{n-1}, i)$ under the policy π.*
3. *(π, V_β) satisfies Eq. (2.15).*

Proof: Necessity. Suppose that $h_n = (k_{n-1}, i) = (i_0, a_0, \ldots, a_{n-1}, i)$ is a realized history under policy π. Then from Lemma 2.9, $V_\beta(\pi^{k_{n-1}}, i) = V_\beta(i)$. Define a new policy π^* by using policy $\pi^{k_{n-1}}$ if the initial state is i, and using policy π otherwise. Then

$$V_\beta(\pi^*, j) = \begin{cases} V_\beta(\pi, j) = V_\beta(j), & \text{if } j \neq i \\ V_\beta(\pi^{k_{n-1}}, i) = V_\beta(i), & \text{if } j = i. \end{cases}$$

So, π^* is also optimal, which together with Lemma 2.10 implies the following,

$$\pi_n(A^*(i) \mid h_n) = \pi_0^{k_{n-1}}(A^*(i) \mid i) = \pi_0^*(A^*(i) \mid i) = 1.$$

Thus 1 and 2 are true, and 3 follows 2 of Theorem 2.5 and $V_\beta = V_\beta(\pi)$.

Sufficiency. From the proof of Lemma 2.9 we have

$$V_\beta = \sum_{t=0}^{n-1} \beta^t \pi_0 p \cdots \pi_t r + \beta^n \pi_0 p \cdots \pi_{n-1} p V_\beta,$$

in which by letting $n \to \infty$ we know that $V_\beta = V_\beta(\pi)$. So π is optimal. □

Statement 2 of Theorem 2.7 is equivalent to saying that π chooses only the optimal actions in a realized history, or equivalently, all actions in any realized history belong to the optimal action sets: $a_k \in A^*(i_k)$. From Theorem 2.7, we know that the set of all optimal stationary policies is exactly $\times_i A^*(i)$. We can say that statements 1 and 3 of Theorem 2.7 characterize the structure of the set of all optimal policies.

From Theorem 2.7, all properties of optimal policies (e.g., the convex combination of optimal policies) discussed in the literature (see [28]) can be obtained easily. The details are omitted here.

4. Successive Approximation

In this section, we consider the successive approximation:

$$\begin{aligned} V_0(i) &= 0,\ i \in S \\ V_{n+1}(i) &= TV_n(i) \\ &= \sup_{a \in A(i)} \{r(i,a) + \beta \sum_j p_{ij}(a) V_n(j)\},\ \ i \in S,\ \ n \geq 0. \end{aligned} \tag{2.17}$$

Certainly, $V_n(i)$ is the maximal discounted expected total reward when the state is i and n periods are remaining. Due to the nonuniqueness of solutions of the optimality equation, V_n may not tend to the optimal value function V_β. But we have the following theorem for it.

Theorem 2.8: *We have the following three statements.*

1. $\liminf_{n\to\infty} V_n(i) \geq V_\beta(i),\ i \in S.$

2. $\lim_{n\to\infty} V_n = V_\beta$, *if* $V_\beta \geq 0$ *or*

$$\limsup_{n\to\infty} \inf_\pi \sum_{t=n+1}^{\infty} \beta^t E_{\pi,i} r(X_t, \Delta_t) \geq 0,\ \ \forall i \in S. \tag{2.18}$$

3. *If* $r \leq 0$, *then* $V_\infty = \lim_{n\to\infty} V_n$ *exist and* $V_\infty \geq V_\beta$. *Moreover,* $V_\infty = V_\beta$ *if and only if* V_∞ *is a solution of the optimality equation.*

Proof: 1. Because $V_n(i)$ is the maximal expected discounted total reward when the state is i and n periods are remaining, we know that for each policy π and state i,

$$V_{n+1}(i) \geq V_n(\pi, i) := \sum_{t=0}^{n} \beta^t E_{\pi,i} r(X_t, \Delta_t), \quad n \geq 0.$$

By taking $\liminf_{n\to\infty}$ in the above inequality, we have that

$$\liminf_{n\to\infty} V_n(i) \geq V_\beta(\pi, i), \quad i \in S, \ \pi \in \Pi.$$

Due to the arbitrariness of π, $\liminf_{n\to\infty} V_n(i) \geq V_\beta(i)$ for $i \in S$.

2. Due to 1, it suffices to show that

$$\limsup_{n\to\infty} V_n(i) \leq V_\beta(i), \quad i \in S.$$

Now if $V_\beta \geq 0$, then $V_\beta = TV_\beta \geq T0 = V_1$. By the induction method we can prove that $V_n \leq V_\beta$ for each $n \geq 0$. Thus $\limsup_{n\to\infty} V_n(i) \leq V_\beta(i)$ for $i \in S$.

If Eq. (2.18) is true, then for each policy π and state i,

$$\begin{aligned} V_\beta(i) &\geq V_\beta(\pi, i) \\ &= V_n(\pi, i) + \sum_{t=n+1}^{\infty} \beta^t E_{\pi,i} r(X_t, \Delta_t) \\ &\geq V_n(\pi, i) + \inf_{\pi} \sum_{t=n+1}^{\infty} \beta^t E_{\pi,i} r(X_t, \Delta_t). \end{aligned}$$

Then,

$$V_\beta(i) \geq V_n(i) + \inf_{\pi} \sum_{t=n+1}^{\infty} \beta^t E_{\pi,i} r(X_t, \Delta_t), \quad i \in S.$$

Letting $\limsup_{n\to\infty}$ in the above formula results in

$$V_\beta(i) \geq \limsup_{n\to\infty} V_n(i), \quad i \in S.$$

3. When the reward function $r \leq 0$ is nonpositive, it is clear that $0 \geq T0 = V_1$. Then with the induction method we can prove that V_n is decreasing in n. So V_∞ exists and it is larger than or equal to V_β due to 1.

Now, if $V_\infty = V_\beta$, surely V_∞ is a solution of the optimality equation. However, if V_∞ is a solution of the optimality equation, then $V_\infty \leq V_\beta$ from Lemma 2.8. So $V_\infty = V_\beta$. □

The main problems in successive approximation include (a) whether it is convergent (i.e., $\lim_n V_n(i) = V_\beta(i)$ for all $i \in S$), and if convergent then (b) how fast it converges. The theorem above answers the first problem in part. As for the second problem, some bounded condition may be needed on the reward, for example, those given in the next section.

5. Sufficient Conditions

This section gives some sufficient conditions for Conditions 2.1 and 2.2. In fact, almost all conditions presented in the literature are sufficient for Conditions 2.1 and 2.2.

First, we give another definition of the criterion function. Let $r^{\pm}(i,a) = \max\{0, \pm r(i,a)\}$. We define

$$\begin{aligned} V_\beta^{\pm}(\pi, i) &= \sum_{n=0}^{\infty} \beta^n E_{\pi,i}\{r^{\pm}(X_n, \Delta_n)\}, \ \ \pi \in \Pi, \ \ i \in S, \\ V_\beta(\pi, i) &= V_\beta^{+}(\pi, i) - V_\beta^{-}(\pi, i), \ \ \pi \in \Pi, \ \ i \in S. \end{aligned} \tag{2.19}$$

It is apparent that for each $\pi \in \Pi$ and $i \in S$, $V_\beta(\pi, i)$ is well defined if and only if $V_\beta^+(\pi, i)$ or $V_\beta^-(\pi, i)$ is finite. In this case, $E_{\pi,i}\{r^+(X_n, \Delta_n)\}$ or $E_{\pi,i}\{r^-(X_n, \Delta_n)\}$ is also finite for each i and n. Hence,

$$E_{\pi,i}\{r(X_n, \Delta_n)\} = E_{\pi,i}\{r^+(X_n, \Delta_n)\} - E_{\pi,i}\{r^-(X_n, \Delta_n)\}$$

is well defined, and so

$$\begin{aligned} V_\beta(\pi, i) &= \sum_{n=0}^{\infty} \beta^n \{E_{\pi,i}\{r^+(X_n, \Delta_n)\} - E_{\pi,i}\{r^-(X_n, \Delta_n)\} \\ &= \sum_{n=0}^{\infty} \beta^n E_{\pi,i}\{r(X_n, \Delta_n)\}. \end{aligned}$$

Theorem 2.9: *Conditions 2.1 and 2.2 hold if $V_\beta^+(\pi, i) < \infty$ or $V_\beta^-(\pi, i) < \infty$ for each π and i.*

Proof: Under the given condition, $V_\beta(\pi, i)$ is well defined and so Condition 2.1 is true. Now, because both $r^+(i,a)$ and $r^-(i,a)$ are nonnegative, one can obtain that for each policy π,

$$V_\beta^{\pm}(\pi, i) = \int_{A(i)} \pi_0(da|i)\{r^{\pm}(i,a) + \beta \sum_j p_{ij}(a) V_\beta^{\pm}(\pi^{i,a}, j)\}, \ \ i \in S.$$

This implies that

$$V_\beta(\pi, i) = V_\beta^{+}(\pi, i) - V_\beta^{-}(\pi, i)$$

$$
\begin{aligned}
&= \int_{A(i)} \pi_0(da|i)\{[r^+(i,a) - r^-(i,a)] \\
&\quad +\beta \sum_j p_{ij}(a)[V_\beta^+(\pi^{i,a}, j) - V_\beta^-(\pi^{i,a}, j)]\} \\
&= \int_{A(i)} \pi_0(da|i)\{r(i,a) + \beta \sum_j p_{ij}(a) V_\beta(\pi^{i,a}, j)\}.
\end{aligned}
$$

Hence, Condition 2.2 is also true. □

The condition given in Theorem 2.9 is true for the positive reward, the negative reward, and the bounded reward with $\beta \in (0,1)$. Moreover we present the following two sufficient conditions by using the ideas presented in Hu and Hu [80].

Condition (UR$^+$): *The following three statements are true.*

1. $\beta \in (0,1)$, $\sum_j p_{ij}(a) r^+(j,f) < +\infty$ *for each* $(i,a) \in \Gamma$ *and* $f \in F$.

2. There is a nonnegative function μ_+ *on* S *and a constant* $c_+ \in (0,1)$ *such that*

$$
\begin{aligned}
&|\sum_j p_{ij}(a) r^+(j, f(j)) - r^+(i,a)| \le \mu_+(i), \quad (i,a) \in \Gamma, \quad f \in F, \\
&\beta \sum_j p_{ij}(a) \mu_+(j) \le c_+ \mu_+(i), \quad (i,a) \in \Gamma, \quad f \in F.
\end{aligned}
$$

3. $U_+(i) := \sup\{r^+(i,a) | a \in A(i)\}$ *is finite for each* $i \in S$.

Condition (UR$^-$): *This condition is exactly Condition (UR$^+$) except that* $r^+(i,a), \mu_+(i), c_+$ *and* $U_+(i)$ *are replaced, respectively, by* $r^-(i,a), \mu_-(i)$, c_-, *and* $U_-(i)$.

Due to Eq. (2.3), $U_+(i)$ is finite for all $i \in S$. But we cannot conclude that $U_-(i)$ is finite.

Conditions (UR$^+$) and (UR$^-$) generalize that the reward function is bounded above and below, respectively. When $r(i,a) \le 0, r^+(i,a) = 0$ and we can take that $\mu_+(i) = U_+(i) = 0$ and $c_+ = \beta$. Then Condition (UR$^+$) is true. When $r(i,a) \ge 0$, Condition (UR$^-$) is also true.

From the results in Hu and Hu [80], we know that if Condition (UR$^+$) or (UR$^-$) is true then

$$
\pm V_\beta(\pi, i) \le (1-\beta)^{-1} U_\pm(i) + \beta(1-\beta)^{-1}(1-c_\pm)^{-1} \mu_\pm(i), \qquad (2.20)
$$

where $\rho_\pm = \beta(1-\beta)^{-1}(1-c_\pm)^{-1}$, and so both Conditions 2.1 and 2.2 are true. This shows the following theorem.

Theorem 2.10: *Conditions 2.1 and 2.2 hold if either Condition (UR$^+$) is true or Condition (UR$^-$) is true.* □

Moreover, Condition 2.2 will be true if the criterion is defined by

$$V_\beta(\pi, i) = E_{\pi,i}\left\{\sum_{n=0}^{\infty} \beta^n r(X_n, \Delta_n)\right\}, \quad \pi \in \Pi, \quad i \in S, \tag{2.21}$$

where the random series is defined by the mean square limitation of the partial sum $\sum_{n=0}^{N} \beta^n r(X_n, \Delta_n)$ as N tends to infinity. The definition (2.21) implies the definition of (2.1) by the properties of the mean square convergence. However, the reverse is not true in general, and a counterexample can be found in Guo [46].

Theorem 2.11: *If $V_\beta(\pi, i)$ is defined by (2.21), then Condition 2.1 implies Condition 2.2.*

Proof: This theorem is proved in Guo [46], but we give a simpler proof here. From (2.21) and the definition of mean square convergence, one can get that

$$\begin{aligned}
V_\beta(\pi, i) &= E_{\pi,i}\{r(X_0, \Delta_0) + \beta \sum_{n=1}^{\infty} \beta^{n-1} r(X_n, \Delta_n)\} \\
&= \int_{A(i)} \pi_0(da|i) r(i, a) + \sum_{a \in A(i)} \pi_0(a|i) \sum_j p_{ij}(a) \\
&\quad \cdot \beta E_\pi\{\sum_{n=1}^{\infty} \beta^{n-1} r(X_n, \Delta_n) | X_0 = i, \Delta_0 = a, X_1 = j\} \\
&= \int_{A(i)} \pi_0(da|i)\{r(i, a) + \beta \sum_j p_{ij}(a) E_{\pi^{i,a},j} \sum_{n=0}^{\infty} \beta^n r(X_n, \Delta_n)\} \\
&= \int_{A(i)} \pi_0(da|i)\{r(i, a) + \beta \sum_j p_{ij}(a) V_\beta(\pi^{i,a}, j)\}.
\end{aligned}$$

This completes the proof. □

From Theorems 2.9 to 2.11 we know that Conditions 2.1 and 2.2 are weaker.

6. Notes and References

There are three classical cases for DTMDP models with the total reward criterion: (1) the discounted MDP models where the reward function is uniformly bounded and the discount factor $\beta \in (0, 1)$, (2) the positive MDP models where the reward function is nonnegative and the discount factor $\beta = 1$, and (3) the negative MDP models where the reward function is nonpositive and the discount factor $\beta = 1$. The pioneering works on these three cases include, respectively, Blackwell [6], [7], Strauch [132], and others. The studies on these three cases are, respectively, done. But the differences among them are significant as pointed out in Feinberg [38]. Later, a so-called General Convergence

Condition was presented to unify these three cases. This condition means that the expected total reward under the positive part of the reward function is finite for each policy and initial state. One can see it, for example, in [27] and [38].

The standard results for MDP models, as pointed out in Chapter 1, include (1) the model's well definition, (2) the validity of the optimality equation, (3) ε-optimal policies from the optimality equation, and (4) algorithms for computing the optimal value and approximating optimal policies.

In order to obtain these standard results, some conditions are required. First, in order to ensure that the Markov decision process model is well defined, there are many papers that present various conditions. For example, for the discounted DTMDP models, Harrison [49], Lippman [92], Wessels [145], and Hu and Hu [80] presented various conditions under which the criterion $V_\beta(\pi, i)$ is well defined and finite and, in fact, is bounded in π for each i. For positive MDP models and negative MDP models, it is often assumed that the optimal value function is less than positive infinity (see [27]) or larger than negative infinity (see [115]). Certainly, in each of these cases, the model is well defined under the presented conditions.

In the literature, the general and most usual method to study a MDP model, after having presented conditions, is to show the standard results 1, 2, 3, and 4 successively. In the proof, some other conditions may be needed. This can be seen, for example, in [27], [58], [115], and [38]. Feinberg [38] surveyed the criterion of total reward.

The structure of optimal policies is a generalization of the standard result 3. Quelle [103] discussed some properties of optimal policies for discrete time MDPs, and Dong and Liu [28] discussed the structure of optimal policies in discounted semi-Markov decision processes with unbounded rewards.

In contrast to the usual method in the literature, this chapter discussed what properties the MDP model will have under the necessary condition which says that the criterion is well defined. The contents of this chapter are from [74] and [76].

Problems

1. Consider the discounted expected total reward in finite horizons. Let N be the number of horizons. Please discuss whether or not Condition 2.2 can be derived from Condition 2.1. Check this problem again when N is a random variable, which is independent of the MDP models (we call this case a stochastic terminated MDP).

2. A Stock Option Model. Consider the problem of buying an option for a given stock. Let P_n be the price of the stock at the nth day for $n = 0, 1, 2, \cdots$. Suppose that $\{P_n\}$ satisfies the random-walk model. That is, there are independent random variables $\xi_1, \xi_2, \cdots$ with identical distribution function F such

that

$$P_{n+1} = P_n + \xi_n, \ \ n \geq 0.$$

Here, P_0 is the initial price and is independent of $\{\xi_n, n \geq 0\}$. Moreover, we suppose that one has an option to buy one share of the stock at a fixed price r at the initial day and then exercise the option in some day in the future.

Surely, a strategy is to tell when to exercise the option. This is obviously based on the price of the stock, if r is given. Hence, one's problem is to find a strategy to maximize his expected profit from exercising the option.

Let $V(p)$ be the maximal expected revenue when the current price of the stock is p.

1) Write the optimality equation.

2) Show $V(p) - p$ to be decreasing in p under the condition that the mean of ξ_n is finite.

3) Show under the condition that the mean of ξ_n is finite, that the optimal strategy to be as follows: if the current price is p then to exercise the option if and only if $p \geq p^*$, for some number p^*.

4) Whether or not the results above are still true when the mean of ξ_n is infinite.

3. Sequential Investment Problem. Suppose one has an amount M money and consider to invest his money in the future N periods. But the opportunity for investment is not deterministic. Suppose that at each period, an investment opportunity occurs with probability p, which is independent of the past and of the amount of the remaining money. When an investment opportunity occurs, if he invests x then he will earn a revenue $r(x)$, including his investment. Assume that both his investment and his return at any period cannot be invested again in the future. What is the optimal strategy for this problem?

Let $V_n(X)$ be the maximal expected profit when there remain n periods, X money for the future investment and an investment opportunity occurs.

1) Write the optimality equation.

2) Assume that $r(x)$ is nondecreasing, concave and satisfies $r(0) = 0$. Show that $V_n(X)$ is also concave in X.

4. Sequential Consumption Problem. An individual will earn ξ_n at the nth period for $n = 0, 1, \cdots, N$, where N may be interpreted as his life. His problem is to determine at each period how much from his wealth he can consume. Let W be his wealth at the beginning of a period, then his consumption y should not be larger than W. If he consumes y, then his utility is $u(y)$. His objective is to maximize his expected total utility during his life (this implies that any wealth after his life remains no utility for him). Assume ξ_n are independent and identically distributed according to F. Please set this problem up as a MDP and write the optimality equation. Moreover, if the individual knows his life is

random with the distribution function G, then what are the MDP model and its corresponding optimality equation?

5. Sequential Investment/Consumption Problem. We combine the problems discussed in Problem 3 and Problem 4. An individual will earn ξ_n at the nth period for $n = 0, 1, \cdots, N$. At each period, he can consume some from his wealth and invest his remaining wealth. Assume that the return $r(x)$ is obtained and can be used by the individual at the beginning of the next period. Hence, at the beginning of each period, he should determine his consumption y and his investment x with $x + y \leq W$, in order to maximize his expected total utility during his life. Please set this problem up as a MDP model and write the corresponding optimality equation.

6. Optimal Stopping Problem. The situation of an individual is divided into several states $i = 0, 1, 2, \cdots$. At each period with state i, he can choose to stop, in which case he receives $r(i)$ and the problem terminates, or to pay $c(i)$, in which case the problem enters into the next period at which the state becomes j with probability p_{ij}. We assume that both the terminate reward $r(i)$ and the continuation cost $c(i)$ are nonnegative.

1) Please set the above problem as a Markov decision process model and write the optimality equation. (Note: one can introduce a state ∞ to represent that the problem is terminated.)

2) Let $R = \sup_i r(i)$ and $C = \inf_i c(i)$. Show that the optimal value of the finite horizons tends to that of the infinite horizons when $R < \infty$ and $C > 0$.

3) Let S^* be a state subset such that it is optimal to stop whenever the state is in S^*. Discuss when S^* is closed, i.e., $p_{ij} = 0$ for all $i \in S^*$ and $j \notin S^*$?

7. Inventory. Consider a seller who sells some product to buyers. The demand for the product for each period is j with probability p_j for $j = 0, 1, \cdots$. In order to satisfy the demand, the seller should order the products from his supplier. Suppose that a fixed cost K is incurred for each order and a cost c for each item, while the holding cost for each item in one period is h and the penalty cost for each shortage item in one period is p. Moreover, the shortage items is backordered and is satisfied when additional inventory becomes available (in this case the inventory is negative). Please help the seller to determine when does he should to order and how many will he order to minimize his expected total cost, by using Markov decision processes.

8. How to test Software for Optimal Software Reliability Assessment. In the final phase of software testing, the concerned software is subjected to testing for validation or acceptance and reliability assessment is conducted by using the resulting test data. For highly reliable software, only few failures can be observed in the final phase of the software testing, and the number of tests that can be applied is usually limited because of stringent schedule constraints. The

software code is frozen. No debugging is invoked during the testing even if software failures are observed. No modifications are applied to the software under the test. The central problem of concern is how to apply a given number of tests such that the resulting software reliability estimate is best trustable. Since observed failures are few, point estimations of software reliability may suffer large fluctuations and it is desirable to keep the corresponding variance under control. Ideally, the variance of a software reliability estimator is minimized so that the resulting estimate may be as stable as possible. In order to formulate the problem of concern, we have the following assumptions.

1) The software under test or reliability assessment is frozen.
2) The input domain of the software under test comprises m equivalence classes of test cases or input values, denoted $C_1, C_2, \ldots, C_m$.
3) Each test case or input value will lead the software under test to failure or success, and

$$P\{\text{a software failure is observed} \mid \text{a test case or input value of } C_i \text{ is applied}\} = \theta_i,$$

   for $i = 1, 2, \ldots, m$.
4) The output of the software under test when executed against a selected test case is independent of the history of testing.
5) A software test includes selecting a test case or an input value from the input domain, executing the test case, collecting the resulting testing data, and updating the software reliability estimate if necessary. A total of n tests are allowed to be used to test the software.
6) All actions or distinct software tests are admissible each time.
7) The operational profile of the software under reliability assessment can be described as $\langle C_i, p_i, i = 1, 2, \ldots, m\rangle$, that is, p_i denotes the probability that an input value is selected from C_i in the phase of software operation, and $\sum_{i=1}^{m} p_i = 1$.

Set this problem up as a Markov decision process model. (Note: This problem is from Cai, Li and Liu [14]. Interested readers may read the article for the details.)

Chapter 3

DISCRETE TIME MARKOV DECISION PROCESSES: AVERAGE CRITERION

In this chapter, we study average optimality in the discrete time Markov decision processes with countable state space and measurable action sets. The average criterion differs from the discounted criterion. In the discounted criterion, the reward at period n should be discounted to period 0 by multiplying β^n. Hence, the smaller the period n is, the more important the reward of period n in the criterion will be. The reverse is also true; that is, the larger the period n is, the less important the reward of period n in the criterion will be. Contrary to it, in the average criterion, the reward in any period accounts for nothing in the criterion. Here, only the future trend of the reward is considered. Therefore, in the discounted criterion the former horizons are mainly considered, whereas in the average criterion only the future horizons are considered. This difference leads to different methods to study the average criterion. For the discounted criterion, we use methods for the convergence of series, and for the average criterion, we use many more results, especially the limiting properties, in Markov chains. Hence, we meet many more difficulties, which give richer results for the average criterion. The contents of this chapter consist of three parts. First, some lemmas from mathematical analysis and Markov chain theory are given. Then, the optimality equation and the optimality inequality for the average criterion are studied with sufficient condition to ensure them, respectively, and the standard results are obtained for both cases.

1. Model and Preliminaries

The model of the discrete time Markov decision processes discussed in this chapter is as follows,

$$\{S, (A(i), \mathcal{A}(i)), p_{ij}(a), r(i,a), V\}. \tag{3.1}$$

The first four elements are the same as those in Chapter 2 for the total reward criterion. That is, the state space S is countable and the action set $A(i)$ is nonempty with a measurable structure $\mathcal{A}(i)$ the same as that in the last chapter. When the system is in state i and an action $a \in A(i)$ is taken at some period, the system will transfer to state j at the next period with probability $p_{ij}(a)$, and incur a finite reward $r(i,a)$. But it is assumed that $\sum_j p_{ij}(a) = 1$ for all $(i,a) \in \Gamma$. In fact, if this is not true then we can introduce a fictitious state i^* and a fictitious action a^* such that $A(i^*) = \{a^*\}$, $p_{i^*i^*}(a^*) = 1, r(i^*,a^*) = 0$, $p_{ii^*}(a) = 1 - \sum_j p_{ij}(a)$, and $r(i,a^*) = 0$ for all $(i,a) \in \Gamma$. Surely, this would not influence the values of the average criterion and the discounted criterion.

The problem is to maximize the long-running average reward per unit time, $V(\pi,i)$, which is defined by

$$V(\pi,i) = \liminf_{N\to\infty} \frac{1}{N+1} V_{1,N}(\pi,i), \quad i \in S, \quad \pi \in \Pi. \tag{3.2}$$

The optimal value is defined by

$$V^*(i) = \sup\{V(\pi,i) | \pi \in \Pi\}, \quad i \in S.$$

A policy $\pi^* \in \Pi$ is optimal if $V(\pi^*,i) = V^*(i)$ for all $i \in S$.

We need several lemmas for studying the average criterion. First, we introduce the Fatou lemma and Control Convergence Theorem, which are famous for changing order of limits and integrations (Theorems 3.3 and 3.4, page 104 in [157]).

Lemma 3.1: *For a series of Lebesgue measurable functions* $\{f_n(x)\}$,

1. (Fatou lemma) *If* $f_n(x)$ *is nonnegative for each* $n = 0, 1, 2, \ldots$ *then*

$$\int_0^\infty \liminf_{n\to\infty} f_n(x)dx \le \liminf_{n\to\infty} \int_0^\infty f_n(x)dx.$$

2. (Control Convergence Theorem) *Suppose that* $f(x) = \lim_{n\to\infty} f_n(x)$ *exists and there is a measurable and integrable function* $g(x)$ *such that* $|f_n(x)| \le g(x)$ *for all* x *and* n. *Then,* $f(x)$ *is also integrable and*

$$\int_0^\infty \lim_{n\to\infty} f_n(x)dx = \lim_{n\to\infty} \int_0^\infty f_n(x)dx.$$

Lemma 3.2 (Jensen Inequality): *For any convex function* $f(x)$ *and a random variable* X, *the following inequality is true whenever the expectations in it are well defined,*

$$Ef(X) \ge f(EX).$$

The above lemma (Corollary 1, page 103 in [22]) concerns the change of order of the expectation and the function f. The following lemma is standard for the limiting properties in the theory of Markov chains, where $P(\pi_0) = (P_{ij}(\pi_0))$ is a matrix with $P_{ij}(\pi_0) = \int_{A(i)} p_{ij}(a)\pi_0(da|i)$ for $i, j \in S$. One can find it in Theorems 5.4.4 and 5.4.5, page 124 in [93].

Lemma 3.3: *For any stochastic stationary policy π_0, the limit $P^*(\pi_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n(\pi_0)$ exists and*

$$P^*(\pi_0)P(\pi_0) = P(\pi_0)P^*(\pi_0) = P^*(\pi_0)P^*(\pi_0) = P^*(\pi_0).$$

Moreover, when the reward function is bounded,

$$V(\pi_0, i) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} E_{\pi_0,i} r(X_n, \Delta_n) = [P^*(\pi_0) r(\pi_0)]_i, \quad i \in S.$$

The next lemma is the discrete type of Abel theorem (Theorem 1, page 181 in [149]).

Lemma 3.4 (Abel Theorem): *Suppose that $\{C_n, n = 0, 1, 2, \ldots\}$ is a series. Then,*

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} C_n \le \liminf_{\beta\uparrow 1} (1-\beta) \sum_{n=0}^{\infty} \beta^n C_n,$$

where the equality is true whenever the limit in the left-hand side exists.

The importance of the lemma above is that it sets up a relationship between the discounted criterion and the average criterion. For any given policy π and state i, if we let $C_n = E_{\pi,i} r(X_n, \Delta_n)$ for $n \ge 0$, then we get from the Abel theorem that

$$\liminf_{\beta\uparrow 1} (1-\beta) V_\beta(i) \ge \liminf_{\beta\uparrow 1} (1-\beta) V_\beta(\pi, i) \ge V(\pi, i), \quad i \in S, \pi \in \Pi.$$

By taking the supremum over π above, we get

$$\liminf_{\beta\uparrow 1} (1-\beta) V_\beta(i) \ge V^*(i), \quad i \in S. \tag{3.3}$$

Moreover, for any stationary policy f and state $i \in S$, we have due to Lemma 3.3,

$$V(f, i) = \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} E_{f,i} r(X_n, \Delta_n),$$

where the limit exists. Hence, due to the Abel theorem,

$$V(f,i) = \lim_{\beta\uparrow 1}(1-\beta)V_\beta(f,i), \;\; i \in S.$$

With this, we may conjecture that the inequality in Eq. (3.3) may become an equality, at least under certain conditions.

Without considering any condition, we assume first that all limits

$$\rho(i) := \lim_{\beta\uparrow 1}(1-\beta)V_\beta(i), \;\; i \in S$$

exist. Due to the above discussions, $\rho(i)$ may be the optimal value funciton for the average criterion; that is, $\rho(i) = V^*(i)$ for $i \in S$. By multiplying both sides of the discounted optimality equation by $(1-\beta)$, we obtain

$$(1-\beta)V_\beta(i) = \sup_{a\in A(i)}\{(1-\beta)r(i,a) + \beta\sum_j p_{ij}(a)(1-\beta)V_\beta(j)\}, \;\; i \in S.$$

Letting $\beta \uparrow 1$ results in (we assume that the orders of the limit with $\sup_{a\in A(i)}$ and with $\sum_j$ can be changed)

$$\rho(i) = \sup_{a\in A(i)}\sum_j p_{ij}(a)\rho(j), \;\; i \in S. \tag{3.4}$$

On the other hand, we subtract both sides of the discounted optimality equation by $\beta V_\beta(i)$ and get

$$\begin{aligned}(1-\beta)V_\beta(i) &= V_\beta(i) - \beta V_\beta(i) \\ &= \sup_{a\in A(i)}\{r(i,a) + \beta\sum_j p_{ij}(a)V_\beta(j) - \beta V_\beta(i)\} \\ &= \sup_{a\in A(i)}\{r(i,a) + \beta\sum_j p_{ij}(a)[V_\beta(j) - V_\beta(i)]\}, \;\; i \in S.\end{aligned}$$

In order to take the limit with $\beta \uparrow 1$ in the above equation, we assume that for some fixed state $0 \in S$ all limits

$$h(i) = \lim_{\beta\uparrow 1}[V_\beta(i) - V_\beta(0)], \;\; i \in S$$

exist. Then, by taking $\beta \uparrow 1$ in the former equation (it is still assumed that the orders of the limits with $\sup_{a\in A(i)}$ and with $\sum_j$ can be changed), we obtain that

$$\begin{aligned}\rho(i) &= \sup_{a\in A(i)}\{r(i,a) + \sum_j p_{ij}(a) \\ &\quad \cdot \lim_{\beta\uparrow 1}\{[V_\beta(j) - V_\beta(0)] - [V_\beta(i) - V_\beta(0)]\}\}\end{aligned}$$

$$= \sup_{a\in A(i)} \{r(i,a) + \sum_j p_{ij}(a)[h(j) - h(i)]\}$$
$$= \sup_{a\in A(i)} \{r(i,a) + \sum_j p_{ij}(a)h(j)\} - h(i), \quad i \in S.$$

This is equivalent to

$$\rho(i) + h(i) = \sup_{a\in A(i)} \{r(i,a) + \sum_j p_{ij}(a)h(j)\}, \quad i \in S. \tag{3.5}$$

For the average criterion, the optimality equations consist of Eq. (3.4) and Eq. (3.5), where the variables are $\{\rho(i), h(i), i \in S\}$. We call these two equations the set of average criterion optimality equations (ACOEs for short).

We have seen from the above discussions a deep relation between the discounted criterion and the average criterion. In the next section, we discuss it in detail for a special case.

2. Optimality Equation

In this section, we consider a special case where $\rho(i)$ is a constant; that is, $\rho(i) = \rho$ for all $i \in S$. In this case, Eq. (3.4) becomes trivial, and Eq. (3.5) can be simplified as

$$\rho + h(i) = \sup_{a\in A(i)} \{r(i,a) + \sum_j p_{ij}(a)h(j)\}, \quad i \in S, \tag{3.6}$$

where ρ is a constant and h is a function on S. A solution of the above equation is $\{\rho, h(i), i \in S\}$. We call Eq. (3.6) the average criterion optimality equation (ACOE for short). We interpret $h(i)$ as a terminal reward at state i; that is, the system will receive a reward $h(i)$ if it terminates at state i. We consider two systems. The first one is a two-period system, which runs in the first period as that in the MDPs and then terminates in the next period. Then its optimal expected reward is exactly the right-hand side of ACOE. The second one is a one-period system in which the one-period reward is simply the constant ρ also with the same terminating reward $h(i)$. So, its reward is just $\rho + h(i)$ when the initial state is i. Hence, ACOE says that these two systems result in the same expected reward.

In the following, we first discuss properties of ACOE and optimal policies. Then, we present a set of conditions to ensure existence of solutions of ACOE. For the main condition in the set, we give a sufficient recurrent condition based on the theory of Markov chains.

2.1 Properties of ACOE and Optimal Policies

In this subsection, we assume that there is a solution of ACOE. For any solution $\{\rho, h\}$ of ACOE, we let for $(i, a) \in \Gamma$,

$$G(i,a) = \rho + h(i) - \{r(i,a) + \sum_j p_{ij}(a)h(j)\}$$

be the deviation when we use the one-period system to approximate the two-period system when action a is chosen. Then, $G(i, a)$ is nonnegative and its infinimum over $a \in A(i)$ is zero. Certainly, $G(i, a)$ depends on the solution $\{\rho, h\}$. But we do not point out the concrete ρ, h hereafter inasmuch as it is clear from context.

Similarly to the discounted criterion, we let

$$A^*(i) = \{a \in A(i) | G(i,a) = 0\}$$

be the optimal action set at state $i \in S$.

For any stochastic stationary policy $\pi_0 \in \Pi_s$, Let $R(\pi_0)$ be the set of all positive recurrent states [23] under the Markov chain with state transition matrix $P(\pi_0)$. Let $G(\pi_0, i) = \int_{a \in A(i)} G(i,a)\pi_0(da|i)$ for $i \in S$.

The first main theorem in this subsection is the following one.

Theorem 3.1: *Suppose that $V(\pi, i)$ is well defined for each policy π and state i and that ACOE (3.6) has a solution $\{\rho, h\}$ satisfying the following condition,*

$$\lim_{n\to\infty} \frac{1}{n} E_{\pi,i} h(X_n) = 0, \quad \pi \in \Pi_m, \quad i \in S. \tag{3.7}$$

Then,

1. *$\rho = V^*(i)$ for all $i \in S$, and for each $\varepsilon \geq 0$, any f that achieves ε-supremum of ACOE (3.6) is ε-optimal.*
2. *If the limits of the right-hand side in the following equation exist, then policy π is optimal if and only if it satisfies the following condition,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} E_{\pi,i} G(X_n, \Delta_n) = 0, \quad i \in S.$$

3. *If a stochastic stationary policy π_0 is optimal, then $\pi_0(A^*(i)|i) = 1$ for each positive recurrent state $i \in R(\pi_0)$. The reverse is also true if the state space S is finite or $\pi_0(A^*(i)|i) = 1$ for each state $i \in S$.*

Proof: It follows from Lemma 2.7 in Chapter 2 that the condition (3.7) is also true for any policy $\pi \in \Pi$. Due to the definition of $G(i, a)$, we have for each

$(i,a) \in \Gamma$ that

$$\rho + h(i) = r(i,a) + \sum_j p_{ij}(a)h(j) + G(i,a).$$

Thus, for each $n \geq 0$,

$$\rho + h(X_n) = r(X_n, \Delta_n) + \sum_j p_{X_n,j}(\Delta_n)h(j) + G(X_n, \Delta_n).$$

For any policy π and state i, by taking expectation $E_{\pi,i}$ in the above equation and summing it from $n = 0$ to $N-1$, we get

$$\begin{aligned} \rho &= \frac{1}{N}\sum_{n=0}^{N-1} E_{\pi,i} r(X_n, \Delta_n) + \frac{1}{N}\sum_{n=0}^{N-1} E_{\pi,i} G(X_n, \Delta_n) \\ &\quad + \frac{1}{N} E_{\pi,i} h(X_N) - \frac{1}{N} h(i). \end{aligned} \tag{3.8}$$

By letting $\liminf_{N\to\infty}$ above, we know from Eq. (3.7) that for $i \in S$,

$$\rho \geq V(\pi, i) + \liminf_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} E_{\pi,i} G(X_n, \Delta_n). \tag{3.9}$$

1. Because G is nonnegative, we obtain $\rho \geq V^*(i)$ for all $i \in S$ due to the arbitrariness of π from the above equation. On the other hand, we know from $\inf_a G(i,a) = 0$ that for each $\varepsilon > 0$ there is f such that $G(i,f) \leq \varepsilon$ for each $i \in S$. Then, due to (3.8) we know that for $i \in S$,

$$\rho \leq \frac{1}{N}\sum_{n=0}^{N-1} E_{f,i} r(X_n, \Delta_n) + \varepsilon + \frac{1}{N} E_{f,i} h(X_N) - \frac{1}{N} h(i).$$

This implies by letting $\liminf_{N\to\infty}$ that

$$\rho \leq V(f,i) + \varepsilon \leq V^*(i) + \varepsilon, \quad i \in S.$$

Moreover, letting $\varepsilon \to 0^+$ we get $\rho \leq V^*(i)$ for all $i \in S$. Hence, $\rho = V^*(i)$ for all $i \in S$, and the above f is an ε-optimal policy.

2. It can be proved from 1 and Eq. (3.8).

3. For any stochastic stationary policy π_0, it is easy to verify that

$$P_{\pi_0}\{X_n = j | X_0 = i\} = P_{ij}^n(\pi_0), \quad i,j \in S, \quad n \geq 0.$$

Then, it follows from the Fatou lemma that

$$\liminf_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} E_{\pi_0,i} G(X_n, \Delta_n)$$

$$
\begin{aligned}
&= \liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} [P^n(\pi_0)G(\pi_0)]_i \\
&= \liminf_{N\to\infty} \sum_j \frac{1}{N} \sum_{n=0}^{N-1} P_{ij}^n(\pi_0)G(\pi_0, j) \\
&\geq \sum_j \liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} P_{ij}^n(\pi_0)G(\pi_0, j) \\
&= [P^*(\pi_0)G(\pi_0)]_i, \quad i \in S. \qquad (3.10)
\end{aligned}
$$

It is well known in the theory of Markov chains that

$$
P_{ij}^*(\pi_0) \begin{cases} > 0, & i = j \in R(\pi_0), \\ = 0, & i \in S, j \notin R(\pi_0). \end{cases} \qquad (3.11)
$$

With this and Eq. (3.9) we get

$$
\rho \geq V(\pi_0, i) + \sum_{j\in R(\pi_0)} P_{ij}^*(\pi_0)G(\pi_0, j), \quad i \in S. \qquad (3.12)
$$

If there is state $j_0 \in R(\pi_0)$ such that $\pi_0(A^*(j_0)|j_0) < 1$, then $G(\pi_0, j_0) > 0$. So, by taking $i = j_0$ in Eq. (3.12) we get

$$
\rho \geq V(\pi_0, j_0) + P_{j_0,j_0}^*(\pi_0)G(\pi_0, j_0) > V(\pi_0, j_0).
$$

This implies that π_0 is not optimal. Hence, $\pi_0(A^*(i)|i) = 1$ for each $i \in R(\pi_0)$.

If S is finite, then the inequality in Eq. (3.10) becomes an equality and the limit in the definition of $V(\pi, i)$ exists. So, the inequality in Eq. (3.12) also becomes an equality. Moreover, if $\pi_0(A^*(j)|j) = 1$ for all $j \in R(\pi_0)$, then $G(\pi_0, j) = 0$ for $j \in R(\pi_0)$. Due to (3.12), $\rho = V(\pi_0, i)$ for $i \in S$. That is, π_0 is optimal.

If $\pi_0(A^*(j)|j) = 1$ for all $j \in S$, then $E_{\pi_0,i}G(X_n, \Delta_n) = 0$ for all $i \in S$. With this and Eq. (3.8) and Eq. (3.7), we know that π_0 is optimal. □

For the discounted criterion, the optimality equation has a unique solution under certain conditions. But this is no longer true here. It is easy to see from ACOE (3.6) that if $\{\rho, h\}$ is a solution of ACOE, then $\{\rho, h+c\}$ is also a solution for any constant c. Result 1 in Theorem 3.1 says that if h satisfies the condition Eq. (3.7) (e.g., h is bounded) then $\rho = V^*(i)$ is unique. In solutions of ACOE, ρ is often the optimal value function, and h is not unique. The condition given in 2 in Theorem 3.1 means that there is no long-running average deviation when approximating the two-period system by the one-period system.

The condition given in Eq. (3.7) can be weakened if we have $\rho \geq V^*(i)$ for all $i \in S$. This can be proved similarly to Theorem 3.1. The conclusion that $\rho \geq V^*(i)$ for $i \in S$ can be seen in Eq. (3.3) and Lemma 3.5 in Subsection 3.2.

Corollary 3.1: *Suppose that ACOE has a solution* $\{\rho, h\}$ *satisfying* $\rho \geq V^*(i)$ *for all* $i \in S$. *Then,*

1. *For* $\varepsilon \geq 0$, *if* f *attains* ε*-supremum of ACOE and*

$$\limsup_{n\to\infty} \frac{1}{n} E_{f,i} h(X_n) \leq 0, \quad i \in S, \tag{3.13}$$

then f *is* ε*-optimal.*

2. *If* f *is optimal and satisfies the condition given in Eq. (3.7), then* $f(i)$ *attains the supremum of ACOE for* $i \in R(f)$. □

Under the condition that $\rho \geq V^*(i)$ for all $i \in S$, Eq. (3.7) is weakened into Eq. (3.13) in the above corollary. For Theorem 3.1 and Corollary 3.1, we give the following remark.

Remark 3.1:

1. *When* h *is bounded, the condition given in Eq. (3.7) is true, and we say that ACOE has a bounded solution.*

2. *If the inequality in Eq. (3.10) becomes an equality for a policy* π_0, *then the first condition in 3 of Theorem 3.1 becomes a necessary and sufficient condition; that is,* π_0 *is optimal if and only if* $\pi_0(A^*(i)|i) = 1$ *for all* $i \in R(\pi_0)$.

3. *We know from Corollary 3.1 that a stationary policy* f *will be optimal if* $f(i)$ *attains the supremum of the optimality equation for positive recurrent state* i, *irrespectively of the nonpositive recurrent states. This differs from the discounted criterion, where it is required that* $f(i)$ *attain the supremum of the optimality equation for each state.*

Comparing with the properties for the discounted criterion (see Theorems 2.5 and 2.6), we have weaker properties for the average criterion, as given in Theorem 3.1 and Corollary 3.1 above. This is further shown in the following example.

Example 3.1: Suppose that the state space is $S = \{1, 2, 3\}$, the action sets are $A(1) = A(2) = \{a, b\}$ and $A(3) = \{a\}$, the state transition probability is given by $p_{12}(a) = p_{23}(a) = p_{33}(a) = 1, p_{13}(b) = p_{21}(b) = 1$, and the reward function is given by $r(1, a) = r(2, a) = r(1, b) = r(2, b) = 0, r(3, a) = 1$.

Let policies be $f_1 = (a, a, a)$, $f_2 = (b, b, a)$, and $f = (a, b, a)$. Then, it is easy to compute that

$$V(f_1) = V(f_2) = (1, 1, 1), \quad V(f) = (0, 0, 1).$$

So, both f_1 and f_2 are optimal policies, whereas f is not optimal, although $f(1) = f_1(1)$, $f(3) = f_1(3)$, and $f(2) = f_2(2)$.

From conclusion 1 in Theorem 3.1 we know that, under certain conditions, the constant ρ in the solutions of the ACOE (3.6) is the optimal value for the average criterion, and is independent of the initial state i. This differs from the discounted criterion, where the optimal value depends on the initial state. The reason for it is that in the average criterion, only the future horizons are considered. But the system in the future horizons will be independent of the initial state, similarly to whether a state is recurrent is independent of the initial state in Markov chains.

2.2 Sufficient Conditions

A natural question is when the ACOE (3.6) has a solution $\{\rho, h\}$. We know from leading Eq. (3.5) that the key for the question is under what conditions the limits $h(i) = \lim_{\beta\uparrow 1}[V_\beta(i) - V_\beta(i_0)]$ and $\rho(i) = \lim_{\beta\uparrow 1}(1-\beta)V_\beta(i)$ exist and the order of the $\lim_{\beta\uparrow 1}$ with $\sup_{a\in A(i)}$ and with $\sum_j$ can be changed. For this, we introduce the following condition.

Condition 3.1: *Suppose there is a sequence of discount factors $\{\beta_m, m \geq 1\}$ that is increasing and tends to 1, a nonnegative function L on S, a positive constant M, and a state $0 \in S$ such that*

1. *$r(i,a)$ is bounded above and below in $a \in A(i)$ for each i, moreover, V_{β_m} satisfies the discounted optimality equation for each $m \geq 1$.*
2. *$|h_m(i)| \leq L(i)$ for all $i \in S$ and $m \geq 1$, where $h_m(i) = V_{\beta_m}(i) - V_{\beta_m}(0)$ is called the relative value function.*
3. *$|(1-\beta_m)V_{\beta_m}(0)| \leq M$ for all $m \geq 1$.*
4. *$\sum_j p_{ij}(a)L(j)$, as a series, is convergent uniformly in $a \in A(i)$, for each $i \in S$.*

For the condition above, we make the following remark.

Remark 3.2:

1. *1 of Condition 3.1 implies that both $\sup_a r(i,a)$ and $\inf_a r(i,a)$ are finite, and for the discounted optimality equation, it is discussed in details in Chapter 2.*
2. *2 of Condition 3.1 is equivalent to the following condition. There is a nonnegative function $L(i,j)$ such that $|V_{\beta_m}(i) - V_{\beta_m}(j)| \leq L(i,j)$ for all $i, j \in S$ and $m \geq 1$. So, the state 0 can be arbitrary.*
3. *4 of Condition 3.1 will be true under one of the following conditions. (a) The state space S is finite. (b) One-step reachable state set $\{j|$ there is an action $a \in A(i)$ such that $p_{ij}(a) > 0\}$ from state i is finite for each $i \in S$ (which is true in many queueing systems and inventory systems). (c) The action sets $A(i)$ are all finite and $\sum_j p_{ij}(a)L(j)$ is convergent for each $a \in A(i)$.*

Under the above condition, we have the following theorem on solutions of the ACOE.

Theorem 3.2: *Under Condition 3.1, there is a solution* $\{\rho, h\}$ *of ACOE that satisfies* $|h(i)| \leq L(i)$ *for all* $i \in S$.

Proof: From 1 of Condition 3.1, V_{β_m} satisfies the discounted optimality equation. Hence, for each $m \geq 1$, there is $f_m \in F$ that attains $(1/m)$-supremum of the optimality equation; that is,

$$\begin{aligned} V_{\beta_m}(i) &= \sup_a \{r(i,a) + \beta_m \sum_j p_{ij}(a) V_{\beta_m}(j)\} \\ &\leq r(i,f_m) + \beta_m \sum_j p_{ij}(f_m) V_{\beta_m}(j) + \frac{1}{m}, \ \ i \in S. \end{aligned}$$

Let $u_m = (1-\beta_m) V_{\beta_m}(0)$ for $m \geq 1$. Then,

$$\begin{aligned} u_m + h_m(i) &= \sup_a \{r(i,a) + \beta_m \sum_j p_{ij}(a) h_m(j)\} \\ &\leq r(i,f_m) + \beta_m \sum_j p_{ij}(f_m) h_m(j) + \frac{1}{m}, \ i \in S. \quad (3.14) \end{aligned}$$

Based on this and Condition 3.1, we can get by using the diagonalization method that there is a subsequence $\{\beta_{m_k}\}$ of $\{\beta_m\}$ such that all the following limits exist,

$$\begin{aligned} &\lim_k u_{m_k} = \rho, && \lim_k h_{m_k}(i) = h(i), \ \ i \in S, \\ &\lim_k r(i, f_{m_k}), \ \ i \in S, && \lim_k p_{ij}(f_{m_k}), \ \ i,j \in S. \end{aligned}$$

For convenience, we assume that the sub-sequence $\{\beta_{m_k}\}$ is exactly the sequence $\{\beta_m\}$ itself. Then, by letting m tend to infinity in Eq. (3.14), we get from 4 of Condition 3.1 and the Fatou lemma that

$$\begin{aligned} \rho + h(i) &\leq \lim_m r(i,f_m) + \sum_j \lim_m p_{ij}(f_m) h(j) \\ &= \lim_m \{r(i,f_m) + \sum_j p_{ij}(f_m) h(j)\} \\ &\leq \sup_a \{r(i,a) + \sum_j p_{ij}(a) h(j)\}, \ \ i \in S. \end{aligned}$$

From the equality in Eq. (3.14) and Condition 3.1.4 again, we get

$$\begin{aligned} \rho + h(i) &\geq \sup_a \lim_m \{r(i,a) + \sum_j p_{ij}(a) h_m(j)\} \\ &\geq \sup_a \{r(i,a) + \sum_j p_{ij}(a) h(j)\}, \ \ i \in S. \end{aligned}$$

Hence, ACOE has a solution $\{\rho, h\}$. Certainly, $|h(i)| \leq L(i)$ for all $i \in S$ from 2 of Condition 3.1. □

In general, if the system is ergodic, that is, the system under each policy has the unique positive recurrent subchain, then the optimal value is a constant and the optimality equation is Eq. (3.6). Otherwise, the optimal value depends on the initial state and so the optimality equations are ACOEs (3.4) and (3.5), which are not discussed in this book. Interested readers can find it in other books and papers.

Remark 3.3: *A simpler case for Condition 3.1 is that $h_m(i)$ bounded and $\{\lim_m p_{ij}(f_m),\ j \in S\}$ are still a probability distribution for each $i \in S$. In this case, there is a bounded solution of ACOE. It is shown in Fernandez-Gaucher and Marcus [41] that when ACOE has a bounded solution, $h_\beta(i) := V_\beta(i) - V_\beta(0)$ is bounded uniformly in β and i.*

Combining Theorems 3.1 and 3.2, we can get conditions for the existence of ε-optimal policies. Such a case is given in the following corollary due to Corollary 3.1.

Corollary 3.2: *Suppose Condition 3.1 is true, $\varepsilon \geq 0$, f attains ε-supremum of ACOE, and $\lim_{n\to\infty}(1/n)E_{f,i}L(X_n) = 0$ for each $i \in S$. Then, f is ε-optimal.* □

2.3 Recurrent Conditions

In Condition 3.1, the key is the existence of the function $L(i)$. For this, we give a recurrent condition based on Markov chains. Here, it is assumed that the reward function is uniformly bounded with a upper bound M.

For any state subset $D \subset S$, we define

$$T_D = \min\{n \geq 1 | X_n \in D\},$$

the first arrival time the system enters the state set D. When $D = \{i\}$ is a singleton, we write $T_D = T_i$. The recurrent condition is given as follows.

Condition 3.2: *Suppose that there is a sequence of discount factors $\{\beta_m, m \geq 1\}$ which is increasing and tends to 1, a sequence of policies $\{f_m \in F, m \geq 1\}$, and a positive function h on S such that for each $m \geq 1$,*

1. *f_m is m^{-1}-optimal for the discounted criterion with the discount factor β_m.*
2. *There is a state $s_m \in S$ such that*

$$E_{f_m,i}T_{s_m} \leq w(i), \quad i \in S. \tag{3.15}$$

We have the following theorem.

Theorem 3.3: *Under Condition 3.2, for each pair of states $i, j \in S$ there is a positive number $L(i,j)$ such that $|V_{\beta_m}(i) - V_{\beta_m}(j)| \le L(i,j)$ for all $m \ge 1$. Therefore, Condition 3.1.2 is true by letting $L(i) = L(i,0)$.*

Proof: For each $m \ge 1$, let $T_m = T_{s_m}$ for convenience. From the definition of the discounted criterion and the fact of $X_{T_m} = s_m$, we have that for $i \in S$,

$$\begin{aligned} V_{\beta_m}(f_m, i) &= E_{f_m,i} \sum_{n=0}^{T_m-1} \beta_m^t r(X_n, \Delta_n) + E_{f_m,i} \sum_{n=T_m}^{\infty} \beta_m^t r(X_n, \Delta_n) \\ &= E_{f_m,i} \sum_{n=0}^{T_m-1} \beta_m^t r(X_n, \Delta_n) + E_{f_m,i} \beta_m^{T_m} V_{\beta_m}(f_m, s_m). \end{aligned}$$

Hence,

$$\begin{aligned} & |V_{\beta_m}(f_m, i) - V_{\beta_m}(f_m, s_m)| \\ \le\ & |E_{f_m,i} \sum_{n=0}^{T_m-1} \beta_m^t r(X_n, \Delta_n)| + |V_{\beta_m}(f_m, s_m)|(1 - E_{f_m,i} \beta_m^{T_m}) \\ :=\ & I_1 + I_2. \end{aligned}$$

By the boundedness of the reward function and Eq. (3.15),

$$I_1 \le E_{f_m,i}(MT_m) \le Mw(i).$$

It is easy to see that $(\beta_m)^x$ is a convex function in x for each $m \ge 1$. Then, from Lemma 3.2 (Jensen Inequality) and Eq. (3.15), we get

$$\begin{aligned} I_2 &\le (1-\beta_m)^{-1} M (1 - \beta_m^{E_{f_m,i} T_m}) \\ &\le (1-\beta_m)^{-1} M (1 - \beta_m^{w(i)}) \\ &\le Mw(i). \end{aligned}$$

Therefore, $|V_{\beta_m}(f_m, i) - V_{\beta_m}(f_m, s_m)| \le 2Mw(i)$, and so

$$\begin{aligned} & |V_{\beta_m}(i) - V_{\beta_m}(j)| \\ \le\ & |V_{\beta_m}(i) - V_{\beta_m}(f_m, i)| + |V_{\beta_m}(f_m, i) - V_{\beta_m}(f_m, s_m)| \\ & + |V_{\beta_m}(f_m, s_m) - V_{\beta_m}(f_m, j)| + |V_{\beta_m}(f_m, j) - V_{\beta_m}(j)| \\ \le\ & \frac{1}{m} + 2Mw(i) + 2Mw(j) + \frac{1}{m} \\ \le\ & 2M[w(i) + w(j)] + 2, \quad \forall i, j \in S, m \ge 1. \end{aligned}$$

Hence, we obtain Theorem 3.3 by letting $L(i,j) = 2M[w(i) + w(j)] + 2$. □

Surely, if $\sum_j p_{ij}(a) w(j)$ is convergent uniformly in $a \in A(i)$ for each $i \in S$, then 4 of Condition 3.1 is true.

It is obvious that when $w(i)$ is bounded, $h_m(i)$ is also bounded. It is not easy to get all f_m in Condition 3.2. A stronger condition than Condition 3.2 is to require that Eq. (3.15) be true for all f with the same state $s_m = s_0$. That is, there is a state $s_0 \in S$ such that $E_{f,i}T_{s_0} \leq h(i)$ for all $i \in S$ and $f \in F$.

In the literature, many recurrent conditions are presented (see [139]). Some of them are given in the following.

Condition 3.3:

C1. *There is a finite state subset D and a positive constant K such that $E_{f,i}T_D \leq K$ for all $i \in S$ and $f \in F$. Moreover, for any $f \in F$, there is at most one closed state subset in the Markov chain with $P(f)$.*

C2. *There is a positive constant K such that for each $f \in F$, there exists a state $j(f)$ satisfying*

$$E_{f,i}T_{j(f)} \leq K, \quad i \in S.$$

C3. *(Simultaneous Doeblin) There is a finite state subset D, an integer $\nu \geq 1$, and a constant $\gamma \in (0,1)$ such that*

$$\sum_{j \in D} p_{ij}^{\nu}(f) \geq \gamma, \quad i \in S, \quad f \in F.$$

Moreover, for any $f \in F$, there is at most one closed state subset in the Markov chain with $P(f)$.

C4. *There is an integer $\nu \geq 1$ and a constant $\gamma \in (0,1)$ such that for each $f \in F$,*

$$\sum_{j} \min\{P_{i_1,j}^{\nu}(f), P_{i_2,j}^{\nu}(f)\} \geq \gamma, \quad i_1, i_2 \in S.$$

C5. *(Ergodicity) There is an integer $\nu \geq 1$ and a constant $\gamma \in (0,1)$ such that for each $f \in F$, there is a probability distribution $\{\eta_j(f), j \in S\}$ on S satisfying*

$$\sum_{j} |P_{ij}^{n}(f) - \eta_j(f)| \leq 2(1-\gamma)^{[n/\nu]}, \quad i \in S, \quad n \geq 1,$$

where $[x]$ is the largest integer that is less than or equal to x.

We have the following results for the conditions above.

Theorem 3.4: *We have the following conclusions.*

1. *Conditions C1, C2, and C3 are equivalent to each other. Moreover, if $P(f)$ is nonperiodic for each $f \in F$, then the five conditions, from C1 to C5, are all equivalent to each other. When one of the five*

conditions is true and the reward function is uniformly bounded, there is a bounded solution of ACOE.

2. *C2 is true if and only if there is a positive constant K such that for any bounded reward function $r(i,a)$, ACOE has a bounded solution $\{\rho, h\}$ satisfying $\|h\| \leq K\|r\|$, where $\|\cdot\|$ is the supremum norm.* □

The proof for 1 above can be seen in Federgruen et al. [42], where some recurrent conditions are given to ensure the existence of bounded solution of ACOE. The proof for 2 can be seen in Cavazos-Cadena [12]. Theorem 3.4 gives some equivalent conditions for the existence of a bounded solution of ACOE. In applications, stronger conditions may be more easily verified.

Thomas in [139] pointed out 23 conditions to ensure the existence of some bounded solutions of ACOE. Among these conditions, that presented by Ross [113] is a weaker one. Condition 3.1 and Condition 3.2 presented here weaken Ross' condition. In the following section, we further weaken Condition 3.1.

3. Optimality Inequalities

By checking carefully the proof of 1 in Theorem 3.1, we can find that if a triple $\{\rho, h(i), f\}$, where ρ is a constant, h is a function on the state space, and $f \in S$ is a stationary policy, satisfies the following inequality,

$$\rho + h(i) \leq \sup_a \{r(i,a) + \sum_j p_{ij}(a)h(j)\}, \qquad i \in S, \tag{3.16}$$

then the same procedure as that in the proof of Theorem 3.1 will lead to the conclusion that $\rho \leq V(f,i) + \varepsilon$ for all $i \in S$. Hence, f is an ε-optimal policy if we have $\rho \geq V^*(i)$ for $i \in S$. This implies that the inequality Eq. (3.16) can lead to the same conclusions as ACOE. On the other hand, sufficient conditions for the validity of Eq. (3.16) may be weaker than those for ACOE.

In this section, we study the inequality Eq. (3.16). We call it the average criterion optimality inequality (ACOI for short).

The concept of ACOI was first presented and studied by Sennott [121], where she presented a set of conditions for countable states and finite actions, to ensure a solution of the ACOI and an ε-optimal policy. Under her condition, the relative value function h satisfies the following condition,

$$-M(i) \leq h(i) \leq N, \qquad i \in S, \tag{3.17}$$

for some constant N and a nonnegative function $M(i)$. Optimal stationary policies can still be obtained from the ACOI. Based on [121], the set of conditions is weakened and generalized to the case of Borel state space and Borel action sets in, for example, [51], [111], and [119]. Later, Sennott [123] presented a new and weaker set of conditions combining the ideas in [121] and [58] (see the

last section of this chapter). Its main contribution is to weaken the condition under which the relative value function h satisfies

$$-M(i) \leq h(i) \leq L(i), \qquad i \in S$$

for some nonnegative functions $M(i)$ and $L(i)$.

On the other hand, Hernandez-Lerma and Lasserre in [51] presented another set of conditions under which the relative value function h satisfies

$$-\infty \leq h(i) \leq N, \qquad i \in S$$

for some constant N.

But there are two limitations on these conditions: (1) $r(i,a)$ is assumed to be nonpositive, and (2) the action sets are all finite, or are all Borel and some continuities about $r(i,a)$ and $p_{ij}(a)$ in a are assumed. In this section, we generalize the conditions presented in [123] in three aspects: (i) The condition about $h(i)$ is generalized into

$$-\infty \leq h(i) \leq L(i), \qquad i \in S$$

for some nonnegative function $L(i)$, which combines the former two conditions about $h(i)$. (ii) The reward function $r(i,a)$ is unbounded from above and from below in $a \in A(i)$ for each $i \in S$. (iii) The action sets $A(i)$ are arbitrary but nonempty and we need no continuity conditions on the reward function and the state transition probability.

3.1 Conditions

We present four conditions here.

The usual condition $r(i,a) \leq 0$, required in the literature, is generalized by the following condition.

Condition 3.4: *There exists a sequence of discount factors $\beta_m \uparrow 1$, a sequence of constants $\varepsilon(m) \downarrow 0$, and policies $\{f_m, m \geq 1\}$ such that*

1. For all $\pi \in \Pi$ and $i \in S$, both the discounted criterion $V_{\beta_m}(\pi, i)$ and the average criterion $V(\pi, i)$ are well defined.

2. We have $V_{\beta_m}(i) \leq r(i, f_m) + \beta_m \sum_j p_{ij}(f_m) V_{\beta_m}(j) + \varepsilon(m)$ for all $i \in S$ and $m \geq 1$.

For the condition above, statement 1 implies that $E_{\pi,i}\{r(X_n, \Delta_n)\}$ is well defined (may be infinity) for each $n \geq 0$ and moreover, both the series $\sum_{n=0}^{\infty} \beta^n E_{\pi,i} r(X_n, \Delta_n)$ and the summation $\sum_{n=0}^{N-1} E_{\pi,i} r(X_n, \Delta_n)$, for each $N \geq 1$, are also well defined, and so the statement is necessary for our study. Whereas statement 2 is true if the following discounted criterion optimality inequality

(DCOI for short) is true,

$$V_{\beta_m}(i) \le \sup_a \{r(i,a) + \beta_m \sum_j p_{ij}(a) V_{\beta_m}(j)\}, \quad i \in S, \quad m \ge 0.$$

In Chapter 2, we show that if we define the objective function by $V_\beta(\pi, i) := E_{\pi,i}\{\sum_{n=0}^{\infty} \beta^n r(X_n, \Delta_n)\}$, then Condition 3.4.1 implies the validity of the discounted criterion optimality equation, and also of Condition 3.4.2. It should be noted that f_m in Condition 3.4.2 is not required to be a $(1 - \beta_m^{-1})\varepsilon(m)$-optimal policy.

Condition 3.4 is true under the usual conditions presented in the literature, for example, those presented in Chapter 2, and is assumed to be true hereafter in this chapter. The second condition is given below.

Condition 3.5: *There exists a state $0 \in S$ such that the quantity $(1 - \beta_m)V_{\beta_m}(0)$ is bounded for $m \ge 0$.*

This condition appeared in Condition 3.1 in the last section. By Condition 3.5, $V_{\beta_m}(0)$ is finite and thus we can define the relative value function by

$$h_{\beta_m}(i) = V_{\beta_m}(i) - V_{\beta_m}(0), \qquad i \in S, m \ge 0.$$

Using $h_{\beta_m}(i)$, Condition 3.4.2 can be rewritten as

$$\begin{aligned} & (1 - \beta_m)V_{\beta_m}(0) + h_{\beta_m}(i) \\ \le \quad & r(i, f_m) + \beta_m \sum_j p_{ij}(f_m) h_{\beta_m}(j) + \varepsilon(m), \quad i \in S. \end{aligned} \tag{3.18}$$

Using the well-known diagonalization method, there exists a subsequence of $\{m\}$ (which is assumed to be $\{m\}$ itself) such that the following two limits exist and are finite,

$$\rho = \lim_{m\to\infty} (1 - \beta_m)V_{\beta_m}(0), \quad \lim_{m\to\infty} p_{ij}(f_m), \qquad i, j \in S.$$

About the third condition, we introduce first the following expression

$$\limsup_{m\to\infty} \sum_j p_{ij}(f_m)V(j) \le \sum_j \lim_{m\to\infty} p_{ij}(f_m)V(j) < +\infty, \quad i \in S, \tag{3.19}$$

and we say that the vector V satisfies Eq. (3.19). By the Fatou lemma, one knows that if V is nonnegative then V satisfies Eq. (3.19) if and only if

$$\lim_{m\to\infty} \sum_j p_{ij}(f_m)V(j) = \sum_j \lim_{m\to\infty} p_{ij}(f_m)V(j) < +\infty, \qquad i \in S.$$

That is, the order of $\lim_m$ and $\sum_j$ can be changed.

Condition 3.6: *The following 1 or 2 holds.*

1. *There exist nonnegative functions $L(i)$ and $M(i)$ such that $-M(i) \le h_{\beta_m}(i) \le L(i)$ for all i and m, and $M + L$ satisfies (3.19).*
2. *There exists a nonnegative function $L(i)$ such that $h_{\beta_m}(i) \le L(i)$ for all i, m. Moreover, L satisfies (3.19), and there exists a subsequence of $\{m\}$ (which is assumed to be $\{m\}$ itself) and $f \in F$ such that*

$$\lim_{m\to\infty} r(i, f_m) = r(i, f), \quad \lim_{m\to\infty} p_{ij}(f_m) = p_{ij}(f), \qquad i, j \in S. \quad (3.20)$$

Let

$$h(i) = \limsup_{m\to\infty} h_{\beta_m}(i), \qquad i \in S.$$

Obviously, under Condition 3.6, $-\infty \le h(i) \le L(i)$ for $i \in S$. Moreover, if Condition 3.6.1 is true, then $-M(i) \le h(i) \le L(i)$ and by the diagonalization method one can assume that

$$h(i) = \lim_{m\to\infty} h_{\beta_m}(i) \quad \text{for all } i \in S.$$

Remark 3.4:

1. *When (a) the action set $A(i)$ is compact and both $r(i, a)$ and $p_{ij}(a)$ satisfy some continuity conditions, or (b) the set $\{f_m(i), m \ge 1\}$ is finite (especially, $A(i)$ is finite) for each i, there exists a limit point f of the sequence $\{f_m\}$; that is, for each i, there exists N_i such that*

$$f_m(i) = f(i), \quad n \ge N_i. \quad (3.21)$$

This implies both 1 and 2 in Condition 3.6.

2. *In each of the following cases, L or $M + L$ satisfies (3.19): (a) The state space S is finite. (b) For each i, the set $\{j | p_{ij}(a) > 0$ for some $a \in A(i)\}$ is finite (e.g., in many queueing models and inventory models). (c) For each i, $A(i)$ is finite (which results in that $\{f_m\}$ has a limit point). (d) L or $M + L$ is bounded.*

The last condition is concerned with a stationary policy f.

Condition 3.7: $\lim_{n\to\infty} (1/n) E_{f,i} L(X_n) = 0$ *for all i.*

This condition for $L = h$ appeared in Theorem 3.1. Obviously, Condition 3.7 implies that for each i, $E_{f,i} L(X_n)$ is finite for sufficiently large n. Thus, we can assume that $E_{f,i} L(X_n)$ is also finite for all i and n. If $\lim_{n\to\infty} (1/n) E_{f,i} \sum_{t=0}^{n-1} r(X_t, f)$ exists for some i, then "lim" in Condition 3.7 can be weakened to "liminf".

Remark 3.5: *Conditions 3.5 and 3.7 are abstracted from [123]. 1 of Condition 3.6 is also from [123], but here it is required that $M + L$ satisfy (3.19), which is essential when $A(i)$ is not finite. Whereas 2 of Condition 3.6 weakens the condition (A4) in [51], where $L(i) = N$ is required.*

3.2 Properties of ACOI and Optimal Policies

First, we have the following lemma.

Lemma 3.5: *If Condition 3.5 and 3.6 hold, then $\rho \geq V^*(i)$ for $i \in S$.*

Proof: We first note that $h_{\beta_m}(i) \leq L(i)$ for all $i \in S$ is true in either 1 or 2 in Condition 3.6. Hence, for each $m \geq 1$,

$$\begin{aligned}(1-\beta_m)V_{\beta_m}(i) &= (1-\beta_m)h_{\beta_m}(i) + (1-\beta_m)V_{\beta_m}(0) \\ &\leq (1-\beta_m)L(i) + (1-\beta_m)V_{\beta_m}(0), \ \ i \in S.\end{aligned}$$

Letting $m \to \infty$ above implies that

$$\limsup_{m\to\infty}(1-\beta_m)V_{\beta_m}(i) \leq \rho, \ \ i \in S.$$

By (3.3), $\rho \geq V^*(i)$ for $i \in S$. □

The above lemma says that ρ, as the limit of $(1-\beta_m)V_{\beta_m}(0)$, is an upper bound of the optimal value function $V^*(i)$. The result of the above lemma appeared as a condition in Corollary 3.1.

In the following theorem, we present conditions under which a stationary policy will be ε-optimal for $\varepsilon \geq 0$.

Theorem 3.5: *Suppose that $\varepsilon \geq 0$ is a constant, f is a stationary policy, $L(i)$ is a nonnegative function, and $v(i)$ is an extended real-valued function. If $v(i) \leq L(i)$ for all $i \in S$, if f and L satisfy Condition 3.7, and if*

$$\rho + v(i) \leq r(i,f) + \sum_j p_{ij}(f)v(j) + \varepsilon, \qquad i \in S, \tag{3.22}$$

then $V(f,i) \geq \rho - \varepsilon$ for $i \in S_v$; that is, f is ε-optimal in S_v, where $S_v := \{i | v(i) > -\infty\}$ and is a closed state set under $P(f)$.

Proof: First, we assume that $S_v = S$; that is, $v > -\infty$. By Eq. (3.22), we know that

$$\rho + v(X_n) \leq r(X_n, f) + E_f(v(X_{n+1})|X_n) + \varepsilon, \qquad n \geq 0.$$

For any state $i \in S$, by taking expectation $E_{f,i}$ above, we get

$$\rho + E_{f,i}v(X_n) \leq E_{f,i}r(X_n,f) + E_{f,i}v(X_{n+1}) + \varepsilon, \ \ n \geq 0, i \in S. \tag{3.23}$$

From Condition 3.7 and

$$E_{f,i}v(X_n) = E_{f,i}L(X_n) + E_{f,i}\{v(X_n) - L(X_n)\}$$

we know that $E_{f,i}v(X_n)$ exists and $E_{f,i}v(X_n) \leq E_{f,i}L(X_n) < \infty$. Now, $V(f,i) \leq \rho$ for $i \in S$ by Lemma 3.5, which together with the definition of $V(f,i)$ implies that $E_{f,i}r(X_n, f) < +\infty$. Note that if $E_{f,i}v(X_n) > -\infty$ for some n, then $E_{f,i}r(X_n, f) > -\infty$ and $E_{f,i}v(X_{n+1}) > -\infty$ by Eq. (3.23). Then it can be proved by the induction method that $E_{f,i}r(X_n, f) > -\infty$ and $E_{f,i}v(X_{n+1}) > -\infty$ for all i and n. Thus both $E_{f,i}r(X_n, f)$ and $E_{f,i}v(X_n)$ are finite for all i and n.

Summing up the terms in Eq. (3.23) for $n = 0, \ldots, N-1$ and dividing it by N yields

$$\begin{aligned}\frac{1}{N}\sum_{n=0}^{N-1} E_{f,i}r(X_n, f) &\geq \rho + \frac{1}{N}v(i) - \frac{1}{N}E_{f,i}v(X_N) - \varepsilon \\ &\geq \rho + \frac{1}{N}v(i) - \frac{1}{N}E_{f,i}L(X_N) - \varepsilon, \quad i \in S.\end{aligned}$$

Taking the $\liminf_n$ of both sides above yields $V(f,i) \geq \rho - \varepsilon$ for $i \in S$, which implies by Lemma 3.5 that f is ε-optimal. When the limit of $(1/N)\sum_{n=0}^{N-1} E_{f,i}r(X_n, f)$ exists, then taking the $\limsup_n$ of both sides yields also $V(f,i) \geq \rho - \varepsilon$ for $i \in S$. So, "lim" in Condition 3.7 can be weakened by "lim inf".

Now, if $S_v \neq S$ then for any i with $v(i) > -\infty$, one has $\sum_{j \in S_v} p_{ij}(f) = 1$ by Eq. (3.22); that is, S_v is a closed set under $P(f)$. So, the problem can be restricted to S_v if the initial state $i \in S_v$, and the results follow as above. □

In Theorem 3.5, we investigate the inequality (3.22), not ACOI (3.16) as in [51], because the inequality is sufficient to get an (ε-)optimal stationary policy. Based on Theorem 3.5, we call (3.22) ACOI(ε), where ε is the given constant. Hence, we have two types of ACOI, given by (3.16) and (3.22), respectively. In some cases, ε may be zero, as 3 in the following lemma.

Next, we investigate the existence of solutions of Eq. (3.22) under Conditions 3.4–3.7.

Lemma 3.6: *Provided that Conditions 3.4 and 3.5 hold,*

1. *If 1 of Conditions 3.6 hold, then*

$$\rho + h(i) \leq \limsup_{m\to\infty}\{r(i, f_m) + \sum_j p_{ij}(f_m)h(j)\}, \qquad i \in S. \tag{3.24}$$

2. *If $h_{\beta_m}(i) \leq L(i)$ for all i and m with $L \geq 0$ satisfying (3.19), then*

$$\rho + h(i) \leq \limsup_{m\to\infty} r(i, f_m) + \sum_j \lim_{m\to\infty} p_{ij}(f_m)h(j), \qquad i \in S. \tag{3.25}$$

3. *If 2 of Conditions 3.6 hold, then*

$$\rho + h(i) \le r(i,f) + \sum_j p_{ij}(f)h(j), \qquad i \in S. \tag{3.26}$$

Proof: 1. If 1 of Conditions 3.6 hold, then by the given conditions and the Fatou lemma one can get

$$\begin{aligned}
&\limsup_{m\to\infty} \sum_j p_{ij}(f_m)\{h_{\beta_m}(j) - h(j)\} \\
&= \limsup_{m\to\infty} \sum_j p_{ij}(f_m)\{[h_{\beta_m}(j) - h(j)] - [L(j) + M(j)] + [L(j) + M(j)]\} \\
&\le \sum_j \lim_{m\to\infty} p_{ij}(f_m)\{[h(j) - h(j)] - [L(j) + M(j)] \\
&\quad + \limsup_{m\to\infty} \sum_j p_{ij}(f_m)[L(j) + M(j)]\} \le 0.
\end{aligned}$$

So $\limsup_{m\to\infty} \sum_j p_{ij}(f_m)h_{\beta_m}(j) \le \liminf_{m\to\infty} \sum_j p_{ij}(f_m)h(j)$ for $i \in S$, and the result follows by taking $\limsup_m$ in Eq. (3.18).

2. It follows the hypothesis and the Fatou lemma that

$$\begin{aligned}
&\limsup_{m\to\infty} \sum_j p_{ij}(f_m)h_{\beta_m}(j) \\
&= \limsup_{m\to\infty} \sum_j p_{ij}(f_m)\{h_{\beta_m}(j) - L(j) + L(j)\} \\
&\le \sum_j \lim_{m\to\infty} p_{ij}(f_m)\{h(j) - L(j)\} + \limsup_{m\to\infty} \sum_j p_{ij}(f_m)L(j) \\
&\le \sum_j \lim_{m\to\infty} p_{ij}(f_m)h(j).
\end{aligned}$$

Letting $\limsup_m$ in Eq. (3.18), one gets Eq. (3.25).

3. It follows from 2 of Conditions 3.6 and 2 above. □

The following theorem is obtained by summarizing the above results.

Theorem 3.6: *Provided that Conditions 3.4–3.6 hold,*

1. *ACOI (3.16) holds if its right-hand side is well defined, and then f is ε-optimal in S_h if f attains the ε-supremum of its right-hand side (for some $\varepsilon \ge 0$) and satisfies Condition 3.7.*
2. *If 1 of Conditions 3.6 hold, and there exists a policy $f \in F$ and a constant $\varepsilon \ge 0$ such that*

$$\begin{aligned}
&\limsup_{m\to\infty}\{r(i,f_m) + \sum_j p_{ij}(f_m)h(j)\} \\
\le\ & r(i,f) + \sum_j p_{ij}(f)h(j) + \varepsilon, \qquad i \in S,
\end{aligned}$$

and f satisfies Condition 3.7, then ACOI(ε) (3.22) is true and f is ε-optimal.

3. *If 2 of Conditions 3.6 hold and f satisfies Condition 3.7, then ACOI(0) (3.26) is true with $\rho = V^*(i)$ for $i \in S_h$, and f is optimal in S_h, where $S_h = \{i | h(i) > -\infty\}$.* □

For 1 in the above theorem, it is apparent that if $\sum_j p_{ij}(a)L(j)$ is well defined for all (i, a), then $\sum_j p_{ij}(a)h(j)$ is also well defined (but may be infinity) under the given conditions and so ACOI (3.16) holds from 3 of Theorem 3.6.

Remark 3.6: *Under the condition of 2 in Lemma 3.6, if there is a stationary policy f and a constant ε such that*

$$\limsup_{m\to\infty} r(i, f_m) + \sum_j \lim_{m\to\infty} p_{ij}(f_m)h(j) \le r(i, f) + \sum_j p_{ij}(f)h(j) + \varepsilon, \quad i \in S,$$

then, ACOI(ε) (3.22) is true.

From the above discussions we know that ACOI is sufficient for obtaining ε-optimal policies. ACOI is weaker than ACOE. Sennott [121] presented an example where ACOI is true and ACOE is not true. Moreover, if we have a stationary policy such that ACOI(ε) is true for some constant ε, then this policy is ε-optimal.

4. Notes and References

There are many papers dealing with the average criterion Markov decision processes, as one can see in the survey paper [1].

There are several methods to study the average criterion. The most popular one is the discount factor vanishing method, where results for the average criterion are obtained by letting the discount factor tend to one in the results for the discounted criterion. This method was first presented by Blackwell [5], where he showed the existence of an optimal policy from the relationship between the discounted criterion and the average criterion for finite MDPs. Taylor [137] derived the average criterion optimality equation (ACOE) from the discounted criterion optimality equation by studying the asymptotic properties of the relative value function $h_\beta(i) = V_\beta(i) - V_\beta(0)$ for a Markovian sequential replacement problem. Ross [113] generalized the method to general cases. Ross presented a set of conditions under which the ACOE has a bounded solution and then any stationary policy achieving the ACOE is optimal. His main condition is that the relative value function is uniformly bounded in β and i. Later, many papers generalized Ross's condition. A survey paper on this area is Thomas [139].

On the other hand, Fernandez-Gaucher and Marcus [41] showed that the boundedness of the relative value function is also necessary for the existence of bounded solutions of the ACOE. Surely, it is too strong to require the existence of bounded solutions of the ACOE.

Sennott [121] presented ACOI, where the inequality is required corresponding to the equality in the optimality equation. She presented a set of conditions to ensure ACOI. The advantage of ACOI to ACOE is that ACOI requires weaker conditions and at the same time is of the same properties as ACOE. After Sennott [121], many authors tried to generalize Sennott's conditions.

Along with Ross's and Sennott's methods and ideas, we further presented weaker conditions to ensure the validity and properties of ACOE and ACOI in this chapter. The contents of this chapter are mainly from Hu [58] and [68].

Problems

1. For the average criterion $V(\pi, i)$, suppose that $V(\pi, i)$ is well defined for each policy π and state i. Check that under this supposition whether or not the results given in Theorem 2.1 is true for the average criterion.

2. Discuss the problems given in Chapter 2 for the average criterion.

3. For the optimal stopping problem (see Problem 6 in Chapter 2), discuss when the average criterion optimality equation is true and when the average criterion optimality inequality is true.

Chapter 4

CONTINUOUS TIME MARKOV DECISION PROCESSES

This chapter discusses continuous time Markov decision processes, where the state space and the action sets are all countable. First, we focus on the total reward criterion for a stationary model by applying the ideas and methods presented in Chapter 2 for DTMDPs. Similar results to those in Chapter 2 are obtained. Then, we deal with a nonstationary model with the total reward criterion. By dividing the time axis into shorter intervals, we obtain the standard results, such as the optimality equation and the relationship between the optimality of a policy and the optimality equation. Finally, we study the average criterion for a stationary CTMDP model by transforming it into a DTMDP model. Thus, the results in DTMDPs can be used directly for CTMDPs for the average criterion.

1. A Stationary Model: Total Reward

1.1 Model and Conditions

The model of the continuous time Markov decision processes discussed here is

$$\{S, A(i), q_{ij}(a), r(i,a), U_\alpha\}. \tag{4.1}$$

Here, the state space S and the action set $A(i)$, available at state i, are countable. $\{q_{ij}(a) \,|\, i, j \in S, a \in A(i)\}$ is the state transition rate family satisfying $q_{ij}(a) \geq 0$ for $i \neq j$ and $\sum_j q_{ij}(a) = 0$ for $(i,a) \in \Gamma = \{(i,a) \mid i \in S, a \in A(i)\}$, and it is assumed that for each $i \in S$,

$$\lambda(i) := \sup\{-q_{ii}(a) \mid a \in A(i)\} < \infty.$$

The reward rate function $r(i,a)$ is extended real-valued. U_α is the criterion of expected discounted total reward with discount rate $\alpha \in (-\infty, +\infty)$, and is defined below. In general, $\alpha \in [0, \infty)$ is required in the literature but is not

necessary here. The discount rate α means that one unit reward at time t values $e^{-\alpha t}$ at time 0. This differs from the discount factor β for the discrete time MDPs. But they are related to $\beta = e^{-\alpha\tau}$ if τ is the length of one period in discrete time MDPs. $\alpha = 0$ means no discounting.

In this chapter, we suppose that the measure about the time variable t is the Lebesgue measure.

A Markov policy $\pi = (\pi_t, t \geq 0) \in \Pi_m$ means that if the system is in state i at time $t \geq 0$, then the action chosen is according to a probability distribution $\pi_t(\cdot \mid i)$ in $A(i)$. Here it is assumed that $\pi_t(a \mid i)$ is Lebesgue measurable for each i and $a \in A(i)$. A stochastic stationary policy $\pi_0 \in \Pi_s$ is a Markov policy $\pi = (\pi_t)$ satisfying $\pi_t = \pi_0$ for all $t \geq 0$. A stationary policy $f \in \Pi_s^d$ is a stochastic stationary policy π_0 such that $\pi_0(f(i) \mid i) = 1$ for some $f(i) \in A(i), i \in S$. The set of decision functions is denoted by $F = \times_i A(i)$. It is obvious that the policy set Π_s^d is equivalent to the set F. For a policy $\pi = (\pi_t)$ and $s \geq 0$, we define a policy $\pi^s = (\pi_t^*) \in \Pi_m$ by $\pi_t^* = \pi_{s+t}$ for $t \geq 0$. π^s is, in fact, the policy π but delayed with a time period of s.

For any policy $\pi = (\pi_t) \in \Pi_m$ and $t \geq 0$, we define a matrix $Q(\pi, t) = (q_{ij}(\pi, t))$ and a vector $r(\pi, t) = (r_i(\pi, t))$ by,

$$q_{ij}(\pi, t) = \sum_{a \in A(i)} q_{ij}(a)\pi_t(a \mid i), \;\; r_i(\pi, t) = \sum_{a \in A(i)} r(i, a)\pi_t(a \mid i).$$

Thus, $q_{ij}(\pi, t)$ and $r_i(\pi, t)$ are, respectively, the state transition rate family and the reward rate function at time t under policy π. It is apparent that $\lambda(i) < \infty$ is necessary to ensure the finiteness of $q_{ij}(\pi, t)$, whereas for $r_i(\pi, t)$, we first assume that it is well defined. Lemma 4.1 below discusses it in detail. If $\pi = \pi_0 \in \Pi_s$, then both $Q(\pi_0, t)$ and $r(\pi_0, t)$ are independent of t, and are denoted, respectively, by $Q(\pi_0) = (q_{ij}(\pi_0))$ and $r(\pi_0) = (r_i(\pi_0))$. The following condition is about the well definition of the process under each policy and is assumed throughout the chapter.

Condition 4.1: *For any Markov policy $\pi \in \Pi_m$, the $Q(\pi, t)$-process $\{P(\pi, s, t),\ 0 \leq s \leq t < \infty\}$ exists uniquely and is the minimal one. Moreover, for any $0 \leq s \leq t \leq u < \infty$,*

$$\begin{aligned}
&\frac{\partial}{\partial t} P(\pi, s, t) = P(\pi, s, t)Q(\pi, t), \\
&P(\pi, s, u) = P(\pi, s, t)P(\pi, t, u), \\
&\sum_j P_{ij}(\pi, s, t) = 1, \;\; P_{ij}(\pi, s, s) = \delta_{ij}, \quad i, j \in S.
\end{aligned}$$

One can find the constructing algorithm for the minimal Q-process in [23] (II. 17) for the stationary case and in [82] for the nonstationary case. Condition

4.1 is true when $q_{ij}(a)$ is bounded, or under the assumptions presented in [129] when $q_{ij}(a)$ is unbounded. In this section we deal mainly with the unbounded case, although the boundedness of q makes our discussions easier.

Now, we generalize the concept of policies. Let $X(t)$ be the state of the process at time t. Given any integer N, real numbers $\{t_i, i = 1, 2, \ldots, N\}$ with $0 = t_0 < t_1 < \cdots < t_N < t_{N+1} = \infty$, and Markov policies $\{\pi^{n,i}, n = 0, 1, 2, \ldots, N, i \in S\} \subset \Pi_m$, we define a policy

$$\pi = (\pi^{n,i}, n = 0, 1, 2, \ldots, N, i \in S)$$

as follows: for $n = 0, 1, 2, \ldots, N$, if $X(t_n) = i$, then $\pi^{n,i}$ is used in the time interval $[t_n, t_{n+1})$; that is, the action is chosen according to $\pi^{n,i}_{t-t_n}(\cdot \mid j)$ at time $t \in [t_n, t_{n+1})$ if $X(t) = j \in S$. Such a policy, denoted by $\pi = (\pi^{n,i})$ for short, is called a (finite) piecewise semi-Markov policy, the set of which is denoted by $\Pi_m(p)$. If all $\pi^{n,i} = f^{n,i} \in F$, then $\pi = (f^{n,i})$ is called a piecewise semi-stationary policy, the set of which is denoted by $\Pi^d_s(p)$. Especially, $\pi \in \Pi_m$ is a piecewise policy with $N = 0$ and $\pi^{0,i}$ is independent of i.

For such a piecewise semi-Markov policy π, if $X(t_n) = i$, then the system in $[t_n, t_{n+1})$ is a Markov process with the transition probability matrix $P(\pi^{n,i}, s, t)$. Thus, the system under a piecewise semi-Markov policy is a special case of piecewise Markov process (see [85]). In detail, for each s and t with $0 \le s \le t$ and $i, j \in S$, suppose that $s \in [t_m, t_{m+1})$ and $t \in [t_n, t_{n+1})$ for some $m \le n$. Then the state transition probability that the system will be in state j at time t provided that the system is in state i at time s and in state k at time t_m is

$$\begin{aligned} P^k_{ij}(\pi, s, t) &:= P_\pi\{X(t) = j \mid X(s) = i, X(t_m) = k\} \\ &= \sum_{j_1} P_{ij_1}(\pi^{m,k}, s - t_m, t_{m+1} - t_m) \\ &\quad \cdot \sum_{j_2} P_{j_1 j_2}(\pi^{m+1,j_1}, 0, t_{m+2} - t_{m+1}) \cdots \\ &\quad \cdot \sum_{j_{n-m}} P_{j_{n-m-1} j_{n-m}}(\pi^{n-1,j_{n-m-1}}, 0, t_n - t_{n-1}) \\ &\quad \cdot P_{j_{n-m} j}(\pi^{n,j_{n-m}}, t_n, t). \end{aligned} \tag{4.2}$$

For $i, j \in S$, let $P_{ij}(\pi, t) = P^i_{ij}(\pi, 0, t)$ be the state transition probability to reach state j at time t from state i at the initial time 0 under policy π.

Under a piecewise semi-Markov policy, the process is divided into several subprocesses. This makes it possible to obtain the optimality equation. This is important, for example, in the proof of Theorems 4.2, 4.3, and 4.5 below. But it will be proved that piecewise semi-Markov policies do not improve the optimality under weak conditions. So, the introduced piecewise semi-Markov policies are only for proving our results.

We define the criterion for a Markov policy $\pi \in \Pi_m$ by

$$U_\alpha(\pi) = \int_0^\infty e^{-\alpha t} P(\pi, t) r(\pi, t) dt, \tag{4.3}$$

where the integral is the Lebesgue integral. It is the expected discounted total reward on the whole time axis under policy π. Let $U_\alpha(\pi, t) := U_\alpha(\pi^t)$ for $t \geq 0$. Obviously,

$$U_\alpha(\pi, t) = \int_t^\infty e^{-\alpha(s-t)} P(\pi, t, s) r(\pi, s) ds \tag{4.4}$$

is the expected discounted, to time t, total reward on the time axis $[t, \infty)$ under π. Similarly, for a piecewise semi-Markov policy $\pi = (\pi^{n,i}) \in \Pi_m(p)$ with $\{t_n, n = 1, 2, \ldots, N\}$ and $t \geq 0$, we define inductively

$$\begin{aligned} U_\alpha^{n,k}(\pi, t, i) &= \int_t^{t_{n+1}} e^{-\alpha(s-t)} \sum_j P_{ij}(\pi^{n,k}, t, s) r_j(\pi^{n,k}, s) ds \\ &\quad + e^{-\alpha(t_{n+1}-t)} \sum_j P_{ij}(\pi^{n,k}, t, t_{n+1}) U_\alpha^{n+1,j}(\pi, t_{n+1}, j), \\ &\qquad t \in [t_n, t_{n+1}),\ n = 0, 1, \ldots, N-1,\ k, i \in S, \\ U_\alpha^{N,k}(\pi, t, i) &= U_\alpha(\pi^{N,k}, t - t_N, i),\ t \geq t_N,\ k, i \in S. \end{aligned} \tag{4.5}$$

Let $U_\alpha^{n,k}(\pi, t_n, i) = 0$ for $t = t_n$ and $k \neq i$, and $U_\alpha(\pi, i) = U_\alpha^{0,i}(\pi, 0, i)$. Let $U_\alpha(\pi)$ be a vector with its ith component $U_\alpha(\pi, i)$.

We have no requirement on the discount rate α. It may be positive (the discounted criterion), zero (the total reward), and negative. Thus, we call U_α uniformly the total reward.

Having defined the criterion, we now present the second condition.

Condition 4.2: *$U_\alpha(\pi)$ is well defined (may be infinite) for each policy $\pi \in \Pi_m(p)$.*

The meaning of Condition 4.2 has three aspects: (1) $\sum_j P_{ij}(\pi, t) r_j(\pi, t)$, and furthermore, the integral in Eq. (4.3), are well defined for each $\pi \in \Pi_m$. (2) $\sum_j P_{ij}(\pi, t, s)\, U_\alpha(\pi', s, j)$ is well defined for every policy $\pi \in \Pi_m$ and $\pi' \in \Pi_m(p)$. (3) The sum of the two terms in Eq. (4.5) is well defined; that is, the case of $\infty - \infty$ would not happen.

Conditions 4.1 and 4.2 require, respectively, the process and the criterion to be well defined for each policy π. We say that the CTMDP model Eq. (4.1) is well defined if both Conditions 4.1 and 4.2 are true. Surely, it is impossible to discuss the CTMDP model if either condition does not hold.

It is well known that Condition 4.2 is true whenever $\alpha > 0$ and $r(i, a)$ is bounded above or below, or $\alpha \geq 0$ and $r(i, a)$ is nonnegative or nonpositive.

For example, when $\alpha > 0$ and $r(i,a)$ has a upper bound M, then for any $\pi \in \Pi_m$,

$$\begin{aligned} & \int_0^\infty e^{-\alpha t}[P(\pi,t)r(\pi,t)]_i^+ dt \\ \leq & \int_0^\infty e^{-\alpha t} \sum_j P_{ij}(\pi,t) r_j^+(\pi,t) dt \leq \alpha^{-1} M. \end{aligned}$$

This together with the definition of the Lebesgue integral implies that $U_\alpha(\pi)$ is well defined and bounded above for $\pi \in \Pi_m$. Thus, $U_\alpha^{n,k}(\pi,t,i)$ defined in Eq. (4.5) is also well defined and bounded above. Condition 4.2 is assumed to be true hereafter.

Because a policy $\pi \in \Pi_m$ is also a piecewise semi-Markov policy with arbitrary N and $t_1, t_2, \ldots, t_N$, we have from Eq. (4.5) that for $\pi \in \Pi_m$ and $t \geq 0$,

$$U_\alpha(\pi) = \int_0^t e^{-\alpha s} P(\pi,s) r(\pi,s) ds + e^{-\alpha t} P(\pi,t) U_\alpha(\pi,t). \tag{4.6}$$

This means that $P(\pi,t)$ can be removed from the integral $\int_t^\infty$; that is,

$$\begin{aligned} & \int_t^\infty e^{-\alpha(s-t)} P(\pi,s) r(\pi,s) ds \\ = & \int_t^\infty e^{-\alpha(s-t)} P(\pi,t) P(\pi,t,s) r(\pi,s) ds \\ = & P(\pi,t) \int_t^\infty e^{-\alpha(s-t)} P(\pi,t,s) r(\pi,s) ds \\ = & P(\pi,t) U_\alpha(\pi,t). \end{aligned}$$

Equation (4.6) is still true for policy $\pi \in \Pi_m(p)$ by defining $r(\pi,s)$ adequately.

Let the optimal value function be

$$U_\alpha^*(i) = \sup\{U_\alpha(\pi,i) \mid \pi \in \Pi_m(p)\}, \quad i \in S.$$

For $\varepsilon \geq 0, \pi^* \in \Pi_m(p)$ and $i \in S$, if $U_\alpha(\pi^*,i) \geq U_\alpha^*(i) - \varepsilon$ (if $U_\alpha^*(i) < +\infty$) or $\geq 1/\varepsilon$ (if $U_\alpha^*(i) = +\infty$), then π^* is called ε-optimal at state i. Here, $1/0 = +\infty$ is assumed. If π^* is ε-optimal at all states $i \in S$ then π^* is called ε-optimal. An 0-optimal policy is simply called an optimal policy.

1.2 Model Decomposition

First, we introduce some concepts. State j can be reached from state i (and write $i \to j$) if there is a policy $\pi \in \Pi_m(p)$ and $t \geq 0$ such that $P_{ij}(\pi,t) > 0$. It is easy to see that $i \to j$ if and only if there are $\pi \in \Pi_m$ and $t \geq 0$ such that $P_{ij}(\pi,t) > 0$, or equivalently there are $n \geq 0$, states $j_1, j_2, \ldots, j_n \in S$ and

$f \in F$ such that $q_{ij_1}(f)q_{j_1j_2}(f)\cdots q_{j_nj}(f) > 0$. It is apparent that if $i \to j$ and $j \to k$, then $i \to k$. For a state subset $S_0 \subset S$ and a state i, if there is a state $j \in S_0$ such that $i \to j$, then we say that S_0 can be reached from state i, which is denoted by $i \to S_0$. Let $S_0^* = \{i \mid i \to S_0\}$ be a set of states that can reach S_0. Because $i \to i$, so $S_0 \subset S_0^*$. A state subset S_0 of S is called a closed set if $q_{ij}(a) = 0$ for all $i \in S_0, a \in A(i)$, and $j \notin S_0$, or equivalently, $(S-S_0)^* = S-S_0$. Similarly to above, S_0 is closed if and only if $P_{ij}(\pi, t) = 0$ for all $i \in S_0, \pi \in \Pi_m(p), j \notin S_0$, and $t \geq 0$.

For any closed subset S_0, if the system's initial state $i \in S_0$, then the system will remain in S_0 irrespective of the policies used. Thus, the restriction of the CTMDP model to S_0,

$$S_0\text{-CTMDPs} := \{S_0, A(i), p_{ij}(a), r(i,a), U_\alpha\}$$

is also a CTMDP model, which is called the sub-CTMDP model induced by S_0. Its policies are restrictions of policies for the original CTMDP model to S_0. It is clear that Conditions 4.1 and 4.2 are also true for the S_0-CTMDP model. Let $U_\alpha^{S_0}(\pi)$ be its total reward criterion. We have the following obvious theorem.

Theorem 4.1: *For any closed subset $S_0 \subset S$, $U_\alpha(\pi, i) = U_\alpha^{S_0}(\pi, i)$ for all $\pi \in \Pi_m(p)$ and $i \in S_0$.* □

The theorem says that the sub-CTMDP model induced by a closed set S_0 is equivalent to the original CTMDP model in the state subset S_0. So when both S_0 and $S-S_0$ are closed, the CTMDP model can be partitioned into two smaller parts: the S_0-CTMDP model and the $(S - S_0)$-CTMDP model. Moreover, if S_0 is closed and $U_\alpha^*(i)$ for $i \in S - S_0$ is known, or an (ε-)optimal policy can be obtained in $S - S_0$, then one needs to discuss only the S_0-CTMDP model. Thus, the state space is partitioned and reduced.

On the other hand, some actions may be eliminated without influencing the optimality of the model. The following definition corresponds to Definition 2.3 for DTMDPs.

Definition 4.1: *Suppose that $A_1(i) \subset A(i)$ for $i \in S$. We call the model with $A(i)$ being replaced by $A_1(i)$ a new CTMDP model (a symbol "$'$" is added). If for any policy π in the original CTMDP model there is a policy π' of the new CTMDPs model such that $U_\alpha(\pi, i) \leq U'_\alpha(\pi', i)$ for all i, then the CTMDP model is equivalent to the new CTMDP model and we say that $A(i)$ can be sized down to $A_1(i)$ for $i \in S$, or all actions in $A(i) - A_1(i)$ can be eliminated for $i \in S$.*

Surely, any policy π' for the new CTMDP model is also a policy for the original CTMDP model and $U'_\alpha(\pi, i) = U_\alpha(\pi, i)$ for all i. So when $A(i)$ can be sized down to $A_1(i)$, the optimal value function of the original CTMDP model

obviously equals that of the new CTMDP model. Thus, these two CTMDP models are equivalent.

For $i \in S$, we denote by

$$U(i) = \sup\{r(i,a) \mid a \in A(i)\},\ L(i) = \inf\{r(i,a) \mid a \in A(i)\}$$

the supremum and the infimum of the reward rate function $r(i,a)$ over the action set $A(i)$, respectively. Let

$$\begin{aligned} S_U &= \{i \mid U(i) = +\infty\}, \\ S_{=\infty} &= \{i \mid \text{there is } \pi \in \Pi_m(p) \text{ such that } U_\alpha(\pi, i) = +\infty\}, \\ S_\infty &= \{i \mid U^*_\alpha(i) = +\infty\} - S_{=\infty}, \\ S_{-\infty} &= \{i \mid U^*_\alpha(i) = -\infty\}, \\ S_0 &= S - S_{=\infty} - S_\infty - S_{-\infty} = \{i \mid -\infty < U^*_\alpha(i) < \infty\}. \end{aligned}$$

These state subsets have their obvious meanings.

The following lemma discusses the infinity of $L(i)$ and $U(i)$.

Lemma 4.1:

1. *For $i \in S_U$, there is a stochastic stationary policy $\pi_0 \in \Pi_s$ such that $r_i(\pi_0) = +\infty$ and so $U_\alpha(\pi_0, i) = +\infty$. Hence, $S_U \subset S_{=\infty}$.*
2. *For $i \in S$ with $L(i) = -\infty$, there is a stochastic stationary policy $\pi_0 \in \Pi_s$ such that $r_i(\pi_0) = -\infty$ and so $U_\alpha(\pi_0, i) = -\infty$.*
3. *For $i \in S, L(i) = -\infty$ and $U(i) = +\infty$ cannot be true simultaneously.*

Proof: 1. For $i \in S_U$, if there is an action $a \in A(i)$ such that $r(i,a) = +\infty$, then we define $\pi_0(a \mid i) = 1$. Otherwise, there are actions $a_n \in A(i)$ for $n \geq 1$, which are different from each other, such that $r(i, a_n) \geq n$. Fixing a constant $\delta \in (0,1)$, let $c = \sum_{n=1}^{\infty} n^{-(1+\delta)} < \infty$ and define $\pi_0(a_n \mid i) = (cn^{1+\delta})^{-1}$ for $n \geq 1$. For $i \notin S_U, \pi_0(\cdot \mid i)$ can be defined arbitrarily. Then, it is easy to prove that for any $i \in S_U, r_i(\pi_0) = +\infty$, and so $U_\alpha(\pi_0, i) = +\infty$ due to $-q_{ii}(\pi_0) < +\infty$.

2. This can be proved similarly to 1.

3. If $L(i) = -\infty$ and $U(i) = +\infty$ for some $i \in S$, let $\pi_0^{(1)}$ and $\pi_0^{(2)}$ be, respectively, the policies in 1 and 2. Then for policy π_0 defined by $\pi_0(\cdot \mid i) := 0.5\pi_0^{(1)}(\cdot \mid i) + 0.5\pi_0^{(2)}(\cdot \mid i), r_i(\pi_0)$ is undefined and so $U_\alpha(\pi_0, i)$ is also undefined. This contradicts Condition 4.2. □

The following lemma deals with eliminating the worst actions and partitioning the state space.

Lemma 4.2:

1. *$S^*_{=\infty} = S_{=\infty}$ and so $S' := S - S_{=\infty}$ is closed.*

2. *For $i \in S' - S_{-\infty}$, $A(i)$ can be sized down to*

$$A_1(i) = \{a \in A(i) \mid r(i,a) > -\infty \text{ and } \sum_{j \in S_{-\infty}} q_{ij}(a) = 0\}. \quad (4.7)$$

After the reduction, $S^*_{-\infty} = S_{-\infty}$ and so $S'' := S' - S_{-\infty}$ becomes closed.

3. *For $i \in S''$, $A_1(i)$ can further be sized down to*

$$\begin{aligned} A_2(i) \;=\; & \{a \in A_1(i) \mid \text{there is } \pi \in \Pi_m \text{ with } U_\alpha(\pi, i) > -\infty \\ & \text{such that the Lebesgue measure of } \{s \in [0,t] \mid \\ & \pi_s(a \mid i) > 0\} \text{is positive for each } t > 0\}. \end{aligned} \quad (4.8)$$

After this reduction, $S^*_\infty = S_\infty$ and so $S_0 := S'' - S_\infty$ is closed.

Proof: 1. For any state $i \in S^*_{=\infty}$, there is a state $j_0 \in S_{=\infty}$, a policy $\pi^* \in \Pi_m$, and $t^* \geq 0$ such that $P_{ij_0}(\pi^*, t^*) > 0$ by the definition. Taking any policy $\pi' \in \Pi_m(p)$ with $U_\alpha(\pi', j_0) = +\infty$, we define a policy π by using π^* in $[0, t^*)$ and π' in $[t^*, \infty)$. Then it is easy to see that $U_\alpha(\pi, i) = +\infty$ and so $i \in S_{=\infty}$. Thus, $S^*_{=\infty} = S_{=\infty}$ and $S' := S - S_{=\infty}$ is closed.

2. For any policy $\pi \in \Pi_m$, state $i \in S' - S_{-\infty}$ and action $a \in A(i)$, it can be assumed from Eq. (4.5) that there is $t^* > 0$ with $\pi_t(a \mid i) > 0$ for $t \leq t^*$.

If $r(i,a) = -\infty$, then $r_i(\pi, t) = -\infty$ for $t \leq t^*$, which implies that $U_\alpha(\pi, i) = -\infty$ due to Eq. (4.5).

If there is a state $j_0 \in S_{-\infty}$ with $q_{ij_0}(a) > 0$, then $P_{ij_0}(\pi, t^*) > 0$ by the construction of the minimal Q-process. Then due to $U^*_\alpha(j_0) = -\infty$ and Eq. (4.5), we have also $U_\alpha(\pi, i) = -\infty$.

So, $A(i)$ can be sized down to $A_1(i)$ for $i \in S' - S_{-\infty}$. It is apparent that $S^*_{-\infty} = S_{-\infty}$ after this reduction.

3. First, it should be noted that Eq. (4.8) is also true for $\pi \in \Pi_m(p)$. Thus, it is apparent that $A_1(i)$ can be reduced as $A_2(i)$. After this reduction, for any state $i \in S^*_\infty$, if there is a state $j_0 \in S_\infty$ and an action $a \in A_2(i)$ such that $q_{ij_0}(a) > 0$, then there is a policy π with $U_\alpha(\pi, i) > -\infty$ from the definition of $A_2(i)$, and there is $t^* > 0$ such that $P_{ij_0}(\pi, t^*) > 0$ by the construction of the minimal Q-process. Thus we can get from Eq. (4.5) that

$$\begin{aligned} c : \;=\; & \int_0^{t^*} e^{-\alpha s} \sum_j P_{ij}(\pi, s) r_j(\pi, s) ds \\ & + e^{-\alpha t^*} \sum_{j \neq j_0} P_{ij}(\pi, t^*) U_\alpha(\pi, t^*, j) > -\infty. \end{aligned}$$

Now, for any constant $M > 0$, we take any policy $\pi^M \in \Pi_m(p)$ with $U_\alpha(\pi^M, j_0) > M$ and define a policy $\pi^{*M} = (\pi^{0,j}, \pi^{1,j}, j \in S)$ by $\pi^{0,j} = \pi, \pi^{1,j} = \pi^{t^*}$ for

$j \neq j_0, \pi^{1,j_0} = \pi^M$, and $t_1 = t^*$. Then from Eq. (4.5) we have

$$U_\alpha(\pi^{*M}, i) \geq c + e^{-\alpha t^*} P_{ij_0}(\pi, t^*)M.$$

By letting $M \to \infty$, we get that $U^*_\alpha(i) = +\infty$; that is, $i \in S_\infty$. So $S^*_\infty = S_\infty$. This completes the proof. □

From Theorem 4.1 and Lemma 4.2, we have the following theorem, which is one of the main results in this chapter.

Theorem 4.2: *The state space S can be partitioned into four subsets: $S_{-\infty}$, $S_{=\infty}$, S_∞, and S_0, for which*

1. *In $S_{-\infty}$, $U^*_\alpha(i) = -\infty$ and each policy is optimal.*
2. *In $S_{=\infty}$, $U^*_\alpha(i) = \infty$ and there is an optimal policy (in fact, there is an optimal stochastic stationary policy in S_U).*
3. *In S_∞, $U^*_\alpha(i) = \infty$, and $U_\alpha(\pi, i) < \infty$ for each π. Thus there is no optimal policy.*
4. *In $S_0, U^*_\alpha(i)$ is finite and S_0 is closed after eliminating some worst actions. Moreover, the original CTMDP model in S_0 is equivalent to the following CTMDP model,*

$$S_0\text{-CTMDPs} = \{S_0, A_2(i), q_{ij}(a), r(i,a), U_\alpha\}. \tag{4.9}$$

Because $i \in S_{-\infty}$ when $A_2(i)$ is empty, the S_0-CTMDP model is well defined. Furthermore, it has the following properties,

$$-\infty < U^*_\alpha(i) < +\infty, \ -\infty < r(i,a) \leq U(i) < +\infty, \ \forall (i,a). \tag{4.10}$$

It is easy to see that all the above results restricted to $\Pi_s(p)$ are also true.

Based on the above theorem, we mainly discuss the S_0-CTMDP model in the following two subsections. So, we write S_0 and $A_2(i)$ by S and $A(i)$, respectively, for convenience, unless a special statement is given.

1.3 Some Properties

In this subsection, we discuss some properties of the S_0-CTMDP model given in Eq. (4.9) and simplify the expression of $A_2(i)$. First, we give several lemmas on Markov chains. The following lemma is from [23] (II. Sections 15–17).

Lemma 4.3: *Suppose that $P(t) = (p_{ij}(t))$ is a homogeneous state transition probability matrix on a countable state space S with a finite transition rate family $Q = (q_{ij})$. Let $q_i = -q_{ii}$. Then there are nonnegative continuous functions $g_{ij}(t)$ for $i, j \in S$, on $[0, \infty)$, such that*

$$p_{ij}(t) = e^{-q_i t} \int_0^t e^{q_i s} q_i g_{ij}(s) ds + e^{-q_i t} \delta_{ij}, \quad i, j \in S, t \geq 0,$$

where δ_{ij} denotes the Kronecker delta function and for $s > 0, t \geq 0$,

$$\begin{aligned} \lim_{s \to 0^+} g_{ij}(s) &= (1 - \delta_{ij}) q_{ij}/q_i, \\ \sum_j g_{ij}(s) &= 1, \\ g_{ij}(s+t) &= \sum_k g_{ik}(s) p_{kj}(t). \end{aligned}$$

The lemma above characterizes the state transition probability $p_{ij}(t)$ by the nonnegative function $g_{ij}(t)$. Based on it, we now show the following lemma, which characterizes the relationship between the finiteness of a function $\sum_j p_{ij}(t) u_j$ and the finiteness of a constant $\sum_j q_{ij} u_j$.

Lemma 4.4: *Suppose that $P(t) = (p_{ij}(t)), Q$ and $g_{ij}(t)$ are as in Lemma 4.3, $\sup_i q_i < \infty, u$ is a finite nonnegative function in $S, Z \subset S$, and $i \in S$. If $\sum_{j \in Z} p_{ij}(t^*) u_j$ is finite for some $t^* > 0$ then*

$$h_i(t) := q_i e^{q_i t} \sum_{j \in Z} g_{ij}(t) u_j$$

is finite and continuous in $[0, t^*)$ and $\sum_{j \in Z} q_{ij} u_j < \infty$. Otherwise, $h_i(t) = +\infty$ for all $t > 0$.

Proof: From Lemma 4.3,

$$\sum_{j \in Z} p_{ij}(t^*) u_j = e^{-q_i t^*} \int_0^{t^*} h_i(s) ds + e^{-q_i t^*} u_i \chi_Z(i),$$

where χ_Z is the indicator function of the set Z. Again, from Lemma 4.3, $g_{ij}(s+t) \geq g_{ij}(s) p_{jj}(t)$ for $s > 0, t \geq 0$ and $j \in S$, which is also true for $s = 0$ by the continuity of $g_{ij}(t)$. For the boundedness of q_i and [23] (Eq. (6), pp. 130), we know that for any $\varepsilon \in (0,1)$, there is a constant $\delta' > 0$ such that $p_{jj}(t) > 1 - \varepsilon$ for each $j \in S$ and $t < \delta'$. So,

$$g_{ij}(s+t) \geq (1 - \varepsilon) g_{ij}(s), \quad s \geq 0, t < \delta', j \in S \tag{4.11}$$

and therefore

$$h_i(s+t) \geq (1 - \varepsilon) h_i(s), \quad s \geq 0, t < \delta'. \tag{4.12}$$

We prove the lemma by the following two cases.

1. Suppose that $\sum_{j \in Z} p_{ij}(t^*) u_j$ is finite. Then, $\int_0^{t^*} h_i(s) ds$ is finite and so $h_i(t)$ is finite for almost everywhere (a.e. for short) $t \in [0, t^*]$. This implies that there is a constant $\delta < \delta'$ such that $h_i(\delta)$ is finite. From Eq. (4.11),

$$g_{ij}(t) \leq \frac{1}{1 - \varepsilon} g_{ij}(t + (\delta - t)) = \frac{1}{1 - \varepsilon} g_{ij}(\delta), \quad j \in S, \ t < \delta.$$

So $\sum_{j\in Z} g_{ij}(t)u_j$, as a series, is uniformly convergent in $[0,\delta]$ by the finiteness of $h_i(\delta)$. It can be proved similarly that this series is uniformly convergent in any subinterval of $[0,t^*]$ with its length being less than or equal to δ'. This results in the series being uniformly convergent in $[0,t^*)$. Because $g_{ij}(t)$ is continuous, $h_i(t)$ is also continuous in $[0,t^*)$. Now, $\sum_{j\in Z} q_{ij}u_j < \infty$ follows $h_i(0) < \infty$.

2. Suppose that $\int_0^t h_i(t)dt = \infty$ for each $t > 0$. Then, there is a decreasing sequence $t_n \to 0$ such that $h_i(t_n) \to +\infty$ as $n \to +\infty$. Fixing any $0 < t < \delta'$, one has that $t_n < t$ for sufficiently large n, and thus from Eq. (4.12),

$$h_i(t) = h_i(t_n + (t - t_n)) \geq (1-\varepsilon)h_i(t_n),\ \ 0 < t_n < t.$$

Letting $n \to \infty$ implies that $h_i(t) = \infty$, which is still true for all $t > 0$ due to Eq. (4.12). □

The following lemma discusses the derivatable of $\sum_j p_{ij}(t)u_j$.

Lemma 4.5: *Using the notations in Lemma 4.3, suppose that* $\sup_i q_i < \infty, u$ *is a finite function in* $S, t^* > 0$ *and* $i \in S$. *If* $\sum_j p_{ij}(t)u_j$ *is finite in* $[0,t^*]$, *then its derivative is well defined and continuous in* $[0,t^*)$ *and*

$$\frac{d}{dt}\{\sum_j p_{ij}(t)u_j\} = \sum_j \frac{d}{dt}p_{ij}(t)u_j = \sum_j\{-q_i p_{ij}(t) + q_i g_{ij}(t)\}u_j.$$

Proof: Applying Lemma 4.3 to both the positive, and the negative parts of $\sum_j p_{ij}(t)u_j$ results in

$$\sum_j p_{ij}(t)u_j = e^{-q_i t}\{\int_0^t e^{q_i s}q_i \sum_j g_{ij}(s)u_j ds + u_i\},\ \ t \in [0,t^*].$$

But it follows from Lemma 4.4 that the integrand above is continuous. So $\sum_j p_{ij}(t)u_j$ is differentiable and its derivative

$$\frac{d}{dt}(\sum_j p_{ij}(t)u_j) = -q_i\sum_j p_{ij}(t)u_j + q_i\sum_j g_{ij}(t)u_j$$

is continuous in $[0,t^*)$. On the other hand, by applying Lemma 4.3 again, we get that

$$\sum_j \frac{d}{dt}p_{ij}(t)u_j = \sum_j\{-q_i p_{ij}(t) + q_i g_{ij}(t)\}u_j,\ \ t \in [0,t^*].$$

This completes the lemma. □

Having the above three lemmas for preparation, we can now prove the following theorem.

Theorem 4.3:

1. *$P(\pi,t)U_\alpha^* < \infty$ is well defined for each $\pi \in \Pi_m(p)$ and $t > 0$.*

2. *For $\pi \in \Pi_m(p), t > 0$ and $i \in S$, if $\sum_j P_{ij}(\pi,t)U_\alpha^*(j) = -\infty$, then $U_\alpha(\pi^*, i) = -\infty$ for any piecewise semi-Markov policy $\pi^* = (\pi^{n,j}) \in \Pi_m(p)$ with $\pi^{0,i} = \pi$ and $t_1 = t$, especially, $U_\alpha(\pi, i) = -\infty$.*

Proof: 1. For $\pi \in \Pi_m, t > 0$ and any $\pi^* = (\pi^{0,i}, \pi^{1,i}, i \in S)$ with $\pi^{0,i} = \pi$ and $t_1 = t$,

$$\begin{aligned} U_\alpha(\pi^*, i) &= \int_0^t e^{-\alpha s} \sum_j P_{ij}(\pi, s) r_j(\pi, s) ds \\ &\quad + e^{-\alpha t} \sum_j P_{ij}(\pi, t) U_\alpha(\pi^{1,j}, j) \\ &\leq U_\alpha^*(i) < \infty \end{aligned}$$

is well defined for $i \in S$. This implies that the term $\sum_j P_{ij}(\pi,t)\ U_\alpha(\pi^{1,j}, j)$ above is also well defined and less than infinity. Now for any $\varepsilon > 0$ and $j \in S$, taking $\pi^{1,j}$ with $U_\alpha(\pi^{1,j}, j) \geq U_\alpha^*(j) - \varepsilon$, we have for any subset Z of S,

$$\begin{aligned} \sum_{j \in Z} P_{ij}(\pi,t) U_\alpha(\pi^{1,j}, j) &\leq \sum_{j \in Z} P_{ij}(\pi,t) U_\alpha^*(j) \\ &\leq \sum_{j \in Z} P_{ij}(\pi,t) U_\alpha(\pi^{1,j}, j) + \varepsilon, \quad i \in S. \end{aligned}$$

Hence, $\sum_j P_{ij}(\pi,t)U_\alpha^*(j) < \infty$ is well defined for $i \in S$ from the convergent definition of series. It can similarly be proved for $\pi \in \Pi_m(p)$.

2. For $\pi \in \Pi_m(p), t > 0$ and $i \in S$, if $\sum_j P_{ij}(\pi,t)U_\alpha^*(j) = -\infty$, then $\sum_j P_{ij}(\pi,t)\ U_\alpha(\pi^j, j) = -\infty$ for any $(\pi^j, j \in S) \subset \Pi_m(p)$. So $U_\alpha(\pi^*, i) = -\infty$ for the given policy π^*. □

We let $S^+ := \{i \in S \mid U_\alpha^*(i) \geq 0\}$ and $S^- := \{i \in S \mid U_\alpha^*(i) < 0\}$ be the state subsets with nonnegative and negative optimal values, respectively. Based on the theorem above, we have the following corollaries.

Corollary 4.1: *Suppose that $\sup_{i \in S} \inf_{a \in A(i)}(-q_{ii}(a))$ is finite. Then for any $i \in S$ and $a \in A(i)$, $\sum_j q_{ij}(a)U_\alpha^*(j) < \infty$ is well defined.*

Proof: The condition given in the corollary implies that there is a decision function $f^* \in F$ such that $Q(f^*)$ is bounded. Then, for any given $i \in S$ and $a \in A(i)$, let f satisfy $f(i) = a$ and $f(j) = f^*(j)$ for $j \neq i$. Then $Q(f)$ is bounded and $\sum_{j \in S^+} P_{ij}(f,t)U_\alpha^*(j) < \infty$ due to Theorem 4.3. Hence, the corollary follows Lemma 4.4. □

The condition supposed in the corollary above is certainly true when $q_{ij}(a)$ is uniformly bounded or the state space is finite.

The result of the above corollary may be used when we concern the well definition of the optimality equation below. The following corollary is important when we show the optimality equation in the next subsection.

Corollary 4.2: *Suppose that $f \in F$ with bounded $Q(f)$, $[P(f,t)U_\alpha^*]_i$ is finite in $t \in [0,t^*]$ for some given $i \in S$ and $t^* > 0$. Then $[P(f,t)U_\alpha^*]_i$ is continuously differentiable in $[0,t^*)$ and*

$$\begin{aligned} \frac{d}{dt}[\sum_j P_{ij}(f,t)U_\alpha^*(j)] &= \sum_j \frac{d}{dt}P_{ij}(f,t)U_\alpha^*(j), \\ \sum_j [P(f,t)Q(f)]_{ij}U_\alpha^*(j) &= \sum_j P_{ij}(f,t)[Q(f)U_\alpha^*]_j,\ t \in [0,t^*). \end{aligned} \quad (4.13)$$

Proof: The results except for Eq. (4.13) follow Lemma 4.5 and Corollary 4.1. Eq. (4.13) can be obtained from the arguments below. It can be seen that

$$\begin{aligned} & \sum_{k\in S} P_{ik}(f,t)\{\sum_{j\in S^+} q_{kj}(f)U_\alpha^*(j)\} \\ = & \sum_{k\in S^+} P_{ik}(f,t)\{\sum_{j\in S^+} q_{kj}(f)U_\alpha^*(j)\} + \sum_{k\in S^-} P_{ik}(f,t)\sum_{j\in S^+} q_{kj}(f)U_\alpha^*(j) \\ = & \sum_{j\in S^+}\{\sum_{k\in S^+} P_{ik}(f,t)q_{kj}(f) + \sum_{k\in S^-} P_{ik}(f,t)q_{kj}(f)\}U_\alpha^*(j) \\ = & \sum_{j\in S^+}\{\sum_k P_{ik}(f,t)q_{kj}(f)\}U_\alpha^*(j) \\ = & \sum_{j\in S^+}\frac{d}{dt}P_{ij}(f,t)U_\alpha^*(j) \end{aligned}$$

is finite. Similarly,

$$\sum_{k\in S} P_{ik}(f,t)\{\sum_{j\in S^-} q_{kj}(f)U_\alpha^*(j)\} = \sum_{j\in S^-}\{\sum_k P_{ik}(f,t)q_{kj}(f)\}U_\alpha^*(j)$$

is also finite. By subtracting the latter from the former, we get that

$$\sum_k P_{ik}(f,t)\{\sum_j q_{kj}(f)U_\alpha^*(j)\} = \sum_j\{\sum_k P_{ik}(f,t)q_{kj}(f)\}U_\alpha^*(j)$$

is well defined and is finite. So Eq. (4.13) is true. □

We conjecture that the result in Corollary 4.2 is also true for $\pi \in \Pi_m$, but it needs Lemma 4.3 to hold for a nonhomogeneous Markov process, which is not known to us.

To conclude this subsection, we simplify the expression for $A_2(i)$ defined in Eq. (4.8), and the notations $A_1(i), A_2(i)$, and so on, are the same as those in the previous subsections.

Theorem 4.4: *If $q_{ij}(a)$ is uniformly bounded, then*

$$A_2(i) \subset \{a \in A_1(i) \mid \sum_j q_{ij}(a)U_\alpha^*(j) > -\infty\}, \ \ i \in S. \tag{4.14}$$

Moreover, if $h_i(t)$ is finite and continuous whenever $\sum_{j\in Z} q_{ij}u_j$ is finite in Lemma 4.4, then

$$A_2(i) = \{a \in A_1(i) \mid \sum_j q_{ij}(a)U_\alpha^*(j) > -\infty\}, \ \ i \in S. \tag{4.15}$$

Proof: First, we show that $Q(\pi,t)U_\alpha^* < +\infty$ is well defined. If there is a policy π and a state i such that $\sum_{j\in S^+} q_{ij}(\pi,t)U_\alpha^*(j) = +\infty$ in a set with positive measure, for example, in $[0,t^*]$ for some $t^* > 0$, then due to the construction of the minimal Q-process one can get that $\sum_{j\in S^+} P_{ij}(\pi,t)U_\alpha^*(j) \geq +\infty$ for $t > 0$, which contradicts 1 of Theorem 4.3. So, $Q(\pi,t)U_\alpha^* < +\infty$ is well defined.

Now, for any $i \in S$ and $a \in A_1(i)$, if $\sum_j q_{ij}(a)U_\alpha^*(j) = -\infty$, then $\sum_{j\in S^+} q_{ij}(a)\ U_\alpha^*(j) < \infty$ and $\sum_{j\in S^-} q_{ij}(a)U_\alpha^*(j) = -\infty$. Thus for any policy π satisfying the condition given in Eq. (4.8), $\sum_j q_{ij}(\pi,t)U_\alpha^*(j) = -\infty$ for each $t > 0$ with the Lebesgue measure of $\{s \in [0,t] \mid \pi_s(a|i) > 0\}$ being positive. Then, $\sum_j P_{ij}(\pi,t)U_\alpha^*(j) = -\infty$, which together with Theorem 4.3 implies that $U_\alpha(\pi,i) = -\infty$. So, Eq. (4.14) is true.

Now, suppose that $h_i(t)$ is finite and continuous whenever $\sum_{j\in Z} q_{ij}u_j$ is finite in Lemma 4.4. For any $i \in S$ and $a \in A_1(i)$ with $\sum_j q_{ij}(a)U_\alpha^*(j) > -\infty$, we say that there is $f^* \in F$ with

$$\sum_j q_{ij}(a)r(j,f^*) > -\infty. \tag{4.16}$$

Otherwise, $\sum_j q_{ij}(a)r(j,f) = -\infty$ for each f, which implies that for any π and t we have that $\sum_j q_{ij}(a)r_j(\pi,t) = -\infty$. Again from the construction of the minimal Q-process, $\sum_j P_{ij}(\pi,t)r_j(\pi,t) = -\infty$ for $t > 0$. So, $U_\alpha(\pi,i) = -\infty$ for each π due to Eq. (4.5), and so $A_2(i)$ is empty. This is a contradiction. Thus, there is f^* satisfying Eq. (4.16). Now, let f satisfy $f(i) = a$ and $f(j) = f^*(j)$ for $j \neq i$. Then, both $\sum_j q_{ij}(f)r(j,f)$ and $\sum_j q_{ij}(f)U_\alpha^*(j)$ are finite, which implies that both $\sum_j P_{ij}(f,t)U_\alpha^*(j)$ and $\sum_j P_{ij}(f,t)r(j,f)$ are finite and thus continuous from Lemma 4.4. Hence, $U_\alpha(f,i)$ exists and is finite due to Eq. (4.5) and $a \in A_2(i)$. □

Remark 4.1: *(1) From the above theorem, if S^- is finite, or $U_\alpha^*(i)$ is bounded below in i, then Eq. (4.15) is true when $q_{ij}(a)$ is uniformly*

bounded. (2) S^- is empty if U_α^* is nonnegative, especially if the reward function is nonnegative. (3) $U_\alpha^*(i)$ is bounded below in i if $\alpha > 0$ and the reward function is bounded below.

From result 3 of Lemma 4.3 we know that with the action set $A_2(i)$, S_0 is closed. This together with Eq. (4.14) lets us conclude that when $q_{ij}(a)$ is uniformly bounded, $A_1(i)$ can be reduced as

$$\begin{aligned} A_2^*(i) &= \{a \in A_1(i) | \sum_j q_{ij}(a) U_\alpha^*(j) > -\infty \\ &\quad \text{and } q_{ij}(a) = 0 \text{ for all } j \notin S_0\}, i \in S_0. \end{aligned} \tag{4.17}$$

The results obtained in this subsection are the preparation for showing the validity of the optimality equation in the next subsection.

1.4 Optimality Equation and Optimal Policies

This subsection discusses the optimality equation and the optimality of policies achieving the optimality equation for S_0-CTMDPs, under the assumption that $\{q_{ij}(a)\}$ is uniformly bounded; that is,

$$\lambda = \sup\{-q_{ii}(a) \mid i \in S, a \in A(i)\} < \infty.$$

For $\pi \in \Pi_m(p), t \geq 0$ and a finite function $u = (u(i))$ on the state space S, we define

$$U_\alpha(\pi, t, u) = \int_0^t e^{-\alpha s} P(\pi, s) r(\pi, s) ds + e^{-\alpha t} P(\pi, t) u$$

whenever the right-hand side is well defined. $U_\alpha(\pi, t, u)$ means the expected discounted total reward in $[0, t]$ under policy π with the terminal reward $u(i)$ at epoch t. For notational simplicity, we denote by $U_\alpha^*(\pi, t) = U_\alpha(\pi, t, U_\alpha^*)$, which is well defined due to Theorem 4.3. Certainly, $U_\alpha^*(\pi, t)$ is the expected discounted total reward if π is used in $[0, t]$ and then an optimal policy is used after t.

First, we have the following lemma.

Lemma 4.6: *$U_\alpha^* = \sup\{U_\alpha^*(\pi, t) \mid \pi \in \Pi_m(p)\}$ for each $t \geq 0$ and $U_\alpha^*(\pi, t)$ is nonincreasing in t for any $\pi \in \Pi_m(p)$.*

Proof: 1. From Eq. (4.6), $U_\alpha^* \leq \sup\{U_\alpha^*(\pi, t) \mid \pi \in \Pi_m(p)\}$. On the other hand, for any given $t > 0$ and $\varepsilon > 0$, we take a policy $\pi^i \in \Pi_m(p)$ with $U_\alpha(\pi^i, i) \geq U_\alpha^*(i) - \varepsilon$ for $i \in S$. Then, for each policy π we define a piecewise semi-Markov policy $\pi(\varepsilon)$ by using π in [0, t) and using π^i in $[t, +\infty)$ if $X(t) = i$. Thus

$$U_\alpha^*(\pi, t) \leq U_\alpha(\pi(\varepsilon)) + e^{-\alpha t} \varepsilon \cdot \boldsymbol{e} \leq U_\alpha^* + e^{-\alpha t} \varepsilon \cdot \boldsymbol{e},$$

where $\boldsymbol{e}$ is the vector with all components being one. Due to the arbitrariness of π and ε, we conclude that $\sup_\pi U^*_\alpha(\pi,t) \le U^*_\alpha$. So the first result is true.

2. It follows from Eqs. (4.5) and (4.6) that for each $\pi \in \Pi_m$ and $t' < t$,

$$\begin{aligned} & \int_{t'}^{t} e^{-\alpha s} P(\pi,t') P(\pi,t',s) r(\pi,s) ds \\ = \;& P(\pi,t') \int_{t'}^{t} e^{-\alpha s} P(\pi,t',s) r(\pi,s) ds. \end{aligned}$$

With the equation above, 1, and Eq. (4.6), we have that for $t' < t$,

$$\begin{aligned} U^*_\alpha(\pi,t) &= \int_0^{t'} e^{-\alpha s} P(\pi,s) r(\pi,s) ds \\ &\quad + e^{-\alpha t'} P(\pi,t') \int_{t'}^{t} e^{-\alpha(s-t')} P(\pi,t',s) r(\pi,s) ds \\ &\quad + e^{-\alpha t'} e^{-\alpha(t-t')} P(\pi,t') P(\pi,t',t) U^*_\alpha \\ &\le \int_0^{t'} e^{-\alpha s} P(\pi,s) r(\pi,s) ds + e^{-\alpha t'} P(\pi,t') U^*_\alpha \\ &= U^*_\alpha(\pi,t'). \end{aligned}$$

Obviously, the above holds also for $\pi \in \Pi_m(p)$ from Eq. (4.5). □

We need the following condition for the optimality equation.

Condition 4.3: *For each $i \in S$ and $a \in A(i)$, there is $f \in F$ and $t > 0$ such that $f(i) = a$ and $U^*_\alpha(f,t,i) > -\infty$.*

The above condition requires that each action should be used by a stationary policy whose criterion value is not negative infinite.

Remark 4.2: *Two sufficient conditions for Condition 4.3 are as follows. (1) the conditions given in Theorem 4.4, especially, when S^- is finite or U^*_α is bounded below (see Remark 4.1). (2) For each $i \in S$, $A(i)$ can be sized down to*

$$A'(i) = \{a \in A(i) \mid \sup_{f \in F: f(i)=a} U_\alpha(f,i) > -\infty\}.$$

This means that action $a \in A(i)$ should be eliminated if any stationary policy f using a will have a negative infinite criterion value. In fact, if $A(i)$ can be sized down to $A'(i)$, then from Lemma 4.6 we have that $U^*_\alpha \ge U^*_\alpha(f,t) \ge U_\alpha(f) > -\infty$ for each $f \in F$ and $t > 0$.

Theorem 4.5: *Under Condition 4.3, U^*_α satisfies the following optimality equation,*

$$\alpha U^*_\alpha(i) = \sup_{a \in A(i)} \{r(i,a) + \sum_j q_{ij}(a) U^*_\alpha(j)\}, \;\; i \in S. \tag{4.18}$$

Proof: 1. For any given i and a, due to Condition 4.3, there are f and $t^* > 0$ such that $f(i) = a$ and $U^*_\alpha(f, t^*, i) > -\infty$. Then with Lemma 4.6 we have that $U^*_\alpha(f, t, i) > -\infty$ for $t \le t^*$. Thus, due to Lemma 4.6 and Corollary 4.2 we can get that for $t \in [0, t^*)$,

$$\begin{aligned} 0 &\ge \frac{d}{dt} U^*_\alpha(f, t, i) \\ &= e^{-\alpha t} \sum_j P_{ij}(f, t)\{r(j, f) + \sum_k q_{jk}(f) U^*_\alpha(k) - \alpha U^*_\alpha(j)\}. \end{aligned}$$

The right-hand side of the above equation is continuous in $t < t^*$. Taking $t = 0$ in the above, we get that

$$\alpha U^*_\alpha(i) \ge r(i, a) + \sum_j q_{ij}(a) U^*_\alpha(j).$$

Due to the arbitrariness of i and a, we get further

$$\alpha U^*_\alpha(i) \ge \sup_{a \in A(i)} \{r(i, a) + \sum_j q_{ij}(a) U^*_\alpha(j)\}, \quad i \in S.$$

2. In the following, we show the reverse of the above inequality is also true. If the above inequality is strict for some i_0, then there is $\varepsilon^* > 0$ such that

$$\alpha U^*_\alpha(i) \ge r(i, a) + \sum_j q_{ij}(a) U^*_\alpha(j) + \varepsilon_i, \quad \forall i, a,$$

where $\varepsilon_{i_0} = \varepsilon^*$ and $\varepsilon_i = 0$ for $i \ne i_0$. We define a vector ε with its ith component being ε_i. So,

$$\alpha U^*_\alpha \ge r(\pi, t) + Q(\pi, t) U^*_\alpha + \varepsilon, \quad \forall \pi \in \Pi_m(p), \ t \ge 0. \tag{4.19}$$

Fixing any $t^* > 0$, for any policy $\pi \in \Pi_m(p)$, if $U^*_\alpha(\pi, t^*, i_0) > -\infty$ then $U^*_\alpha(\pi, t, i_0)$ and so $\sum_j P_{i_0 j}(\pi, t) U^*_\alpha(j)$ are finite in $t \in [0, t^*]$. Due to the uniform boundedness of $\{q_{ij}(a)\}$, we know that $\sum_j P_{i_0 j}(\pi, t) q_{jj}(\pi, t)\ U^*_\alpha(j)$ is finite for $t \in [0, t^*]$. Hence,

$$\begin{aligned} & \sum_j P_{i_0 j}(\pi, t)[\sum_k q_{jk}(\pi, t) U^*_\alpha(k)] \\ = & \sum_j P_{i_0 j}(\pi, t) \sum_{k \ne j} q_{jk}(\pi, t) U^*_\alpha(k) + \sum_j P_{i_0 j}(\pi, t) q_{jj}(\pi, t) U^*_\alpha(j) \\ = & \sum_k [\sum_{j \ne k} P_{i_0 j}(\pi, t) q_{jk}(\pi, t)] U^*_\alpha(k) + \sum_k P_{i_0 k}(\pi, t) q_{kk}(\pi, t) U^*_\alpha(k) \\ = & \sum_k [\sum_j P_{i_0 j}(\pi, t) q_{jk}(\pi, t)] U^*_\alpha(k) \end{aligned}$$

is well defined. Premultiplying Eq. (4.19) by $e^{-\alpha t}P(\pi,t)$, we can get that

$$\begin{aligned} 0 \;\geq\; & \sum_j \frac{d}{dt}[e^{-\alpha t}P_{i_0 j}(\pi,t)U_\alpha^*(j)] \\ & + e^{-\alpha t}\sum_j P_{i_0 j}(\pi,t)[r_j(\pi,t)+\varepsilon_j], \quad t \leq t^*. \end{aligned} \tag{4.20}$$

Now,

$$\begin{aligned} & \sum_j e^{-\alpha t}\frac{d}{dt}P_{i_0 j}(\pi,t)U_\alpha^*(j) \\ = \; & \sum_j e^{-\alpha t}[\sum_{k\neq j} P_{i_0 k}(\pi,t)q_{kj}(\pi,t) + P_{i_0 j}(\pi,t)q_{jj}(\pi,t)]U_\alpha^*(j) \\ = \; & \sum_{j\in S^+} e^{-\alpha t}\sum_{k\neq j} P_{i_0 k}(\pi,t)q_{kj}(\pi,t)U_\alpha^*(j) \\ & + \sum_{j\in S^-} e^{-\alpha t}\sum_{k\neq j} P_{i_0 k}(\pi,t)q_{kj}(\pi,t)U_\alpha^*(j) \\ & + \sum_{j\in S^+} e^{-\alpha t}P_{i_0 j}(\pi,t)q_{jj}(\pi,t)U_\alpha^*(j) \\ & + \sum_{j\in S^-} e^{-\alpha t}P_{i_0 j}(\pi,t)q_{jj}(\pi,t)U_\alpha^*(j). \end{aligned}$$

Because $\sum_{k\neq j} P_{i_0 k}(\pi,t)q_{kj}(\pi,t) \geq 0$ and $q_{jj}(\pi,t) \leq 0$, we have from the above formula that

$$\begin{aligned} & \int_0^{t^*} \sum_j e^{-\alpha t}\frac{d}{dt}P_{i_0 j}(\pi,t)U_\alpha^*(j) \\ = \; & \sum_j \int_0^{t^*} e^{-\alpha t} \\ & \cdot[\sum_{k\neq j} P_{i_0 k}(\pi,t)q_{kj}(\pi,t) + P_{i_0 j}(\pi,t)q_{jj}(\pi,t)]U_\alpha^*(j)dt \\ = \; & \sum_j \int_0^{t^*} e^{-\alpha t}\frac{d}{dt}P_{i_0 j}(\pi,t)U_\alpha^*(j)dt. \end{aligned}$$

Similarly, we have

$$\begin{aligned} & \int_0^{t^*} \sum_j e^{-\alpha t}P_{i_0 j}(\pi,t)U_\alpha^*(j)dt \\ = \; & \sum_j \int_0^{t^*} e^{-\alpha t}P_{i_0 j}(\pi,t)U_\alpha^*(j)dt. \end{aligned}$$

So,

$$\begin{aligned}
& \int_0^{t^*} \sum_j \frac{d}{dt}[e^{-\alpha t} P_{i_0 j}(\pi, t) U_\alpha^*(j)] dt \\
= & \sum_j \int_0^{t^*} \frac{d}{dt}[e^{-\alpha t} P_{i_0 j}(\pi, t) U_\alpha^*(j)] dt \\
= & \sum_j [e^{-\alpha t} P_{i_0 j}(\pi, t) U_\alpha^*(j)] \mid_0^{t^*} \\
= & \sum_j e^{-\alpha t^*} P_{i_0 j}(\pi, t^*) U_\alpha^*(j) - U_\alpha^*(i_0).
\end{aligned}$$

Integrating Eq. (4.20) in $[0, t^*]$ implies that

$$\begin{aligned}
U_\alpha^*(i_0) &\geq U_\alpha^*(\pi, t^*, i_0) + \int_0^{t^*} e^{-\alpha t} \sum_j P_{i_0 j}(\pi, t) \varepsilon_j dt \\
&= U_\alpha^*(\pi, t^*, i_0) + \int_0^{t^*} e^{-\alpha t} P_{i_0 i_0}(\pi, t) dt \varepsilon^* \\
&\geq U_\alpha^*(\pi, t^*, i_0) + \int_0^{t^*} e^{-\alpha t} e^{-\lambda t} dt \varepsilon^*, \qquad (4.21)
\end{aligned}$$

where the last inequality results from the construction of the minimal Q-process. So

$$U_\alpha^*(i_0) > \sup\{U_\alpha^*(\pi, t^*, i_0) \mid \pi \in \Pi_m(p)\},$$

which contradicts Lemma 4.6. Thus Eq. (4.18) is true. □

The concept of policies is generalized in this chapter. But it is often our pleasure to restrict an ($\varepsilon \geq 0$) optimal policy to a smaller and simpler policy set. To do this, our first result is the following theorem, which says that the optimality can be restricted to Π_m, the set of Markov policies, if and only if the optimal value function restricted to Π_m also satisfies the optimality equation (Eq. (4.18)). Let

$$U_\alpha^m = \sup\{U_\alpha(\pi) \mid \pi \in \Pi_m\}$$

be the optimal value function within Markov policies. We affirm that U_α^m is finite. In fact, if $U_\alpha^m(i_0) = -\infty$ for some $i_0 \in S$, then it is easy to see from Eq. (4.5) that $U_\alpha(\pi, i_0) = -\infty$ for each $\pi \in \Pi_m(p)$. Thus, $U_\alpha^*(i_0) = -\infty$, which is a contradiction. Moreover, $U_\alpha^m \leq U_\alpha^* < \infty$. So U_α^m is finite.

Theorem 4.6: *$U_\alpha^* = U_\alpha^m$ if and only if U_α^m is a solution of the optimality equation.*

Proof: It suffices to prove the sufficiency. If U_α^m is a solution of Eq. (4.18), then for each $\pi \in \Pi_m$ and $i \in S$ with $U_\alpha(\pi, i) > -\infty$, we have from Eq. (4.6)

that $U_\alpha(\pi, t, U_\alpha^m, i)$ is finite. So it can be proved as Eq. (4.21) that $U_\alpha^m(i) \geq U_\alpha(\pi, t, U_\alpha^m, i)$ for $t \geq 0$. With this, we can conclude from Eq. (4.5) that $U_\alpha(\pi) \leq U_\alpha^m$ for each $\pi \in \Pi_m(p)$. So $U_\alpha^* = U_\alpha^m$. □

In order to obtain some properties for the optimality equation, we define a set, denoted by W, of finite functions $u = (u(i))$ on S satisfying the following conditions. For each $\pi \in \Pi_m(p)$ and $i \in S$, $\sum_j P_{ij}(\pi, t)u(j) < \infty$ is well defined for all $t \geq 0$. Moreover, we know for each $t \geq 0$, and $i \in S$ that $\sum_j P_{ij}(\pi, t)u(j) > -\infty$ whenever $\sum_j P_{ij}(\pi, t)U_\alpha^*(j) > -\infty$. W is non-empty for $U_\alpha^* \in W$. It is clear that $U_\alpha(\pi, t, u) < \infty$ is well defined for each $u \in W$.

Lemma 4.7: *Suppose that $\varepsilon \geq 0, \beta + \alpha \geq 0, u \in W$, $\pi \in \Pi_m$, and $i \in S$. If π and u satisfy the following two conditions then $u(i) \leq U_\alpha(\pi, i) + (\beta + \alpha)^{-1}\varepsilon$.*

$$\alpha u \leq r(\pi, t) + Q(\pi, t)u + e^{-\beta t}\varepsilon e, \quad \text{a.e. } t \geq 0, \tag{4.22}$$

$$\liminf_{t\to\infty} e^{-\alpha t} \sum_j P_{ij}(\pi, t)u(j) \leq 0. \tag{4.23}$$

Proof: From the given conditions, it can be proved as Eq. (4.21) that

$$u(i) \leq U_\alpha(\pi, t, u, i) + \int_0^t e^{-(\beta+\alpha)s} ds \cdot \varepsilon.$$

By letting $\liminf_{t\to\infty}$ above, due to Eq. (4.23), we have $u(i) \leq U_\alpha(\pi, i) + (\beta + \alpha)^{-1}\varepsilon$. □

With the lemma above, we now prove the following theorem, which compares the optimal value function with other solutions of the optimality equation under certain conditions.

Theorem 4.7: *Suppose that $u \in W$ is a solution of the optimality equation (Eq. (4.18)). Then, for any given state $i \in S$,*

1. *If for some $\beta > -\alpha$ and each $\varepsilon > 0$, there is a policy $\pi \in \Pi_m(p)$ with $U_\alpha(\pi, i) > -\infty$ satisfying Eqs. (4.22) and (4.23), then $u(i) \leq U_\alpha^*(i)$.*
2. *$U_\alpha^*(i)$ is the smallest solution of the optimality equation that satisfies the following Eq. (4.24) for each $\pi \in \Pi_m(p)$ with $U_\alpha(\pi, i) > -\infty$,*

$$\limsup_{t\to\infty} e^{-\alpha t} \sum_j P_{ij}(\pi, t)u(j) \geq 0. \tag{4.24}$$

Proof: 1. $U_\alpha(\pi, i) > -\infty$ implies that $U_\alpha^*(\pi, t, i)$ and so $U_\alpha(\pi, t, u, i)$ are finite for all $t \geq 0$ due to the definition of W. Thus, from Lemma 4.7,

$$u(i) \leq U_\alpha(\pi, i) + (\beta + \alpha)^{-1}\varepsilon \leq U_\alpha^*(i) + (\beta + \alpha)^{-1}\varepsilon.$$

So $u(i) \leq U_\alpha^*(i)$ from the arbitrariness of ε.

2. Suppose that u is a solution of Eq. (4.18) and satisfies Eq. (4.24) for each $\pi \in \Pi_m(p)$ with $U_\alpha(\pi, i) > -\infty$. Then, we have

$$\alpha u \geq r(\pi, t) + Q(\pi, t)u, \quad t \geq 0, \pi \in \Pi_m(p).$$

Thus, it can be proved as Eq. (4.21) that for each $\pi \in \Pi_m(p)$ with $U_\alpha(\pi, i) > -\infty$, $u(i) \geq U_\alpha(\pi, t, u, i)$ for each $t \geq 0$. Letting $\limsup_{t\to\infty}$ implies $u(i) \geq U_\alpha(\pi, i)$ due to Eq. (4.24). Hence, $u(i) \geq U_\alpha^*(i)$.

On the other hand, U_α^* often satisfies Eq. (4.24) for $\pi \in \Pi_m(p)$ with $U_\alpha(\pi, i) > -\infty$. In fact, from Eq. (4.6) we know that if $U_\alpha(\pi, i) > -\infty$ then $\sum_j P_{ij}(\pi, t)U_\alpha^*(j)$ is also finite for each $t \geq 0$ and

$$\begin{aligned} & \limsup_{t\to\infty} e^{-\alpha t} \sum_j P_{ij}(\pi, t)U_\alpha^*(j) \\ \geq \quad & \limsup_{t\to\infty} e^{-\alpha t} \sum_j P_{ij}(\pi, t)U_\alpha(\pi, t, j) = 0. \end{aligned}$$

This completes the proof. □

Equation (4.23) is true if $u \leq 0$ or if $\alpha > 0$ and u is bounded above, and Eq. (4.24) is true if $u \geq 0$ or if $\alpha > 0$ and u is bounded below.

It is clear that there is often a policy $\pi = (f_t) \in \Pi_m^d$ satisfying Eq. (4.22), but $U_\alpha(\pi, i) > -\infty$ may be not true.

The following corollary can be proved easily by Theorem 4.7 and Lemma 4.7.

Corollary 4.3: *Provided that Eq. (4.18) holds,*

1. *for any given $f \in F$, if f attains the supremum of Eq. (4.18), f and U_α^* satisfy Eq. (4.23) and $U_\alpha(f) > -\infty$, then f is optimal.*
2. *For any policy $\pi^* \in \Pi_m(p)$, if $U_\alpha(\pi^*)$ is a solution of Eq. (4.18) then π^* is optimal.*
3. *If for any $\varepsilon > 0$, there is a policy $\pi \in \Pi_m^d$ with $U_\alpha(\pi) > -\infty$, π and U_α^* satisfy Eqs. (4.22) and (4.23) for each $i \in S$, then $U_\alpha^* = \sup\{U_\alpha(\pi) \mid \pi \in \Pi_m^d\}$.*
4. *If $\alpha > 0, \varepsilon \geq 0, f \in F$ attains the ε-supremum of Eq. (4.18), f and U_α^* satisfy Eq. (4.23) with $U_\alpha(f) > -\infty$, then f is $\alpha^{-1}\varepsilon$-optimal. Moreover, if such f exists for each $\varepsilon > 0$, then $U_\alpha^* = \sup\{U_\alpha(f) \mid f \in F\}$.*
5. *If $U_\alpha^* \leq 0$, then U_α^* is the largest solution of Eq. (4.18) in W satisfying conditions given in 1 of Theorem 4.7.*
6. *U_α^* is the smallest solution of Eq. (4.18) in W satisfying Eq. (4.24) for $\pi \in \Pi_m(p)$ and $i \in S$ with $U_\alpha(\pi, i) > -\infty$.*

7. *If $\alpha > 0$ and the reward rate is uniformly bounded, U_α^* is the unique bounded solution of Eq. (4.18).*

In the above corollary, conclusions 1 and 2 give sufficient conditions for a policy to be optimal, 3 and 4 characterize the optimality within smaller policy sets, and 5 and 6 characterize the optimal value function as the largest solution or the smallest solution of the optimality equation under certain conditions.

Corollary 4.4: *For each $f \in F$ and $i \in S$ with $U_\alpha(f,i) > -\infty$, we have that $\sum_j q_{ij}(f)U_\alpha(f,j)$ is finite and*

$$\alpha U_\alpha(f,i) = r(i,f) + \sum_j q_{ij}(f)U_\alpha(f,j). \tag{4.25}$$

Moreover, $U_\alpha(f,j)$ is finite for $j \in S$ with $q_{ij}(f) > 0$.

Proof: Eq. (4.6) for $f \in F$ and $i \in S$ can be rewritten as

$$\begin{aligned} U_\alpha(f,i) &= \int_0^t e^{-\alpha s} \sum_j P_{ij}(f,s) r(j,f) ds \\ &\quad + e^{-\alpha t} \sum_j P_{ij}(f,t) U_\alpha(f,j). \end{aligned}$$

If $U_\alpha(f,i) > -\infty$ then from Corollary 4.2 we have that for $t \geq 0$,

$$\begin{aligned} 0 &= \frac{d}{dt} U_\alpha(f,i) \\ &= e^{-\alpha t} \sum_j P_{ij}(f,t)\{r(j,f) + [Q(f)U_\alpha(f)]_j - \alpha U_\alpha(f,j)\}. \end{aligned}$$

Its right-hand side is continuous and so is true for $t = 0$. Then, Eq. (4.25) is true. □

With the above corollary, $U_\alpha(f,i)$ can be solved through the set of linear equations (4.25) under the given conditions.

To conclude this subsection, we discuss the CTMDP model Eq. (4.1) restricted to $\Pi_s^d(p)$, the set of piecewise semi-stationary policies. In this case, Lemma 4.2 is still true except that "$\leq U(i)$" should be deleted in Eq. (4.10), and $A_2(i)$ defined by Eq. (4.8) should be redefined by

$$\begin{aligned} A_2(i) &= \{a \in A_1(i) \mid \text{there is } f \in F \text{ such that } f(i) = a \\ &\qquad \text{and } U_\alpha(f,i) > -\infty\}. \end{aligned}$$

Thus, Condition 4.3 is trivial. By noting that Corollaries 4.2 and 4.4 also hold for stationary policies f, the following theorem can be proved similarly to

Theorems 4.5 and 4.6, where we let

$$\begin{aligned} U_\alpha^{*d} : &= \sup\{U_\alpha(\pi) \mid \pi \in \Pi_s^d(p)\}, \\ U_\alpha^{s} : &= \sup\{U_\alpha(f) \mid f \in F\} \end{aligned}$$

be the optimal value functions among the policy sets $\Pi_s^d(p)$ and F, respectively.

Theorem 4.8: *If the CTMDP model of Eq. (4.1) is restricted to $\Pi_s^d(p)$, then U_α^{*d} satisfies the optimality equation (4.18). Moreover, U_α^s satisfies the optimality equation if and only if $U_\alpha^{*d} = U_\alpha^s$.*

In the previous subsections, we applied the methods and ideas from DTMDPs in Chapter 2 to stationary CTDMPs. Under the conditions that the model is well defined, we decompose the model into three subparts. Within the subpart where the optimal value if finite, we study the optimality equation and optimal policies.

In the following two sections, we study a nonstationary CTMDP model with the total reward criterion and a stationary CTMDP model with the average criterion.

2. A Nonstationary Model: Total Reward

In Section 1, we studied a stationary CTMDP model. In this section, we investigate a nonstationary CTMDP model with the expected total reward criterion.

2.1 Model and Conditions

In this section, we study the following nonstationary CTMDP model,

$$\{S, (A(i), i \in S), q(t), r(t), U\}.$$

Here, S is the countable state space and we assume $S = \{0, 1, 2, \ldots\}$. For $i \in S$, $A(i)$ is the feasible action set when the system is in state i. We assume $A(i)$ to be also countable. We define the set $\Gamma = \{(i, a) \mid a \in A(i), i \in S\}$. $q(t) = \{q_{ij}(t, a) \mid (i, a) \in \Gamma, j \in S\}$ is the transition rate family at time t; that is, if the system is in state i at time t and action $a \in A(i)$ is used in time interval $[t, t + \Delta t]$ for Δt small enough, then the probability that the system will transfer to state j at time $t + \Delta t$ is

$$P_{ij}(t, t + \Delta t) = \delta_{ij} + q_{ij}(t, a)\Delta t + o(\Delta t).$$

We suppose that $q(t)$ satisfies the following two conditions.

1. $q_{ij}(t, a) \geq 0$ for $i, j \in S, i \neq j, a \in A(i), t \geq 0$.
2. $\sum_j q_{ij}(t, a) = 0$ for $(i, a) \in \Gamma, t \geq 0$.

The reward rate obtained by the system at time t is $r_i(t, a)$. U is the expected total reward, which is defined in the following.

Policies and policy sets are the same as those in the stationary CTMDP model.

For a policy $\pi = (\pi_t) \in \Pi_m$, we define a matrix $Q(\pi, t) = (q_{ij}(\pi, t))$ and a column vector $r(\pi, t) = (r_i(\pi, t))$ as follows.

$$\begin{aligned} q_{ij}(\pi, t) &= \sum_{a \in A(i)} q_{ij}(t, a)\pi_t(a \mid i), \ \ i, j \in S, t \geq 0, \\ r_i(\pi, t) &= \sum_{a \in A(i)} r_i(t, a)\pi_t(a \mid i), \ \ i, j \in S, t \geq 0. \end{aligned}$$

$\{q_{ij}(\pi, t)\}$ and $r_i(\pi, t)$ are, respectively, the state transition rate family and the reward rate of the process under the policy π at time t.

Similar to that in the stationary CTMDPs, for each $\pi \in \Pi_m$, the $Q(\pi, t)$-process $\{P(\pi, s, t), 0 \leq s \leq t < \infty\}$ is probably not unique. So, we make the following condition.

Condition 4.4: *1. For each $\pi \in \Pi_m, i, j \in S, q_{ij}(\pi, t)$ is almost everywhere (a.e.) continuous.*

2. There exists a function $Q(t)$ that is integrable in every finite time interval such that

$$-q_{ii}(t, a) \leq Q(t), \ \text{ a.e. } t, (i, a) \in \Gamma.$$

By this condition, we know from Liu et al. [88] that for each $\pi \in \Pi_m$, there exists a unique $Q(\pi, t)$-process $\{P(\pi, s, t), 0 \leq s \leq t < \infty\}$ satisfying:

1. $\frac{\partial}{\partial t} P(\pi, s, t) = P(\pi, s, t)Q(\pi, t), \ 0 \leq s \leq t < \infty.$
2. $\sum_j P_{ij}(\pi, s, t) = 1, \ 0 \leq s \leq t < \infty.$
3. $P_{ij}(\pi, s, s) = \delta_{ij}, \ 0 \leq s < \infty.$
4. $P(\pi, s, u) = P(\pi, s, t)P(\pi, t, u), \ 0 \leq s \leq t \leq u < \infty.$

For the above condition, we make the following remark.

Remark 4.3: *(1) If $q_{ij}(t, a)$ is uniformly bounded (in i, j, t, a) and is Lebesgue measurable, then we know from Kakumanu [82] that for each $\pi \in \Pi_m$, the $Q(\pi, t)$-process uniquely exists and satisfies 1–4 above. (2) $Q(t)$ may be unbounded. For example, $Q(t) = t$.*

From Condition 4.4, we have the following lemma.

Lemma 4.8: *For any given constant $\beta \in (0, 1)$, there exists a sequence $\{t_n, n \geq 0\}$ such that*

1. $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} < \cdots, \lim_n t_n = +\infty.$

2. $\int_{t_n}^{t_{n+1}} 2Q(t)dt \leq \beta$ for all $n \geq 0$.

Proof: From Condition 4.4, we know that for each $n \geq 0$, $\beta_n := \int_n^{n+1} Q(t)dt$ is finite. So, there exist $n = t_{n,0} < t_{n,1} < \cdots < t_{n,k_n} = n+1$ such that

$$\int_{t_{n,k}}^{t_{n,k+1}} Q(t)dt \leq \frac{\beta}{2}, \quad k = 0, 1, \ldots, k_n - 1.$$

We then get t_n by listing $t_{n,k}$. That is, let $s_0 = 0$ and $s_n = \sum_{l=0}^{n-1} k_l$ for $n = 1, 2, \ldots$. Then, $t_0 = 0$ and $t_{s_n+k} = t_{n,k}$ for $k = 1, 2, \ldots, k_n$ and $n \geq 0$. It is apparent that such $\{t_n\}$ satisfies Lemma 4.8. □

Lemma 4.8 above is important. It ensures that the operators introduced later will be contracted.

About the reward rate function, we introduce the following condition.

Condition 4.5: *For each $(i, a) \in \Gamma$, $r_i(t, a)$ is Lebesgue measurable in t. Moreover, there exists an integrable function $r(t)$ for $t \in [0, \infty)$ such that*

$$|r_i(t, a)| \leq r(t), \quad \text{a.e. } t, \forall (i, a) \in \Gamma.$$

Now, we define the criterion by

$$U(\pi, t) = \int_t^\infty P(\pi, t, s) r(\pi, s) ds, \quad t \geq 0, \pi \in \Pi_m.$$

$U(\pi, t)$ represents the expected total reward in the time interval (t, ∞) under policy π. Obviously, due to Condition 4.5,

$$|U(\pi, t)| \leq \int_t^\infty r(s) ds \cdot \boldsymbol{e},$$

where $\boldsymbol{e}$ is a column vector with all components being 1. Hence, $U(\pi, t, i)$ is well defined and uniformly bounded in $\pi \in \Pi_m$ and $i \in S$ for each $t \geq 0$.

Let the optimal value function be

$$U^*(t) = \sup\{U(\pi, t) : \pi \in \Pi_m\}, \quad t \geq 0.$$

Moreover, let $U(\pi) = U(\pi, 0)$. For $\pi^* \in \Pi_m$ and $\varepsilon \geq 0$, if $U(\pi^*) \geq U(0) - \varepsilon \boldsymbol{e}$, then we call π^* an ε-optimal policy. If $\varepsilon = 0$, we call π^* an optimal policy.

2.2 Optimality Equation

We need a well-structured space, denoted by Ω, that includes all $U(\pi, t)$. This is defined as the set of all real column functions $x(t) = \{x_i(t), i \in S, t \in [0, \infty)\}$ satisfying the following three conditions.

1. $x_i(t)$ is absolutely continuous in t for each $i \in S$. So, its differential, denoted by $x_i'(t)$, exists a.e. and is measurable.

2. $x_i(t)$ converges to zero uniformly in $i \in S$ as t tends to infinity.

3. There exists a function $N(t)$ that is integrable in every finite time interval such that $|x_i'(t)| \leq N(t)$, a.e. t, for $i \in S$.

It should be noted that $N(t)$ above may depend on $x(t)$. It is easy to conclude that each $x(t) \in \Omega$ is uniformly bounded.

First, we have the following lemma.

Lemma 4.9: *For $\pi \in \Pi_m(c), \varepsilon \geq 0, U(t) \in \Omega, \alpha > 0$,*

1. *If $-U'(t) \leq r(\pi,t)+Q(\pi,t)U(t)+\varepsilon e^{-\alpha t}\boldsymbol{e}$, a.e. t, then $U(t) \leq U(\pi,t)+\alpha^{-1}e^{-\alpha t}\varepsilon\boldsymbol{e}$ for all $t \geq 0$.*

2. *If $-U'(t) \geq r(\pi,t)+Q(\pi,t)U(t)-\varepsilon e^{-\alpha t}\boldsymbol{e}$, a.e. t, then $U(t) \geq U(\pi,t)-\alpha^{-1}e^{-\alpha t}\varepsilon\boldsymbol{e}$ for all $t \geq 0$.*

Proof: We only prove 1 in the following and 2 can be proved similarly. For $U(t) \in \Omega$, there exists a function $N(t)$ that is integrable in every finite time interval such that $|U_i'(t)| \leq N(t)$, a.e. t, for $i \in S$. For $U_i(t)$ uniformly bounded, there exists a constant K such that $|U_i(t)| \leq K$.

Premultiplying the given condition by $P_{ij}(\pi,s,t)$, we get

$$\begin{aligned} & -\sum_j P_{ij}(\pi,s,t)U_j'(t) \\ \leq\ & \sum_j P_{ij}(\pi,s,t)r_j(\pi,t) + \sum_j P_{ij}(\pi,s,t)\sum_k q_{jk}(\pi,t)U_k(t) \\ & + e^{-\alpha t}\varepsilon, \ \text{ a.e. } t,\ i \in S. \end{aligned}$$

Because

$$\begin{aligned} & \sum_j\sum_k |P_{ij}(\pi,s,t)q_{jk}(\pi,t)U_k(t)| \\ \leq\ & \sum_j\sum_k P_{ij}(\pi,s,t)|q_{jk}(\pi,t)|K \leq 2KQ(t) \end{aligned}$$

is finite a.e., $\sum_j$ and $\sum_k$ can be exchanged for a.e. t. Let $[\cdot]_t'$ denote the differential of the function $[\cdot]$ in t. Then,

$$\begin{aligned} & -\sum_j [P_{ij}(\pi,s,t)U_j(t)]_t' \\ =\ & -\sum_j \left\{ \sum_k P_{ik}(\pi,s,t)q_{kj}(\pi,t)U_j(t) + P_{ij}(\pi,s,t)U_j'(t) \right\} \\ \leq\ & \sum_j P_{ij}(\pi,s,t)r_j(\pi,t) + e^{-\alpha t}\varepsilon, \ \text{ a.e. } t,\ i \in S. \end{aligned} \tag{4.26}$$

For $n = \infty, 0, 1, 2, \ldots, i \in S$, we define a function $x_{n,i}$ as follows,

$$x_{n,i}(s,t) = -\sum_{j=0}^{n} [P_{ij}(\pi,s,t)U_j(t)]'_t .$$

Then, similarly to Eq. (4.26) we have

$$\begin{aligned} |x_{n,i}(s,t)| &\leq \sum_{j=0}^{\infty} P_{ij}(\pi,s,t)N(t) + \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} P_{ij}(\pi,s,t)|q_{jk}(\pi,t)|K \\ &\leq 2KQ(t) + N(t), \text{ a.e. } t, i \in S. \end{aligned}$$

This implies that $\lim_{n\to\infty} x_{n,i}(s,t) = x_{\infty,i}(s,t)$. Because $2KQ(t) + N(t)$ is integrable in every finite interval, it follows the Lebesgue Control Convergence Theorem that

$$\begin{aligned} \int_s^T x_{\infty,i}(s,t)dt &= -\sum_j \int_s^T [P_{ij}(\pi,s,t)U_j(t)]'_t \, dt \\ &= \sum_j [P_{ij}(\pi,s,s)U_j(s) - P_{ij}(\pi,s,T)U_j(T)] \\ &= U_i(s) - \sum_j P_{ij}(\pi,s,T)U_j(T), \quad i \in S, \end{aligned}$$

for $0 \leq s \leq T < \infty$. With this in mind, by integrating Eq. (4.26) in $t \in [s,T]$ and letting $T \to \infty$, we get from the fact that $U_j(T)$ tends to zero uniformly in $j \in S$,

$$U_i(s) \leq \int_s^{\infty} \sum_j P_{ij}(\pi,s,t)r_j(\pi,t)dt + \alpha^{-1}e^{-\alpha s}\varepsilon, \quad s \geq 0, i \in S.$$

This completes the proof. □

We use operators to study the optimality equation. For each $n \geq 0$, let M_n be the set of all bounded measurable column vector functions on $[t_n, t_{n+1}]$. The distance d in M_n is defined by

$$d(x,y) = \sup\{|x_i(t) - y_i(t)| : t_n \leq t \leq t_{n+1}, i \in S\}, \quad x, y \in M_n.$$

Obviously, (M_n, d) is a Banach space. Moreover, for $\pi \in \Pi_m$, we define an operator $T_{n\pi}$ by

$$(T_{n\pi}x)_i(t) = \int_t^{t_{n+1}} \{r_i(\pi,s) + \sum_j q_{ij}(\pi,s)x(s)_j\}ds + z_{n+1}(i),$$

for $i \in S, t \in [t_n, t_{n+1}], x \in M_n$, where z_{n+1} is any given bounded column vector. It is easy to show by Lemma 4.8 that $T_{n\pi}$ is a contraction operator with the module β on M_n for each $\pi \in \Pi_m$.

Theorem 4.9: *For $\pi \in \Pi_m, U(\pi,t)$ is the unique solution in Ω of the following equation,*

$$-U'(t) = r(\pi,t) + Q(\pi,t)U(t), \text{ a.e. } t. \tag{4.27}$$

Proof: Suppose that $U(t) \in \Omega$ is a solution of Eq. (4.27), then from Lemma 4.9 we know that $U(t) = U(\pi,t)$. So, the solution of Eq. (4.27), if it exists, is unique.

For $T_{n\pi}$ a contraction mapping, we can prove that $U(\pi,t)$ is a solution in Ω of Eq. (4.27) in a similar way as that in Theorem 4.10 later. □

To prove the optimality equation, it is necessary to define the following operators $T_n (n \geq 0)$. For any given bounded column vector z_{n+1}, define for $x \in M_n$,

$$(T_n x)(t)_i = \int_t^{t_{n+1}} \sup_{a \in A(i)} \{r_i(s,a) + \sum_j q_{ij}(s,a)x_j(s)\}ds + z_{n+1}(i) \tag{4.28}$$

for $i \in S, t \in [t_n, t_{n+1}]$. From Conditions 4.4 and 4.5, the integrand above is measurable and has an integrable upper bound function $r(t) + 2KQ(t)$, where K is a bound of $x_i(t)$. Then, $T_n x \in M_n$. Moreover, due to Lemma 4.8, we know that for any $x, y \in M_n$,

$$\begin{aligned} |T_n x_i(t) - T_n y_i(t)| &\leq \int_{t_n}^{t_{n+1}} \sup_{a \in A(i)} \left| \sum_j q_{ij}(s,a)[x_j(s) - y_j(s)] \right| ds \\ &\leq \int_{t_n}^{t_{n+1}} 2Q(s)d(x,y)ds \leq \beta d(x,y), \quad i \in S. \end{aligned}$$

So, $d(T_n x, T_n y) \leq \beta d(x,y)$ and T_n is also a contraction mapping in M_n (for all $n \geq 0$).

Lemma 4.10: *For each $n \geq 0$ and any given bounded vector z_{n+1}, there exists a unique $U_n(t) \in M_n$ such that the following two statements hold.*

1. *$U_n(t) = T_n U_n(t)$ for $t \in [t_n, t_{n+1}]$ and $U_n(t_{n+1}) = z_{n+1}$.*

2. *For every $i \in S, U_{n,i}(t)$ is absolutely continuous in $t \in [t_n, t_{n+1}]$ and*

$$-U'_{n,i}(t) = \sup_{a \in A(i)} \{r_i(t,a) + \sum_j q_{ij}(t,a)U_{n,j}(t)\}, \text{a.e. } t \in [t_n, t_{n+1}]. \tag{4.29}$$

Proof: 1 follows the Banach Fixed Theorem and 2 is equivalent to 1 due to Eq. (4.28). □

Equation (4.29) says in some degree that the optimality equation is true in each interval $[t_n, t_{n+1})$. In order to show the optimality equation in the whole real line, we need the following lemma.

Lemma 4.11: *For each $U(t) \in \Omega$ and $\varepsilon > 0$, there is a policy $\pi = (f_t) \in \Pi_m^d$ such that*

$$\begin{aligned} & r_i(t, f_t(i)) + \sum_j q_{ij}(t, f_t(i)) U_j(t) \\ \geq & \sup_{a \in A(i)} \{r_i(t,a) + \sum_j q_{ij}(t,a) U_j(t)\} - \varepsilon, \quad t \geq 0, i \in S. \end{aligned}$$

Proof: Because $A(i)$ is countable, we let $A(i) = \{a_1^i, a_2^i, \ldots\}$ for $i \in S$. Let

$$h_n(t,i) := r_i(t, a_n^i) + \sum_j q_{ij}(t, a_n^i) U_j(t).$$

Then, for each $i \in S$ and $n \geq 0$, $h_n(t,i)$ is measurable in t. So, $\sup_n h_n(t,i)$ is also measurable, and for each $i \in S$ and $n \geq 1$,

$$J'_{n,i} := \{t \geq 0 \mid \sup_m h_m(t,i) - h_n(t,i) \leq \varepsilon\}$$

is a measurable set. If we define the disjoint sets $J_{n,i}$ as follows,

$$J_{1,i} = J'_{1,i}, \;\; J_{n,i} = J'_{n,i} - \bigcup_{k=1}^{n-1} J'_{k,i}, \; n \geq 2,$$

then the policy $\pi = (f_t)$ defined as follows is the required,

$$f_t(i) = a_n^i, \;\; \text{if } t \in J_{n,i}, n \geq 1, \; i \in S, \; t \geq 0.$$

This completes the proof. □

Shrev and Bertsekas [131] showed a generalized result for the measurable selectors. But in this book the lemma above is enough. Having the above results we can prove the optimality equation in the following.

Theorem 4.10: *$U^*(t)$ is the unique solution of the following optimality equation in Ω.*

$$-U_i'(t) = \sup_{a \in A(i)} \{r_i(t,a) + \sum_j q_{ij}(t,a) U_j(t)\}, \;\; \text{a.e. } t, i \in S. \tag{4.30}$$

Moreover, $U^*(t) = \sup\{U(\pi, t) : \pi \in \Pi_m^d\}$ for $t \geq 0$, and for every $\varepsilon > 0$ there exist ε-optimal policies.

Proof: First, we prove the existence of a solution of Eq. (4.30). Define that

$$z_n = \sup\{U(\pi, t_n) | \pi \in \Pi_m\}, \;\; n \geq 0.$$

Then, for each $n \geq 0$, by Lemma 4.10, there exists a unique $U_n(t) \in M_n$ such that $U_n(t)$ is absolutely continuous in $t \in [t_n, t_{n+1}]$ and satisfies Eq. (4.29). Certainly, $U_n(t)$ also satisfies

$$U_{n,i}(t) = \int_t^{t_{n+1}} \sup_{a \in A(i)} \{r_i(s,a) + \sum_j q_{ij}(s,a) U_{n,j}(s)\} ds + z_{n+1}(i) \quad (4.31)$$

for $i \in S$ and $t \in [t_n, t_{n+1}]$. We define a function U_* as follows,

$$U_*(t) = U_n(t), \text{ if } t \in [t_n, t_{n+1}), \ n \geq 0.$$

Then, $U_*(t)$ is differentiated a.e. $t \in [0, \infty)$ and satisfies Eq. (4.30) because each $U_n(t)$ satisfies Eq. (4.29).

In order to prove that $U_*(t) \in \Omega$, we first prove that $U_{*,i}(t)$ converges uniformly to zero. It is clear that

$$|z_n(i)| \leq M_n = \int_{t_n}^{\infty} r(t) dt, \ i \in S, n \geq 0.$$

Let

$$K_n = \sup\{|U_{n,i}(t)|, t \in [t_n, t_{n+1}], i \in S\}, \ n \geq 0.$$

Then, due to Eq. (4.31)

$$K_n \leq \int_{t_n}^{t_{n+1}} [r(s) + 2Q(s)K_n] ds + M_{n+1} \leq M_n + \beta K_n.$$

So, $K_n \leq (1-\beta) - 1 M_n$, which together with $\lim_{n\to\infty} M_n = 0$ implies that $U_{*,i}(t)$ converges uniformly to zero. Moreover, $|U_{*,i}(t)| \leq (1-\beta)^{-1} M_0$.

Second, because $U_*(t)$ satisfies Eq. (4.29), we have

$$|U'_{*,i}(t)| \leq r(t) + 2Q(t)(1-\beta)^{-1} M_0, \text{ a.e. } t, i \in S.$$

So, $U_*(t)$ satisfies condition 3 in the definition of Ω (at the beginning of this subsection).

Now we prove that $U_*(t)$ is absolutely continuous. It follows Eq. (4.29) that for any $\pi = (\pi_t) \in \Pi_m$,

$$-U'_n(t) \geq r(\pi, t) + Q(\pi, t) U_n(t), \text{ a.e. } t \in [t_n, t_{n+1}] \, n \geq 0.$$

With this, it can be proved as Lemma 4.9 that for $n \geq 0$,

$$U_n(t) - P(\pi, t, t_{n+1}) U_n(t_{n+1}) \geq \int_t^{t_{n+1}} P(\pi, t, s) r(\pi, s) ds, \ t \in [t_n, t_{n+1}].$$

This implies due to $U_n(t_{n+1}) = z_{n+1}$ that

$$U_n(t) \geq \sup_{\pi} \left\{ \int_t^{t_{n+1}} P(\pi, t, s) r(\pi, s) ds + P(\pi, t, t_{n+1}) z_{n+1} \right\}, \quad t \in [t_n, t_{n+1}], n \geq 0. \quad (4.32)$$

On the other hand, from Lemma 4.11 we know that for any $\alpha > 0$, there is $\pi^* \in \Pi_m^d$ such that for any $n \geq 0$,

$$-U_n'(t) \leq r(\pi^*, t) + Q(\pi^*, t)U_n(t) + e^{-\alpha t}\boldsymbol{e}, \text{ a.e. } t \in [t_n, t_{n+1}].$$

Again, it can be proved as Lemma 4.11 that

$$\begin{aligned} U_n(t) &\leq \int_t^{t_{n+1}} P(\pi^*, t, s)r(\pi^*, s)ds + P(\pi^*, t, t_{n+1})z_{n+1} \\ &\quad + \alpha^{-1}[e^{-\alpha t} - e^{-\alpha t_{n+1}}]\boldsymbol{e} \\ &\leq \sup_{\pi \in \Pi_m^d} \left\{ \int_t^{t_{n+1}} P(\pi, t, s)r(\pi, s)ds + P(\pi, t, t_{n+1})z_{n+1} \right\} \\ &\quad + \alpha^{-1}[e^{-\alpha t} - e^{-\alpha t_{n+1}}]\boldsymbol{e}. \end{aligned} \tag{4.33}$$

Letting $\alpha \to \infty$, one can get together with Eq. (4.32) that

$$U_n(t) = \sup_{\pi \in \Pi_m^d} \left\{ \int_t^{t_{n+1}} P(\pi, t, s)r(\pi, s)ds + P(\pi, t, t_{n+1})z_{n+1} \right\}, \quad t \in [t_n, t_{n+1}],\ n \geq 0. \tag{4.34}$$

By the definition of z_n and Eq. (4.34), $U_n(t_n) \geq z_n$. This implies by Eq. (4.33) that for $n \geq 0$,

$$\begin{aligned} z_n \leq U_n(t_n) &\leq \int_{t_n}^{t_{n+1}} P(\pi^*, t_n, s)r(\pi^*, s)ds + P(\pi^*, t_n, t_{n+1})z_{n+1} \\ &\quad + \alpha^{-1}[e^{-\alpha t_n} - e^{-\alpha t_{n+1}}]\boldsymbol{e}. \end{aligned}$$

By using the induction method, we can obtain that for $N > n \geq 0$,

$$\begin{aligned} U_n(t_n) &\leq \int_{t_n}^{t_{n+N}} P(\pi^*, t_n, s)r(\pi^*, s)ds + P(\pi^*, t_n, t_{n+N})z_{n+N} \\ &\quad + \alpha^{-1}[e^{-\alpha t_n} - e^{-\alpha t_{n+N}}]\boldsymbol{e}. \end{aligned}$$

Letting $N \to \infty$, one gets from $\lim_{n\to\infty} z_n = 0$ that

$$\begin{aligned} z_n \leq U_n(t_n) &\leq U(\pi^*, t_n) + \alpha^{-1}e^{-\alpha t_n}\boldsymbol{e} \\ &\leq z_n + \alpha^{-1}e^{-\alpha t_n}\boldsymbol{e}, \quad n \geq 0. \end{aligned} \tag{4.35}$$

From the arbitrariness of α, we have

$$U_n(t_n) = z_n, \quad n \geq 0.$$

This together with the facts of $U_n(t_{n+1}) = z_{n+1}$ and $U_n(t)$ is absolutely continuous in $[t_n, t_{n+1}]$ implies that $U_*(t)$ is absolutely continuous in $[0, \infty)$.

Therefore, $U_*(t) \in \Omega$.

About the uniqueness of the solution of Eq. (4.30), suppose that $U(t) \in \Omega$ is a solution of Eq. (4.30). Then, one can prove as Eq. (4.34) (t_{n+1} is replaced by ∞) that $U(t) = U^*(t)$ for $t \geq 0$.

Finally, it follows Lemma 4.11 that for each $\varepsilon > 0$ there is policy π^* that satisfies the condition in 1 of Lemma 4.9. Thus, with Lemma 4.11 we know that π^* is $\alpha^{-1}\varepsilon$-optimal. □

The above theorem shows the validity of the optimality equation and the optimality of the deterministic Markov policies.

The criterion considered above is the expected total reward, which is equivalent to the expected discounted total reward for nonstationary CTMDPs.

First, the expected discounted total reward is a special case of expected total reward. We assume that

$$r_i(t,a) = e^{-\alpha t}\hat{r}_i(t,a), \;\; t \geq 0, (i,a) \in \Gamma. \tag{4.36}$$

Here, $\alpha > 0$ is the fixed discount rate and $\hat{r}_i(t,a)$ is the reward rate when the discount is considered. We define $\hat{r}(\pi,t)$ similarly to $r(\pi,t)$. Then

$$\begin{aligned} U(\pi,t) &= \int_t^\infty P(\pi,t,s)r(\pi,s)ds \\ &= e^{-\alpha t}\int_t^\infty e^{-\alpha(s-t)}P(\pi,t,s)\hat{r}(\pi,s)ds \\ &= e^{-\alpha t}\hat{U}_\alpha(\pi,t), \end{aligned} \tag{4.37}$$

where $\hat{U}_\alpha(\pi,t)$ is the expected discounted total reward in $[t,\infty)$ under policy π when the reward rate function is $\hat{r}_i(t,a)$ and the discount rate is α. Due to Eq. (4.30), the optimality equation for the discounted criterion $\hat{U}_\alpha(\pi,t)$ is

$$\begin{aligned} -U_i'(t) &= \sup_{a\in A(i)}\{\hat{r}_i(t,a) + \sum_j q_{ij}(t,a)U_j(t) - \alpha U_i(t)\}, \\ &\qquad\qquad \text{a.e. } t,\; i \in S. \end{aligned} \tag{4.38}$$

Comparing the above equation with the optimality equation (4.30), we see that the term "$-\alpha U_i(t)$" is added here. Let

$$\hat{U}_\alpha^*(t) = \sup\{\hat{U}_\alpha(\pi,t) : \pi \in \Pi_m\},$$

then $e^{-\alpha t}\hat{U}_\alpha^*(t) \in \Omega$, but $\hat{U}_\alpha^*(t)$ does not belong to Ω. In fact, if we define a set $\hat{\Omega} := \{e^{-\alpha t}x | x \in \Omega\}$, then $\hat{U}_\alpha^*(t)$ is the unique solution of the above equation in $\hat{\Omega}$.

Certainly, the discounted total reward is also a special case of the total reward through Eq. (4.36). Hence, for the nonstationary CTMDP model, the discounted criterion is equivalent to the total reward criterion.

At the end of this section, we consider a periodic case of the discounted criterion. For some $T \geq 0$, we say that the nonstationary CTMDP model has period T if $q_{ij}(t,a)$ and $\hat{r}_i(t,a)$ have period T; that is, $q_{ij}(t+T,a) = q_{ij}(t,a)$ and $\hat{r}_i(t+T,a) = \hat{r}_i(t,a)$ for all $t \geq 0$. A policy $\pi = (\pi_t)$ has period T if $\pi_{T+t} = \pi_t$ for all $t \geq 0$. The set of such policies is denoted by $\Pi_m(T)$. $\Pi_m^d(T)$ can be defined similarly.

Theorem 4.11: *Suppose that the nonstationary CTMDP model with the discounted criterion has period T. Then, $\hat{U}_\alpha^*(t)$ has also period T and $\hat{U}_\alpha^*(t) = \sup\{\hat{U}_\alpha(\pi,t) : \pi \in \Pi_m^d(T)\}$. Moreover, for any $\varepsilon > 0$, there exists an ε-optimal policy with period T.*

Proof: Due to Theorem 4.10, it suffices to prove that $\hat{U}_\alpha^*(t)$ has period T. From the given conditions,

$$\begin{aligned} & -\frac{d}{dt}\hat{U}_{\alpha,i}^*(t+T) \\ = & \sup_{a\in A(i)}\{\hat{r}_i(t+T,a) + \sum_j q_{ij}(t+T,a)\hat{U}_{\alpha,j}^*(t+T) - \alpha\hat{U}_{\alpha,i}^*(t+T)\} \\ = & \sup_{a\in A(i)}\{\hat{r}_i(t,a) + \sum_j q_{ij}(t,a)\hat{U}_{\alpha,j}^*(t+T) - \alpha\hat{U}_{\alpha,i}^*(t+T)\}, \text{ a.e. } t,i. \end{aligned}$$

It is clear that $\{\hat{U}_\alpha^*(t+T), t \geq 0\}$ is a solution of Eq. (4.37) in Ω, which implies together with the uniqueness of the solution that $\hat{U}_\alpha^*(t+T) = \hat{U}_\alpha^*(t)$ for $t \geq 0$. □

When $\hat{r}_i(t,a)$ and $q_{ij}(t,a)$ are independent of the time variable t; that is, $\hat{r}_i(t,a) = \hat{r}_i(a)$ and $q_{ij}(t,a) = q_{ij}(a)$, then the model is stationary. In this case, $\hat{U}_\alpha^*(t+T) = \hat{U}_\alpha^*(t)$ for all $t \geq 0$ and $T > 0$. This is to say that $\hat{U}_\alpha^*(t)$ is independent of $t \geq 0$. We denote it by $\hat{U}_\alpha^*$. Thus, the optimality equation (4.37) becomes

$$\alpha U_\alpha(i) = \sup_{a\in A(i)}\{\hat{r}_i(a) + \sum_j q_{ij}(a)U_\alpha(j)\}, \quad i \in S.$$

Another special case is the expected total reward in a finite horizon $[0,T]$. In fact, this happens when the reward rate becomes zero after T; that is, $r_i(t,a) = 0$ for all $t > T$ and $(i,a) \in \Gamma$.

3. A Stationary Model: Average Criterion

In the previous sections, we discussed the criterion of the total reward, which includes the discounted criterion. Now, in this section, we discuss the average criterion for a stationary CTMDP model. The method we used is to transform it into a DTMDP model.

The stationary CTMDP model we discuss in this section is as follows,

$$\{S, A(i), q_{ij}(a), r(i,a), U\}, \tag{4.39}$$

where the state space S is countable and the action set $A(i)$ is nonempty with a measurable structure $\mathcal{A}(i)$, the state transition rate family $q_{ij}(a)$ and the reward rate $r(i,a)$ are as those in Section 1 (we use the same symbols as in Section 1 hereafter). But we assume that both of them are uniformly bounded; that is,

$$\lambda := \sup_{i\in S, a\in A(i)} (-q_{ii}(a)) < \infty, \quad M := \sup_{i\in S, a\in A(i)} |r(i,a)| < \infty.$$

U is the average criterion defined by

$$U(\pi) = \liminf_{T\to\infty} \frac{1}{T} \int_0^T P(\pi, t) r(\pi, t) dt, \quad \pi \in \Pi_m.$$

Let $U^* = \sup\{U(\pi) | \pi \in \Pi_m\}$ be the optimal value. (ε-)optimal policies can be defined as usual.

First, we have the following result on the matrix $Q(\pi_0)$.

Lemma 4.12: *For any stochastic stationary policy $\pi_0 \in \Pi_s$ and positive discount rate $\alpha > 0$, the matrix $\alpha I - Q(\pi_0)$ is invertible and*

$$[\alpha I - Q(\pi_0)]^{-1} = \int_0^\infty e^{-\alpha t} P(\pi_0, t) dt. \tag{4.40}$$

Proof: When the reward rate $r(i,a)$ is uniformly bounded, it is easy to see from Theorem 4.9 and Eq. (4.37) that $U(\pi_0)$ is the unique bounded solution of the following equation,

$$[\alpha I - Q(\pi_0)]u = r(\pi_0).$$

So, the matrix $\alpha I - Q(\pi_0)$ is invertible and

$$U(\pi_0) = [\alpha I - Q(\pi_0)]^{-1} r(\pi_0) = \int_0^\infty e^{-\alpha t} P(\pi_0, t) r(\pi_0) dt,$$

which implies Eq. (4.40) due to the arbitrariness of $r(i,a)$. □

Before discussing the transformation for the average criterion, we first discuss a transformation for the discounted criterion. For this, suppose that the discount rate for the CTMDP model is $\alpha > 0$ and let $U_\alpha(\pi)$ be the corresponding criterion, which is defined in Section 1.

Then, we define the following DTMDP model with the discounted criterion,

$$\{S, A(i), p_{ij}(a), r'(i,a), V'_\beta\}, \tag{4.41}$$

where the state space S and the action set $A(i)$ are the same as in the CTMDP model (4.39), the state transition probability $p_{ij}(a)$ and the reward function $r'(i,a)$ are given, respectively, by

$$p_{ij}(a) = \lambda^{-1} q_{ij}(a) + \delta_{ij}, \;\; r'(i,a) = \frac{r(i,a)}{\lambda + \alpha}$$

and the discount factor is

$$\beta = \frac{\lambda}{\lambda + \alpha},$$

and let $V'_\beta(\pi)$ be the corresponding criterion, which is defined in Chapter 2.

Let $V'^*_\beta = \sup_\pi V'_\beta(\pi)$ be the optimal value. For the two models, we have the following result on their discounted objective functions, where we define matrices $Q(\pi_0) = (q_{ij}(\pi_0))$ and $P(\pi_0) = (p_{ij}(\pi_0))$ as usual

$$q_{ij}(\pi_0) = \sum_{a \in A(i)} q_{ij}(a)\pi_0(a|i), \;\; p_{ij}(\pi_0) = \sum_{a \in A(i)} p_{ij}(a)\pi_0(a|i), \;\; i,j \in S.$$

Lemma 4.13: *Suppose $\alpha > 0$. Then, $V'_\beta(\pi_0) = U_\alpha(\pi_0)$ for each stochastic stationary policy $\pi_0 \in \Pi_s$ and $V'^*_\beta = U^*_\alpha$.*

Proof: By the definition of $p_{ij}(a)$ we have $[\alpha I - Q(\pi_0)] = (\lambda+\alpha)[I - \beta P(\pi_0)]$. Due to Lemma 4.12, the matrix $I - \beta P(\pi_0)$ is invertible and

$$[\alpha I - Q(\pi_0)]^{-1} = \frac{1}{\lambda + \alpha}[I - \beta P(\pi_0)]^{-1}.$$

With this, $r(\pi_0) = (\lambda + \alpha) r'(\pi_0)$, and Lemma 4.12, we obtain that

$$\begin{aligned} U(\pi_0) &= \int_0^\infty e^{-\alpha t} P(\pi_0, t) dt \cdot r(\pi_0) \\ &= [\alpha I - Q(\pi_0)]^{-1} r(\pi_0) \\ &= \frac{1}{\lambda + \alpha}[I - \beta P(\pi_0)]^{-1} \cdot (\lambda + \alpha) r'(\pi_0) \\ &= [I - \beta P(\pi_0)]^{-1} r'(\pi_0) \\ &= V'_\beta(\pi_0). \end{aligned}$$

The discounted criterion optimality equation for the DTMDP model (4.41) is

$$V(i) = \sup_{a \in A(i)} \{r'(i,a) + \beta \sum_j p_{ij}(a) V(j)\}, \;\; i \in S,$$

and that for the CTMDP model of Eq. (4.39) is

$$\alpha U(i) = \sup_{a \in A(i)} \{r(i,a) + \beta \sum_j q_{ij}(a) U(j)\}, \;\; i \in S,$$

It is easy to see that these two equations are equivalent; that is, V is a solution of the former if and only if it is a solution of the latter. Hence, $V'^*_\beta = U^*_\alpha$. $\square$

Lemma 4.13 above says that both models have the same discounted criteria in the set of stochastic stationary policies, the same optimal values, and the same discounted criterion optimality equation.

Now, for the average criterion, we define the following DTMDP model,

$$\{S, A(i), p_{ij}(a), r(i,a), V\}, \tag{4.42}$$

where the state space S, the action set $A(i)$, and the reward function $r(i,a)$ are the same as in the CTMDP model (4.39), the state transition probability $p_{ij}(a)$ is given in the DTMDP model (4.41), and V is the average criterion, as defined in Chapter 3. We also let $V_\beta(\pi)$ be the discounted criterion with the discount factor $\beta = \lambda/(\lambda+\alpha)$.

From Chapter 3, we know that the optimality equation and the optimality inequalities for the average criterion of the DTMDP model of Eq. (4.42) are, respectively,

$$\rho + h(i) = \sup_{a\in A(i)} \{r(i,a) + \sum_j p_{ij}(a)h(j)\}, \quad i \in S, \tag{4.43}$$

$$\rho + h(i) \leq \sup_{a\in A(i)} \{r(i,a) + \sum_j p_{ij}(a)h(j)\}, \quad i \in S, \tag{4.44}$$

$$\rho + h(i) \leq r(i,f) + \sum_j p_{ij}(f)h(j), \quad i \in S. \tag{4.45}$$

The corresponding optimality equation and optimality inequalities for the average criterion of the CTMDP model of Eq. (4.39) are, respectively,

$$\rho = \sup_{a\in A(i)} \{r(i,a) + \sum_j q_{ij}(a)h(j)\}, \quad i \in S, \tag{4.46}$$

$$\rho \leq \sup_{a\in A(i)} \{r(i,a) + \sum_j q_{ij}(a)h(j)\}, \quad i \in S, \tag{4.47}$$

$$\rho \leq r(i,f) + \sum_j q_{ij}(f)h(j), \quad i \in S. \tag{4.48}$$

The following theorem shows the equivalence between the CTMDP model (4.39) and the DTMDP model (4.42) on the average criterion.

Theorem 4.12: *$U(\pi_0) = V(\pi_0)$ for each stochastic stationary policy $\pi_0 \in \Pi_s$, and for any constant ρ and a function h on S, (ρ, h) is a solution of the optimality equation (or the optimality inequalities) for the DTMDP model (4.42) if and only if it is a solution of the optimality equation (or the optimality inequalities) for the CTMDP model (4.39).*

Proof: It follows the definition of $p_{ij}(a)$ and Lemma 4.12 that for each $\pi_0 \in \Pi_s$,

$$V_\beta(\pi_0) = (\lambda + \alpha)V'_\beta(\pi_0) = (\lambda + \alpha)U_\alpha(\pi_0).$$

Then, from the Abel theorem (Lemma 3.4 in Chapter 3) we have

$$\begin{aligned} U(\pi_0) &= \lim_{\alpha \downarrow 0} \alpha U_\alpha(\pi_0) = \lim_{\alpha \downarrow 0} \frac{\alpha}{\lambda + \alpha}(\lambda + \alpha)U_\alpha(\pi_0) \\ &= \lim_{\beta \uparrow 0}(1 - \beta)V_\beta(\pi_0) = V(\pi_0). \end{aligned}$$

The latter results are easy to show by the definition of $p_{ij}(a)$. □

Based on Theorem 4.12, we can directly use the results and methods in DTMDPs to CTMDPs for the average criterion. For example, Conditions 3.4–3.7 for DTMDPs can be transformed into the following conditions for the CTMDP model (4.39).

Condition 4.6: *There exists a series of discount rates $\alpha(m) \downarrow 0$, a series of constants $\varepsilon(m) \downarrow 0$, and a series of decision functions f_m such that*

$$U_{\alpha(m)}(i) \leq r(i, f_m) + \alpha(m) \sum_j q_{ij}(f_m)U_{\alpha(m)}(j) + \varepsilon(m), \ \ i \in S.$$

Condition 4.6 is true when the discounted criterion optimality equation is true and the optimal value is finite. The second condition is given below.

Condition 4.7: *There exists a state $0 \in S$ such that $(1 - \alpha(m))U_{\alpha(m)}(0)$ is bounded for $m \geq 0$.*

By the above condition, $U_{\alpha(m)}(0)$ is finite and thus one can define the relative value function by

$$h_{\alpha(m)}(i) = U_{\alpha(m)}(i) - U_{\alpha(m)}(0), \ \ i \in S, m \geq 0.$$

Using the term $h_{\alpha(m)}(i)$, the inequality in Condition 4.6 can be rewritten as

$$\begin{aligned} & (1 - \alpha(m))U_{\alpha(m)}(0) + h_{\alpha(m)}(i) \\ \leq \ & r(i, f_m) + \alpha(m) \sum_j p_{ij}(f_m)h_{\alpha(m)}(j) + \varepsilon(m), \ \ i \in S. \end{aligned}$$

About the third condition, we introduce first the following condition for a vector U,

$$\limsup_{m \to \infty} \sum_{j \neq i} q_{ij}(f_m)U(j) \leq \sum_{j \neq i} \lim_{m \to \infty} q_{ij}(f_m)U(j) < +\infty, \ \ \forall i \in S. \quad (4.49)$$

Condition 4.8: *One of the following two holds.*

1. *There exist nonnegative functions $L(i)$ and $M(i)$ such that $M + L$ satisfies (4.49), and $-M(i) \le h_{\alpha(m)}(i) \le L(i)$ for all i and m.*

2. *There exists a nonnegative function $L(i)$ such that $h_{\alpha(m)}(i) \le L(i)$ for all i and m. Moreover, L satisfies (4.49), and there exists a subsequence of $\{m\}$ (which is assumed to be $\{m\}$ itself) and a decision function $f \in F$ such that*

$$\lim_{m\to\infty} r(i, f_m) = r(i, f), \quad \lim_{m\to\infty} p_{ij}(f_m) = p_{ij}(f), \qquad i, j \in S.$$

Let

$$h(i) = \limsup_{m\to\infty} h_{\alpha(m)}(i), \qquad i \in S.$$

Obviously, under Condition 4.8, $-\infty \le h(i) \le L(i)$ for $i \in S$. Moreover, if 1 in Condition 4.8 is true, then $-M(i) \le h(i) \le L(i)$ for $i \in S$ and by the diagonalization method we can assume that

$$h(i) = \lim_{m\to\infty} h_{\alpha(m)}(i), \ \ \forall i \in S.$$

About the stationary policy f introduced in the above condition, we need the following condition for it.

Condition 4.9: $\limsup_{t\to\infty} \frac{1}{t} P(f, t) \le 0$ *for all* $i \in S$.

The following theorem is a similar one to Theorem 3.6.

Theorem 4.13: *Provided that Conditions 4.6–4.9 hold.*

1. *ACOI (4.47) holds if its right-hand side is well defined, and then f is ε-optimal in S_h if f attains the ε-supremum of its right-hand side (for some $\varepsilon \ge 0$) and satisfies Condition 4.9.*

2. *If 1 of Condition 4.8 holds, and there exists a policy $f \in F$such that*

$$\begin{aligned} &\limsup_{m\to\infty} \{r(i, f_m) + \sum_j q_{ij}(f_m) h(j)\} \\ \le\ & r(i, f) + \sum_j q_{ij}(f) h(j), \qquad i \in S, \end{aligned}$$

and f satisfies Condition 4.9, then ACOI(0) (4.48) is true and f is optimal.

3. *If 2 of Condition 4.8 holds and f satisfies Condition 4.9, then ACOI(0) (4.48) is true with $\rho = U^*(i)$ for $i \in S_h$, and f is optimal in S_h.*

The results in Section 3.3 in Chapter 3 are all true for the CTMDP model (4.39). The details are omitted here.

At the end of this section, we pointed out that the transformation may still be true when the state transition rate family is not uniformly bounded. Hu and Wang [72] showed all the results in this section under the condition that $\lambda(i) := \sup_{a\in A(i)}\{-q_{ii}(a)\}$ is finite for each $i \in S$, for the discounted criterion and the total reward criterion with positive or negative rewards.

4. Notes and References

Continuous time MDPs were proposed in Howard [55] but they were first formally studied in [82] and [83]. They were also been studied well for discounted criterion with bounded rewards. In [126], Serfozo studied a CTMDP model by a transformation but for only the discounted criterion function among stationary policies. All of them are for bounded transition rates. In [129], Song studied a CTMDP model with unbounded transition rate by using the usual method. In [57], Hu studied a CTMDP model for the discounted criterion also with unbounded transition rates but by using a transformation method, which transforms the CTMDP model into a DTMDP model for the discounted criterion. Under this transformation, the corresponding optimality equation and the discounted criteria among stationary policies in the CTMDP model and those in the DTMDP model are equivalent. So the results for CTMDP can be obtained directly from those for DTMDP. On the other hand, Hou in [53] presented a set of conditions for a discounted CTMDP model with unbounded reward rate. Her conditions are, in fact, a generalization of that of Lippman's [92] for a SMDP model. Recently, Guo and Zhu [47] discussed a denumerable state CTMDP model with the discounted criterion by presenting a set of conditions for the unbounded transition rate and the unbounded reward rate. They illustrated that their condition is weaker than that in the literature. But the method they used is a combination of that of Hu [57] and Hou [53]. In [48], Guo and Zhu discussed the same model and the same condition but on the average criterion, and the standard results are obtained. All the above references are for the countable state case. Doshi [29] studied the arbitrary state space case with the discounted criterion.

But there are few studies on the total reward criterion. The method presented in Hu [57] was used in Hu and Wang [73] to study the nonpositive or nonnegative rewards, but it is restricted to the stationary policies and the method cannot deal with the general reward rate function or the negative discount rate.

There are few discussions concerning nonstationary CTMDPs. Martin-Lof [94] investigated a model in a finite horizon, in which the state space is finite, and the transition rates $q_{ij}(t, a)$ and reward rate functions $r_i(t, a)$ are periodic. Under the assumption that $q_{ij}(t, a)$ and $r_i(t, a)$ are continuous in (t, a) and satisfy a Lipschitz condition, the author proved that there exists an optimal periodic policy.

We first tried to see what results can be obtained under the necessary condition that the CTMDP model with the total reward is well defined in the first section. The method is similar to that in Chapter 2 for the DTMDPs. Then, we studied a nonstationary CTMDP model with the total reward in a general framework. Finally, we studied a stationary CTMDP model with the average criterion by a transformation. Section 1 is from Hu et al [71], Section 2 is from Hu [60] and Hu [66], and Section 3 is from Hu [57] and Hu and Wang [73].

Problems

1. Optimal Service Control of Queueing System $M/M/1$. Consider a queueing system $M/M/1$ in which the customers arrive according to a Poisson process with the arrival rate λ. The arrived customers are served one after another and so they may wait for the service when the server is busy for service. The service times for customers are random with the identical exponential distribution function with the mean of $1/\mu$. We call μ the service rate, which is chosen from a closed interval $[\underline{\mu}, \overline{\mu}]$ at any time. When the service rate μ is chosen then a cost per time unit is $c(\mu)$. On the other hand, the holding cost per time unit $h(i)$ is incurred when there are i customers in the queue (including the one served). Assume that both $c(\mu)$ and $h(i)$ are increasing and concave.

When a larger service rate μ is chosen, then more service cost incurs but more customers are served per time unit and so lower holding cost incurs, while a smaller service rate is chosen then less service cost incurs but less customers are served per time unit and so higher holding cost incurs. Hence, the manager should choose a suitable service rate to minimize the discounted total expected cost or the average expected cost. Set them up as Markov decision processes models and write the optimality equations.

2. Optimal Advertisement. There is an asset to be sold. The owner can make advertisement to attract buyers. Suppose that if the owner invest x per time unit in the advertisement then the buyers arrive according to a Poisson process with the arrival rate $\lambda(x)$, which is increasing and concave in x. The advertisement investment rate can be chosen continuously. Arriving customers have reserves which are independent and identically distributed random variables. Any customer will buy the asset if and only if his reserve is larger than or equals the price p, which is determined initially. Set this problem up as a Markov decision process model to maximize his expected total profit. Based on this, solve further the optimal price p.

3. Optimal Advertisement/Pricing. For the problem above, suppose that the price for the asset can be adjusted continuously. Then both the advertisement investment rate and the price need to be chosen continuously. Which is the Markov decision process model for this case?

4. Revenue Management. Consider a flight with N seats (Note: N is fixed) to be sold within a given time horizon T. The customers arrive according to a Poisson process with the arrival rate λ. They have reserves which are independent and identically distributed random variables. Any customer will buy a seat if and only if his reserve is larger than or equals the price p, which is determined continuously. It should be noted that for this problem the remaining seats after the flight values nothing. Set this problem up as a Markov decision process model to maximize the expected total revenue.

How is the problem when the Poisson process is nonhomogeneous, i.e., the arrival rate $\lambda(t)$ depends on time t?

5. For the stationary MDP model discussed in Section 3, the average criterion is transformed into the discounted criterion. How about this transformation if the state transition rate family is not uniformly bounded, for example, if $\lambda(i) = \sup_{a \in A(i)}(-q_{ii}(a))$ is finite for each $i \in S$ but $\lambda(i)$ may be unbounded. Please study this problem.

Chapter 5

SEMI-MARKOV DECISION PROCESSES

The underlying stochastic processes in DTMDPs are discrete time Markov chains, where the decision epochs are equally periodic or the length of adjacent decision epochs are not considered. Those in CTMDPs are continuous time Markov chains, where the decision is chosen every time. In this chapter, we study a stationary semi-Markov decision processes (SMDPs) model, where the underlying stochastic processes are semi-Markov processes. Here, the decision epoch is exactly the state transition epoch with its length being random. We transform the SMDP model into a stationary DTMDP model for either the total reward criterion or the average criterion, similarly to the stationary CTMDP model with the average criterion discussed in Section 4.3. Thus, the results in DTMDP can be used directly for SMDP for the discounted criterion, the total reward criterion, and the average criterion.

1. Model and Conditions

1.1 Model

The semi-Markov decision process model we discuss in this chapter is given by

$$\{S, (A(i), \mathcal{A}(i)), p_{ij}(a), T(\cdot|i, a, j), r(u, i, a, j, t)\}. \tag{5.1}$$

The meaning of the model above is as follows. The system's state is periodic observable and an action will be chosen when the state is observed. The state space S is countable and the action set $A(i)$, available at state i, is nonempty with a measurable structure $\mathcal{A}(i)$, as that in Chapter 2. When the system enters some state i at some decision epoch, the decision maker chooses an action a from the available action set $A(i)$. Then, the following three things happen.

1. The system will enter state j with probability $p_{ij}(a)$ at the next decision epoch.

2. The duration time of the system at state i, before entering state j, is a random variable with distribution function $T(\cdot|i,a,j)$.

3. The system will receive a reward $r(u,i,a,j,t)$ in the time interval $[0,u]$ for $u \le t$ if the next state is j and the duration time of the system at state i is t.

The system proceeds in the above way repeatedly. The criteria include the discounted criterion, the total reward criterion, and the average criterion, which are defined later.

For the mathematical requirement, it is assumed that the reward function $r(u,i,a,j,t)$ is a bounded variation in u and is Lebesgue measurable in t. A usual and general form of the reward function is

$$r(u,i,a,j,t) = r_1(i,a,j) + \delta_t(u) r_2(i,a,j) + r_3(i,a,j)u, \tag{5.2}$$

where $\delta_t(u) = 1$ for $u = t$ and $\delta_t(u) = 0$ otherwise. The three terms r_i in the right-hand side of the above equation represent different types of reward functions: (a) $r_1(i,a,j)$ is an instantaneous reward received when the system enters state i and action a is chosen (certainly it is random because it depends on the next state j which is random when the system enters i). (b) $r_2(i,a,j)$ is an instantaneous reward received before the system enters the next state j. (c) $r_3(i,a,j)$ is the reward rate received from the system during the period. In practical systems, one or two types of the reward functions my be included.

It is easy to see that when the duration time is a constant, that is, $T(\cdot|i,a,j)$ is degenerative at the same point irrespectively of i,a,j, then the SMDP model becomes a DTMDP model. So, DTMDPs are special cases of SMDPs. On the other hand, we show in this chapter that SMDPs can be transformed into DTMDPs. In fact, the transformation is the key to the methodology in this chapter.

Before defining policies, we define histories. Because the duration time is introduced, a history consists of states, actions, and duration times. Its form is given by

$$h_n = (i_0, a_0, t_0, i_1, a_1, t_1, \ldots, i_{n-1}, a_{n-1}, t_{n-1}, i_n), \quad n \ge 0,$$

where i_k, a_k, t_k are, respectively, the state, the action chosen, and the duration at the state i_k of the system after its kth transition, for $k = 0, 1, \ldots, n-1$, and i_n is the current state of the system after its nth transition. h_n is called a history up to n (i.e., the nth decision epoch), the set of which is denoted by H_n. With the sets $\Gamma = \{(i,a) | a \in A(i), i \in S\}$ and $E = [0,\infty)$, H_n can be rewritten as $H_n = (\Gamma \times E)^n \times S$ for $n \ge 0$.

Similar to that in DTMDPs, a policy is a sequence $\pi = (\pi_0, \pi_1, \ldots)$ such that an action should be chosen according to a probability distribution $\pi_n(\cdot|h_n)$ on $A(i_n)$ whenever a history $h_n \in H_n$ has happened. It is assumed that $\pi_n(\cdot|h_n)$ is measurable in elements a_k and t_k in the history $h_n = (i_0, a_0, t_0, i_1, a_1, t_1,$

$\ldots, i_n)$. The set of all policies is denoted by $\Pi(s)$. Here, "(s)" is used to differentiate the policy set Π for DTMDPs.

For a policy $\pi \in \Pi(s)$, if $\pi_n(\cdot|h_n)$ is irrespective of the time variables $t_0, t_1, \ldots, t_{n-1}$ in the history h_n for each $h_n \in H_n$ and $n \geq 0$, then π has the same form of policies for DTMDPs. The set of these policies is denoted by Π.

A policy $\pi \in \Pi$ is said to be semi-Markov if $\pi_n(\cdot|h_n) = \pi_n(\cdot|i_0, i_n)$ depends only on the initial state i_0 and the current state i_n (and is irrespective of all other elements $a_0, t_0, i_1, \ldots, t_{n-1}$) in h_n for each history $h_n \in H_n$ and $n \geq 0$. Let Π_{sm} be the set of all semi-Markov policies. Moreover, if $\pi_n(\cdot|h_n) = \pi_n(\cdot|i_n)$ for all h_n and n, then π is said to be a Markov policy, the set of which is denoted by Π_m.

Other types of policies, such as stochastic stationary policies ($\pi_0 \in \Pi_s$), deterministic Markov policies ($\pi = (f_0, f_1, \ldots) \in \Pi_m^d$), and stationary policies ($f \in F$) are exactly those in DTMDPs (see Chapter 2).

For $n \geq 0$, we let X_n, Δ_n, and t_n be, respectively, state, action chosen, and duration after the nth state transition. Let

$$T_0 = 0, \;\; T_{n+1} = T_n + t_n, \;\; n \geq 0.$$

Obviously, T_n is the nth decision epoch. We often say that $[T_n, T_{n+1})$ is the nth period and T_n the beginning time of the nth period.

For any given policy $\pi \in \Pi(s)$, let $\mathcal{L}(\pi) = (X_0, \Delta_0, t_0, X_1, \Delta_1, t_1, X_2, \Delta_2, t_2, \ldots)$ be a stochastic sequence to represent states, actions, and duration times of the process under policy π. It proceeds as follows.

a. At the nth decision epoch T_n, we have a history $h_n = (i_0, a_0, t_0, \ldots, i_{n-1}, a_{n-1}, t_{n-1}, i_n)$ with the current state i_n.

b. Then, an action a_n is chosen from $A(i_n)$ according to the probability $\pi_n(\cdot|h_n)$. Let $\Delta_n = a_n$.

c. The system will enter state i_{n+1} at the next decision epoch (i.e., $X_{n+1} = i_{n+1}$) according to a probability $p_{i_n, i_{n+1}}(a_n)$ for $i_{n+1} \in S$.

d. The system will stay at i_n for a duration, denoted by t_n, according to the probability distribution $T(\cdot|i_n, a_n, i_{n+1})$.

e. In this time, a new history $h_{n+1} = (h_n, a_n, t_n, i_{n+1})$ occurs and the system repeats.

Surely, under a Markov policy $\pi \in \Pi_m$, the sequence $\mathcal{L}(\pi)$ is a Markov chain, and may be nonstationary, and under a stochastic stationary policy $\pi_0 \in \Pi_s$, the sequence $\mathcal{L}(\pi)$ is a stationary Markov chain.

1.2 Regular Conditions

As in the semi-Markov processes, we need the following regular condition.

Condition 5.1 (Regular Condition 1): *There exist constants* $\theta \in (0,1)$ *and* $\delta > 0$ *such that*

$$\sum_j p_{ij}(a)T(\delta|i,a,j) \le 1-\theta, \quad \forall (i,a) \in \Gamma.$$

For $(i,a) \in \Gamma$ we let

$$T(t|i,a) = \sum_j p_{ij}(a)T(t|i,a,j)$$

be the distribution function of the duration time at state i if action a is chosen. Then, the Regular Condition 1 says that $T(\delta|i,a) \le 1-\theta$ for all $(i,a) \in \Gamma$. That is, the probability that the duration at any state does not exceed δ is at most $1-\theta$. From the theory of semi-Markov processes, we know that under any policy π, the process $\mathcal{L}(\pi)$ is regular; that is, with probability one there occur only finite state transitions in every finite time interval.

For any $\alpha \ge 0$, we let for $(i,a) \in \Gamma$ and $j \in S$,

$$\begin{aligned}
\beta_\alpha(i,a,j) &= \int_0^\infty e^{-\alpha t}T(dt|i,a,j), \\
\beta_\alpha(i,a) &= \sum_j p_{ij}(a)\beta_\alpha(i,a,j) = \int_0^\infty e^{-\alpha t}T(dt|i,a), \\
\beta_\alpha &= \sup_{i,a} \beta_\alpha(i,a).
\end{aligned}$$

We call $\beta_\alpha(i,a)$ the expected discount factor at (i,a) because one unit reward at T_{n+1} is value $\beta_\alpha(i,a)$ at T_n if the state and the action at T_n are, respectively, i and a when the continuous discount rate is α. It is apparent that $\beta_\alpha \le 1$. But from Chapter 2 we know that $\beta_\alpha < 1$ will result in better results. Thus, we present the following condition, which is also a regular condition.

Condition 5.2 (Regular Condition 2): *There exists* $\alpha > 0$ *such that* $\beta_\alpha < 1$.

The condition above requires that the expected discount factor $\beta_\alpha(i,a)$ be uniformly less than one. Then, we can expect better results for the SMDP model.

The following lemma says that the two regular conditions are equivalent.

Lemma 5.1: *The two regular conditions are equivalent to each other. Therefore, when one of them is true, then* $\beta_\alpha < 1$ *for all* $\alpha > 0$.

Proof: Suppose that Regular Condition 1 is true. Then, for any given $\alpha > 0$,

$$\beta_\alpha(i,a) = \sum_j p_{ij}(a)\{\int_0^\delta e^{-\alpha t}T(dt|i,a,j) + \int_\delta^\infty e^{-\alpha t}T(dt|i,a,j)\}$$

$$\begin{aligned}
&\leq \sum_j p_{ij}(a)\{\int_0^\delta T(dt|i,a,j) + \int_\delta^\infty e^{-\alpha t} T(dt|i,a,j)\} \\
&= \sum_j p_{ij}(a)\{T(\delta|i,a,j) + e^{-\alpha\delta}[1 - T(\delta|i,a,j)]\} \\
&\leq e^{-\alpha\delta} + (1 - e^{-\alpha\delta}) \sum_j p_{ij}(a) T(\delta|i,a,j) \\
&\leq e^{-\alpha\delta} + (1 - e^{-\alpha\delta})(1 - \theta) \\
&:= \beta_\alpha^* < 1.
\end{aligned}$$

On the other hand, suppose that only Regular Condition 2 is true. If Regular Condition 1 is not true then for any constants $\delta > 0$ and $\theta > 0$, there is $(i,a) \in \Gamma$ such that $\sum_j p_{ij}(a)T(\delta|i,a,j) \geq 1-\theta$. By noting that $\sum_j p_{ij}(a)T(\delta|i,a,j) \leq 1$ for each $\delta > 0$, we know from the arbitrariness of θ that

$$\sup_{i,a} \sum_j p_{ij}(a) T(\delta|i,a,j) = 1, \ \ \forall \delta > 0.$$

But for any $(i,a) \in \Gamma$ and $\alpha > 0$,

$$\begin{aligned}
\beta_\alpha(i,a) &\geq \sum_j p_{ij}(a) \int_0^\delta e^{-\alpha t} T(dt|i,a,j) \\
&\geq e^{-\alpha\delta} \sum_j p_{ij}(a) T(\delta|i,a,j).
\end{aligned}$$

Hence,

$$\begin{aligned}
\sup_{i,a} \beta_\alpha(i,a) &\geq e^{-\alpha\delta} \sup_{i,a} \sum_j p_{ij}(a) T(\delta|i,a,j) \\
&= e^{-\alpha\delta} \to 1, \ \text{ as } \delta \to 0^+,
\end{aligned}$$

which contradicts Regular Condition 2. □

In the above, it has been shown that when one regular condition is true, $\beta_\alpha < 1$ for all $\alpha > 0$.

With the above lemma, we say the regular condition is true whenever either Regular Condition 1 or 2 is true. The regular condition is assumed throughout this chapter.

Similarly to $\beta_\alpha(i,a)$, we let for $(i,a) \in \Gamma$,

$$\tau(i,a) = \sum_j p_{ij}(a) \int_0^\infty t T(dt|i,a,j) = \int_0^\infty t T(dt|i,a)$$

be the expected duration at state i if action a is chosen. The following lemma says that the expected duration has a positive lower bound.

Lemma 5.2: $\inf_{i,a} \tau(i,a) \geq \delta\theta > 0$.

Proof: From Regular Condition 1,

$$\begin{aligned}\tau(i,a) &= \int_0^\infty tT(dt|i,a) \geq \int_\delta^\infty tT(dt|i,a) \\ &\geq \delta \int_\delta^\infty T(dt|i,a) \\ &= \delta[1 - T(\delta|i,a)] \geq \delta\theta > 0.\end{aligned}$$

This completes the proof. □

This lemma is useful when we transform the SMDP model into a DTMDP model in the following sections. Obviously, the lemma above is ensured by regular conditions.

1.3 Criteria

We define the criteria in this subsection for both the total reward criterion and the average criterion.

We first consider the total reward criterion with the discount rate $\alpha \geq 0$. The meaning of α is the same as that in CTMDPs (see Eq. (4.1)). But here $\alpha \geq 0$ is assumed. When $\alpha = 0$ there is no discounting and the criterion is the total reward, whereas when $\alpha > 0$ the criterion is the discounted total reward. We call them uniformly the total reward criterion.

We define

$$r_\alpha(i,a,j,t) = \int_0^t e^{-\alpha t} d_u r(u,i,a,j,t), \quad (i,a) \in \Gamma,\ j \in S,\ t \geq 0.$$

Because $r(u,i,a,j,t)$ is bounded variation in u, $r_\alpha(i,a,j,t)$ is well defined. It is the discounted reward received from the system when it enters state i if action a is chosen, the next state is j, and the duration time at state i is t. For simplicity, we write $r(i,a,j,t) = r_0(i,a,j,t)$ when there is no discounting. It should be noted that $r(i,a,j,t)$ and $r(u,i,a,j,t)$ have different meanings. When the reward function is given by Eq. (5.2), then

$$r_\alpha(i,a,j,t) = r_1(i,a,j) + r_2(i,a,j)e^{-\alpha t} + r_3(i,a,j)\frac{1}{\alpha}(1 - e^{-\alpha t}).$$

For $n \geq 0$, $r_\alpha(X_n, \Delta_n, X_{n+1}, t_n)$ is the reward, discounted to T_n, received from the system in the nth period. Furthermore, this reward discounted to the initial time 0 is $e^{-\alpha T_n} r_\alpha(X_n, \Delta_n, X_{n+1}, t_n)$. This is a random variable depending on the policy π used and the initial state i. Its expected value is $E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n, X_{n+1}, t_n)\}$. Thus, we define

$$V_{\alpha,N}(\pi,i) = \sum_{n=0}^{N-1} E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n, X_{n+1}, t_n)\} \tag{5.3}$$

to be the expected discounted total reward in N horizons under policy π from the initial state $i \in S$. We further define

$$V_\alpha(\pi, i) = \sum_{n=0}^{\infty} E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n, X_{n+1}, t_n)\} \tag{5.4}$$

to be the expected discounted total reward (in infinite horizons) under policy π from the initial state $i \in S$.

Moreover, we write $V_N(\pi, i) = V_{0,N}(\pi, i)$ when there is no discounting, and define

$$V(\pi, i) = \liminf_{N\to\infty} \frac{V_N(\pi, i)}{E_{\pi,i} T_N} \tag{5.5}$$

to be the average reward per unit time in infinite horizons under policy π from the initial state $i \in S$.

Regular Conditions ensure that T_n tends to infinity when n tends to infinity. Then, the criteria $V_\alpha(\pi, i)$ and $V(\pi, i)$ defined above are really on the whole time axis $[0, \infty)$.

Let

$$V_\alpha^*(i) = \sup_\pi V_\alpha(\pi, i), \;\; V^*(i) = \sup_\pi V(\pi, i), \;\; i \in S$$

be the optimal values for the total reward criterion and the average criterion, respectively. (ε-)optimal policies for both criteria can be defined as in Chapters 2 and 3 for DTMDP models.

2. Transformation

For $(i, a) \in \Gamma$, we define

$$r_\alpha(i, a) = \sum_j p_{ij}(a) \int_0^\infty r_\alpha(i, a, j, t) T(dt|i, a, j)$$

to be the expected discounted reward received from the system at the beginning of a period when the state i is entered and action a is chosen. We write $r(i, a) = r_0(i, a)$ for $(i, a) \in \Gamma$. $r_\alpha(i, a)$ is similar to $\beta_\alpha(i, a)$ and $\tau(i, a)$ introduced previously.

Throughout this section, it is assumed that $r_\alpha(i, a)$ is well defined although it may be infinite for the considered discount rate $\alpha \geq 0$. We show in the following two subsections that the general reward form $r(u, i, a, j, t)$ can be simplified into the form $r_\alpha(i, a)$. Similarly, the duration time can also be simplified to $\tau(i, a)$. In the following, we transform the SMDP model (5.1) into DTMDP models for the total reward criterion and the average criterion, respectively.

2.1 Total Reward

In this subsection, we discuss the transformation for the total reward. The discount rate $\alpha \geq 0$ is fixed. First, we show that the reward function can be simplified as $r_\alpha(i,a)$.

Lemma 5.3: *For any policy $\pi \in \Pi(s)$, $n \geq 0$, and $\alpha \geq 0$, we have*

$$E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n, X_{n+1}, t_n)\} = E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n)\},\ i \in S. \quad (5.6)$$

This means that if either side of the above equation is well defined then the other side is also well defined and both of them are equal to each other. Therefore,

$$\begin{aligned} V_\alpha(\pi, i) &= \sum_{n=0}^{\infty} E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n)\}, \quad i \in S, \\ V_{\alpha,N}(\pi, i) &= \sum_{n=0}^{N-1} E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n)\}, \quad i \in S \end{aligned}$$

for each $N \geq 1$.

Proof: When one side of Eq. (5.6) is well defined, we have

$$\begin{aligned} & E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n, X_{n+1}, t_n)\} \\ = & \sum_{j \in S} P_{\pi,i}\{X_n = j\} \int_{A(j)} P_{\pi,i}\{\Delta_n \in da | X_n = j\} \\ & \cdot \sum_{j'} p_{jj'}(a) \int_0^\infty T(dt|i,a,j) E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(j,a,j',t)\} \\ = & \sum_{j \in S} P_{\pi,i}\{X_n = j\} \int_{A(j)} P_{\pi,i}\{\Delta_n \in da | X_n = j\} E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(j,a)\} \\ = & E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n)\}. \end{aligned}$$

This completes the proof. □

Based on the lemma above, we define the following DTMDP model,

$$\{S, A(i), \bar{p}_{ij}(a), r_\alpha(i,a), V_\beta^{(D)}\}, \quad (5.7)$$

where the state transition probability is given by

$$\bar{p}_{ij}(a) = \beta_\alpha^{-1} \beta_\alpha(i,a,j) p_{ij}(a).$$

The discount factor is $\beta = \beta_\alpha$ and all other elements, S, $A(i)$, and $r_\alpha(i,a)$, are the same as those for the SMDP model (5.1). Let $V_\beta^{(D)}(\pi)$ and $V_{\beta,N}^{(D)}(\pi)$

be, respectively, the expected discounted total reward in infinite horizons and N horizons under policy π for the DTMDP model above.

The following theorem establishes the equivalence between the DTMDP model (5.7) with the original SMDP model (5.1) in policy set Π.

Theorem 5.1: *For each policy $\pi \in \Pi$, $V_\beta^{(D)}(\pi) = V_\alpha(\pi)$ and $V_{N,\beta}^{(D)}(\pi) = V_{\alpha,N}(\pi)$. So, the DTMDP model (5.7) and the SMDP model (5.1) are equivalent for the total reward criterion over finite or infinite horizons in policy set Π.*

Proof: For each policy $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$, $i \in S$ and $n \geq 0$,

$$
\begin{aligned}
& E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n)\} \\
= & \int_{A(i)} \pi_0(da_0|i) \sum_{i_1} p_{i,i_1}(a_0) \int_0^\infty e^{-\alpha t_0} T(dt_0|i, a_0, i_1) \cdots \\
& \cdot \sum_{i_n} p_{i_{n-1},i_n}(a_{n-1}) \int_0^\infty e^{-\alpha t_{n-1}} T(dt_{n-1}|i_{n-1}, a_{n-1}, i_n) \\
& \cdot \int_{A(i_n)} \pi_n(da_n|i, a_0, i_1, \ldots, i_n) r_\alpha(i_n, a_n) \\
= & \int_{A(i)} \pi_0(da_0|i) \sum_{i_1} p_{i,i_1}(a_0) \beta_\alpha(i, a_0, i_1) \cdots \sum_{i_n} p_{i_{n-1},i_n}(a_{n-1}) \\
& \cdot \beta_\alpha(i_{n-1}, a_{n-1}, i_n) \int_{A(i_n)} \pi_n(da_n|i, a_0, i_1, \ldots, i_n) r_\alpha(i_n, a_n) \\
= & \int_{A(i)} \pi_0(da_0|i) \sum_{i_1} \beta_\alpha \bar{p}_{i,i_1}(a_0) \cdots \sum_{i_n} \beta_\alpha \bar{p}_{i_{n-1},i_n}(a_{n-1}) \\
& \cdot \int_{A(i_n)} \pi_n(da_n|i, a_0, i_1, \ldots, i_n) r_\alpha(i_n, a_n) \\
= & \beta_\alpha^n E_{\pi,i}^{(D)} r_\alpha(X_n, \Delta_n),
\end{aligned}
$$

where $E_{\pi,i}^{(D)}$ represents the expectation in the DTMDP model. Hence, the theorem is true. □

Each of the two equalities in Theorem 5.1 above means that if either side is well defined then the other side is also well defined and both sides are equal. With this theorem, we can directly generalize all results in Chapter 2 to SMDP model (5.1) for the discounted criterion. For example, Conditions 2.1 and 2.2 in Chapter 2 can be rewritten, respectively, as the following.

Condition 5.3: *$V_\alpha(\pi, i)$ is well defined for each policy π and state i.*

Condition 5.4: *For each policy π and state i, we have*

$$V_\alpha(\pi, i) = \int_{A(i)} \pi_0(da|i)\{r_\alpha(i,a) + \sum_j \beta_\alpha(i,a,j)p_{ij}(a)V_\alpha(\pi^{i,a}, j)\}.$$

Let $S_\infty, S_{-\infty}, S_0$ be the state subsets of, respectively, positive infinite, negative infinite, and finite optimal values, the same as those in Chapter 2. Then, based on the equivalence established in Theorem 5.1, all results in DTMDPs can be directly used for SMDPs for the total reward criterion. For example, the following theorem can be obtained from Theorems 2.2 and 2.3 and Corollary 2.1.

Theorem 5.2: *Suppose Conditions 5.3 and 5.4.*

1. *If $\sum_j \beta_\alpha(i,a,j)p_{ij}(a)V_\alpha^*(j)$ is well defined for each $(i,a) \in \Gamma$, then V_α^* satisfies the following optimality equation,*

$$V_\alpha(i) = \sup_{a \in A(i)} \{r_\alpha(i,a) + \sum_j \beta_\alpha(i,a,j)p_{ij}(a)V_\alpha(j)\}, \quad i \in S. \tag{5.8}$$

2. *For each $i \in S_0$, the action set $A(i)$ can be sized down to*

$$A_3(i) = \{a \in A(i) \quad | \quad p_{i,S_{-\infty}}(a) = 0 \text{ and } \sum_{j \in S_0} \beta_\alpha(i,a,j)p_{ij}(a)V_\alpha(j) > -\infty\}.$$

After the reduction, $S_\infty^* = S_\infty, S_{-\infty}^* = S_{-\infty}$, and S_0 is closed and the condition presented in 1 above is satisfied in S_0.

Now, we discuss that $V_\alpha^*(j)$ is also optimal in the larger policy set $\Pi(s)$ in the following corollary.

Corollary 5.1: *Suppose that V_α^* satisfies the optimality equation (5.8) and for each policy $\pi \in \Pi(s) - \Pi$,*

$$\limsup_{n \to \infty} E_{\pi,i}\{e^{-\alpha T_n} V_\alpha^*(X_n)\} \geq 0, \quad i \in S. \tag{5.9}$$

Then, $V_\alpha^*(i) = \sup\{V_\alpha(\pi, i) | \pi \in \Pi(s)\}$ for $i \in S$.

Proof: From the optimality equation (5.8),

$$V_\alpha(i) \geq r_\alpha(i,a) + \sum_j \beta_\alpha(i,a,j)p_{ij}(a)V_\alpha(j), \quad \forall (i,a) \in \Gamma.$$

Then,

$$V_\alpha(X_n) \geq r_\alpha(X_n, \Delta_n) + E\{e^{-\alpha t_n} V_\alpha(X_{n+1}) | X_n, \Delta_n\}, \quad n \geq 0.$$

For any policy $\pi \in \Pi(s) - \Pi$ and state $i \in S$, multiplying the above inequality by $e^{-\alpha T_n}$ and taking expectation under $E_{\pi,i}$ results in

$$\begin{aligned} E_{\pi,i}\{e^{-\alpha T_n} V_\alpha(X_n)\} &\geq E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n)\} \\ &\quad + E_{\pi,i}\{e^{-\alpha T_{n+1}} V_\alpha(X_{n+1})\}, \ \ n \geq 0. \end{aligned}$$

This implies that

$$V_\alpha(i) \geq \sum_{n=0}^{N-1} E_{\pi,i}\{e^{-\alpha T_n} r_\alpha(X_n, \Delta_n)\} + E_{\pi,i}\{e^{-\alpha T_N} V_\alpha(X_N)\}$$

for $N \geq 1$. Letting $N \to \infty$, we know from the given condition that $V_\alpha(i) \geq V_\alpha(\pi, i)$. □

The condition given in Eq. (5.9) appeared in the previous chapters for both DTMDPs and CTMDPs. It is true if the optimal value V_α^* is nonnegative (especially when the rewards are nonnegative), or V_α^* is bounded below and the discount factor α is positive.

2.2 Average Criterion

About the transformation for the average criterion, we first simplify the reward function and the distribution function of the duration times.

Lemma 5.4: *For any policy* $\pi \in \Pi$

$$V(\pi, i) \ = \ \liminf_{N \to \infty} \frac{E_{\pi,i}\{\sum_{n=0}^{N} r(X_n, \Delta_n)\}}{E_{\pi,i}\{\sum_{n=0}^{N} \tau(X_n, \Delta_n)\}}, \ \ i \in S.$$

Proof: We have for $\pi \in \Pi$

$$\begin{aligned} E_{\pi,i} t_n &= \sum_{j \in S} P_{\pi,i}\{X_n = j\} \int_{A(j)} P_{\pi,i}\{\Delta_n \in da |\ X_n = j\} \\ &\quad \cdot \sum_{j'} p_{jj'}(a) \int_0^\infty t T(dt|j, a, j') \\ &= \sum_{j \in S} P_{\pi,i}\{X_n = j\} \int_{A(j)} P_{\pi,i}\{\Delta_n \in da |\ X_n = j\} \tau(j, a) \\ &= E_{\pi,i} \tau(X_n, \Delta_n), \ \ i \in S. \end{aligned}$$

This together with Eq. (5.6) for $\alpha = 0$ implies the lemma. □

The above lemma says that for the average criterion, the reward function and the duration time can be simplified to $r(i, a)$ and $\tau(i, a)$, respectively. That is,

the original SMDP model (5.1) can be transformed into the following SMDP model for the average criterion,

$$\{S, A(i), p_{ij}(a), \tau(i,a), r(i,a), V\}, \tag{5.10}$$

where the duration time, at state i when action a is chosen, is a deterministic number $\tau(i,a)$, not a random variable as in the original SMDP model. The above model is, in fact, a type of DTMDP models.

By noting that $r(i,a)$ is the reward from the system received in a time horizon with length $\tau(i,a)$, we say that the average reward per time is $r(i,a)/\ \tau(i,a)$. With this in mind, we define the following DTMDP model,

$$\{S, A(i), \hat{p}_{ij}(a), \hat{r}(i,a), V\}, \tag{5.11}$$

where the state transition probability and the reward function are, respectively,

$$\hat{p}_{ij}(a) = \frac{\tau_*[p_{ij}(a) - \delta_{ij}]}{\tau(i,a) + \delta_{ij}}, \ \ \hat{r}(i,a) = \frac{r(i,a)}{\tau(i,a)},$$

and τ_* is a constant satisfying the following condition,

$$0 < \tau_* < \tau_0 := \inf\{\tau(i,a)[1 - p_{ii}(a)]^{-1} \mid p_{ii}(a) < 1, (i,a) \in \Gamma\}.$$

From Lemma 5.2, $\tau_0 \geq \delta\theta$. So, the above τ_* exists and $\hat{r}(i,a)$ is bounded when $r(i,a)$ is bounded. It should be noted that the DTMDP model (5.11) differs from that given in Eq. (5.7) in the last subsection for the total reward criterion.

In Chapter 3, we discussed the optimality equation and optimality inequalities for DTMDP with the average criterion. They are rewritten for the DTMDP model (5.11) as follows.

$$\rho + h(i) = \sup_{a \in A(i)} \{\hat{r}(i,a) + \sum_j \hat{p}_{ij}(a) h(j)\}, \ \ i \in S, \tag{5.12}$$

$$\rho + h(i) \leq \sup_{a \in A(i)} \{\hat{r}(i,a) + \sum_j \hat{p}_{ij}(a) h(j)\}, \ \ i \in S, \tag{5.13}$$

$$\rho + h(i) \leq \hat{r}(i,f) + \sum_j \hat{p}_{ij}(f) h(j), \ \ i \in S. \tag{5.14}$$

A solution of each equation above includes a constant ρ and a function h on S. By substituting $\hat{p}_{ij}(a)$ and $\hat{r}(i,a)$ into the above equations, we obtain the optimality equation and the optimality inequalities for the SMDP model (5.1) with the average criterion as follows.

$$h(i) = \sup_{a \in A(i)} \{r(i,a) - \rho\tau(i,a) + \sum_j p_{ij}(a) h(j)\}, \ \ i \in S, \tag{5.15}$$

$$h(i) \leq \sup_{a \in A(i)} \{r(i,a) - \rho\tau(i,a) + \sum_j p_{ij}(a) h(j)\}, \ \ i \in S, \tag{5.16}$$

$$h(i) \leq r(i,f) - \rho\tau(i,f) + \sum_j p_{ij}(f) h(j), \ \ i \in S. \tag{5.17}$$

We say Eq. (5.12) and Eq. (5.15) correspond to each other and each one is called a corresponding equation of the other. Also, we call Eq. (5.13) and Eq. (5.16) corresponding, and Eq. (5.14) and Eq. (5.17) corresponding.

We let $\tau(i) = \sup_{a \in A(i)} \tau(i,a)$ be the supremum of the expected duration over actions $a \in A(i)$ for $i \in S$. The equivalence between the SMDP model (5.1) with the DTMDP model (5.11) is established in the following theorem.

Theorem 5.3: *We have the following equivalence.*

1. *Suppose that $\tau(i)$ is finite for all $i \in S$. If $\{\rho, h\}$ is a solution of Eq. (5.12) or Eq. (5.13), then $\{\rho, \tau_* h\}$ is a solution of their corresponding equations.*
2. *If $\{\rho, h\}$ is a solution of Eq. (5.15) or Eq. (5.16), then $\{\rho, \tau_*^{-1} h\}$ is a solution of their corresponding equations.*
3. *For $f \in F$, $\{\rho, h\}$ is a solution of Eq. (5.14) if and only if $\{\rho, \tau_*^{-1} h\}$ is a solution of its corresponding equation.*

Proof: 1. Suppose that $\{\rho, h\}$ is a solution of Eq. (5.12). Then, by substituting $\hat{p}_{ij}(a)$ and $\hat{r}(i,a)$ into it, we obtain for $i \in S$,

$$0 = \sup_{a \in A(i)} \frac{1}{\tau(i,a)} \{ r(i,a) - \rho \tau(i,a) + \tau_* \sum_j p_{ij}(a) h(j) - \tau_* h(i) \}. \quad (5.18)$$

Because $\tau(i,a)$ is positive, we get

$$\tau_* h(i) \geq r(i,a) - \rho \tau(i,a) + \sum_j p_{ij}(a) \tau_* h(j), \ \ \forall (i,a) \in \Gamma,$$

or equivalently,

$$\tau_* h(i) \geq \sup_{a \in A(i)} \{ r(i,a) - \rho \tau(i,a) + \sum_j p_{ij}(a) \tau_* h(j) \}, \ \ i \in S. \quad (5.19)$$

On the other hand, for any state $i \in S$ and constant $\varepsilon > 0$, there is an action $a \in A(i)$ due to (5.18) such that

$$-\varepsilon \leq \frac{1}{\tau(i,a)} \{ r(i,a) - \rho \tau(i,a) + \sum_j p_{ij}(a) \tau_* h(j) - \tau_* h(i) \}, \ \ i \in S.$$

By multiplying the above inequality by $\tau(i,a)$ and rearranging it, we get

$$\begin{aligned} \tau_* h(i) - \varepsilon \tau(i) &\leq \tau_* h(i) - \varepsilon \tau(i,a) \\ &\leq r(i,a) - \rho \tau(i,a) + \sum_j p_{ij}(a) \tau_* h(j) \\ &\leq \sup_{a \in A(i)} \{ r(i,a) - \rho \tau(i,a) + \sum_j p_{ij}(a) \tau_* h(j) \}, \ \ i \in S. \end{aligned}$$

Because $\tau(i)$ is finite, letting ε tend to zero we obtain due to (5.19) that $\{\rho, \tau_* h\}$ is a solution of Eq. (5.15).

The result for Eq. (5.13) can be proved similarly.

2. Suppose that $\{\rho, h\}$ is a solution of Eq. (5.15). Then,

$$h(i) \geq r(i,a) - \rho\tau(i,a) + \sum_j p_{ij}(a)h(j), \ \ \forall (i,a) \in \Gamma.$$

Dividing the above inequality by $\tau(i,a)$ and adding $[\tau_*^{-1} - \tau(i,a)^{-1}]h(i)$ to both sides, we get

$$\tau_*^{-1}h(i) \geq \sup_{a \in A(i)} \{\hat{r}(i,a) - \rho + \sum_j \hat{p}_{ij}(a)\tau_*^{-1}h(j)\}, \ \ i \in S. \tag{5.20}$$

On the other hand, from Eq. (5.15) we know that for each $n \geq 1$ there is $f_n \in F$ such that

$$h(i) - \frac{1}{n} \leq \hat{r}(i, f_n) - \rho\tau(i, f_n) + \sum_j \hat{p}_{ij}(f_n)h(j), \ \ i \in S.$$

Dividing the above inequality by $\tau(i, f_n)$ and adding $\tau_*^{-1}h(i)$ to both sides, we get

$$\begin{aligned} & \rho + \tau_*^{-1}h(i) - [n\tau(i, f_n)]^{-1} \\ \leq \ & \hat{r}(i, f_n) - \rho\tau(i, f_n) + \sum_j \hat{p}_{ij}(f_n)\tau_*^{-1}h(j) \\ \leq \ & \sup_{a \in A(i)} \{\hat{r}(i,a) - \rho + \sum_j \hat{p}_{ij}(a)\tau_*^{-1}h(j)\}, \ \ i \in S. \end{aligned}$$

Because $\tau(i,a) \geq \delta\theta$, letting n tend to infinity above we obtain from Eq. (5.20) that $\{\rho, \tau_*^{-1}h\}$ is a solution of Eq. (5.12).

The result for Eq. (5.16) can be proved similarly.

3. This can be similarly proved as above. □

The above Theorem 5.3 establishes only the equivalence between the two models for the optimality equation and the optimality inequalities for the average criterion. It should be noted that in the above theorem, conclusion 1 needs the condition that all $\tau(i)$ are finite, which is not needed in conclusions 2 and 3.

There are plenty of results concerning the optimality equation and optimality inequality for DTMDPs with the average criterion (e.g., those in Chapter 3). All of them can be directly used for SMDPs, as those in Section 4.3 for CTMDPs. But the details are omitted here except the following theorem, which is similar to Theorem 3.5.

Theorem 5.4: *Suppose that $\rho \geq V^*(i)$ for all $i \in S$, and there are an extended real function h on S and a stationary policy f satisfying the*

following two conditions for some nonnegative constant ε.

$$h(i) \le r(i,f) - \rho\tau(i,f) + \sum_j p_{ij}(f)h(j) + \varepsilon, \quad i \in S, \tag{5.21}$$

$$\limsup_{n\to\infty} \frac{E_{f,i}h(X_n)}{\sum_{t=0}^{n-1} E_{f,i}\tau(X_t,f)} \le 0, \quad i \in S. \tag{5.22}$$

Then, $\rho \le V(f,i) + (\delta\theta)^{-1}\varepsilon$ *for* $i \in S$ *with finite* $h(i)$.

Proof: We assume first that all $h(i)$ are finite. Then, it can be proved similarly to Theorem 3.5 that $E_{f,i}h(X_n)$ and $E_{f,i}r(X_n,f)$ are well defined and finite for each $i \in S$ and $n \ge 0$. This implies that $E_{f,i}\tau(X_n,f)$ is also finite. Due to Eq. (5.21), we have

$$E_{f,i}h(X_n) \le E_{f,i}r(X_n,f) + \varepsilon - \rho E_{f,i}\tau(X_n,f) + E_{f,i}h(X_{n+1}), \quad i \in S,$$

for each $n \ge 0$. With this we can prove that

$$\rho \le \frac{\sum_{n=0}^{N-1} E_{f,i}r(X_n,f) + N\varepsilon}{\sum_{n=0}^{N-1} E_{f,i}\tau(X_n,f)} + \frac{E_{f,i}h(X_N) - h(i)}{\sum_{n=0}^{N-1} E_{f,i}\tau(X_n,f)}.$$

This together with Lemma 5.2, Eq. (5.22) implies $\rho \le V(f,i) + (\delta\theta)^{-1}\varepsilon$ for $i \in S$.

If there is i such that $h(i)$ is infinite, the result can be proved similarly. □

The above theorem implies that a policy f is $(\delta\theta)^{-1}\varepsilon$-optimal under the given conditions, if it achieves the ε-supremum of the optimality equation (5.15) or the optimality inequality, Eq. (5.16).

Unfortunately, there is no equivalence between the average objective functions for the two models. Beutler and Ross [3] discussed some properties for it.

3. Notes and References

SMDPs were first studied by Howard [56] and Jewell [81] separately. The study of them in the literature is mainly parallel to that for DTMDPs (e.g., see [35], [37] and [92]). Accompanying the study there was presented a transformation for SMDPs into DTMDPs for the average criterion [120], [54] and [36].

Here, we focus on the method of the transformation, for the discounted criterion, the total reward criterion, and the average criterion. It seems that this chapter is the first time SMDPs have been studied systematically by transforming them into DTMDPs. The contents of this chapter are based on the transformation presented in Schweitzer [120], Hordijk et al. [54], and Federgruen and Tijms [36].

Problems

1. Optimal Service Control of Queueing System $M/M/1$. Consider a queueing system $M/M/1$ in which the customers arrive according to a Poisson process with the arrival rate λ. The arrived customers are served one after another and so they may wait for the service when the server is busy for service. The service times for customers are random with the identical exponential distribution function with the mean of $1/\mu$. We call μ the service rate, which is chosen from a closed interval $[\underline{\mu}, \overline{\mu}]$ *when a new customer begins his service* (note this differs from Problem 1 in Chapter 4). When the service rate μ is chosen, a cost rate is $c(\mu)$. On the other hand, the holding cost per time unit $h(i)$ is incurred when there are i customers in the queue (including the one served). Assume that both $c(\mu)$ and $h(i)$ are increasing and concave.

Set this up as a semi-Markov decision process model and write the optimality equations for both the discounted total expected cost and the average expected cost.

2. Optimal Service Control of Queueing System $M/G/1$. This problem is similar to Problem 1 except that the distribution function of the service time is chosen from a set $\{G_a, a \in A\}$ when a new customer begins his service. For minimizing the discounted total expected cost, set this up as a semi-Markov decision process model and write the optimality equations for both the discounted total expected cost and the average expected cost.

3. Optimal Arrival Control of Queueing System $G/M/1$. Consider a queueing system $G/M/1$ in which the inter-arrival time between two adjacent customers is according to a general distribution function F. For any arrival, the manager of the system should determine either to admit his entrance or to reject his entrance. The manager can earn r from serving one customer, while she also earn a holding cost rate $h(i)$ when there are i customers in the queue.

Set this up as a semi-Markov decision process model and write the optimality equations for both the discounted total expected cost and the average expected cost.

Chapter 6

MARKOV DECISION PROCESSES IN SEMI-MARKOV ENVIRONMENTS

In this chapter, we deal with Markov decision processes in semi-Markov environments with the discounted criterion. The model can describe such a system that itself can be modeled by a Markov decision process, but the system is influenced by its environment which is modeled by a semi-Markov process. The influence of the environment on the system occurs when the environment state changes, and consists of the following three things: (1) an instantaneous state (of the system) transition, (2) an instantaneous reward, and (3) the parameters of the Markov decision process change. We study CTMDPs and then SMDPs in semi-Markov environments. Based on them, we study mixed MDPs in a semi-Markov environment, where the underlying MDP model can be either CTMDPs or SMDPs according to which environment states are entered. The criterion considered is the discounted criterion here. The standard results for all the models are obtained.

1. Continuous Time Markov Decision Processes in Semi-Markov Environments

In this section, we first present the model and then show the validity of the optimality equation. After this, we discuss an approximation problem when the kernel of the semi-Markov environment is approximated by another kernel. Finally, we discuss two special environments of Markov and phase type.

1.1 Model

This section deals with the following nonstationary continuous time Markov decision process in a semi-Markov environment (CTMDP-SE for short),

$$\{(K, G), (CTMDP^k, p^k, R^k, k \in K), U\}, \tag{6.1}$$

where

(a) The system's environment is described by a stationary semi-Markov process $\{(J_n, L_n), n \geq 0\}$, where J_n takes values in a countable set K and the elements in K are called the *environment states*. The kernel of $\{(J_n, L_n)\}$ is

$$G_{kk'}(t) = Pr\{J_{n+1} = k', \Delta L_n \leq t \mid J_n = k\},$$

where $\Delta L_n := L_{n+1} - L_n$. Let

$$\psi_{kk'} = G_{kk'}(+\infty), \;\; G_k(t) = \sum_{k' \in K} G_{kk'}(t)$$

be, respectively, the state transition probability from state k to k' and the distribution function of sojourning time at state k.

(b) When the environment state is k, the system can be described by a non-stationary continuous time Markov decision process:

$$CTMDP_k := \{S, A(i), q_{ij}^k(t,a), r^k(t,i,a)\}. \tag{6.2}$$

Here, the state space S and the action set $A(i)$ are all countable. In order to differentiate the environment states from the elements in S, we call the latter the *inner states* of the system. $\{q_{ij}^k(t,a)\}$ is the state transition rate family when the environment state is k. We assume that $q_{ij}^k(t,a) \geq 0$ for $i \neq j$ and $\sum_j q_{ij}^k(t,a) = 0$ for all $(i,a) \in \Gamma$ and $t \geq 0$. $r^k(t,i,a)$ is the reward rate when the environment is in state k for a time period t and the inner state is i and action a is chosen. The detailed meaning of each element is the same as that in the model in Section 4.2.

(c) For $i, j \in S, a \in A(i), k \in K$, $p_{ij}^k :=$ { the inner state at L_{n+1} is j | the inner state at $L_{n+1} - 0$ is i and $J_n = k$} is the instantaneous state transition probability at the epoch where the environment state changes. We assume that p_{ij}^k is independent of n. Here, $\sum_j p_{ij}^k$ may be less than one, and $1 - \sum_j p_{ij}^k$ can be interpreted as the terminated probability of the system that is terminated by the environment state transition at L_{n+1}.

(d) If the state and the action taken at $L_{n+1} - 0$ are i and a, respectively, and $J_n = k$, then the system receives an instantaneous reward $R^k(i,a)$ at $L_{n+1} - 0$.

(e) U is the discounted criterion with the discount rate $\alpha > 0$. It is defined in the following.

Remark 6.1: *If both S and $A(i)$ in Eq. (6.2) depend on k, all the results following can be proved exactly.*

The policy of the model has the form $\pi = (\pi_t^k, k \in K, t \geq 0)$, which means that for $k \in K$ and $n \geq 0, t \geq 0$, the action taken at $L_n + t$ is according to a probability distribution $\pi_t^k(\cdot \mid i)$ on $A(i)$ if the inner state at $L_n + t$ is i and $t < \Delta L_n, J_n = k$. In addition, $\pi_t^k(a \mid i)$ is measurable in t for each i, a, k. Let

Π_m denote the policy set. Subsets of Π_m, such as Π_m^d, Π_s, Π_s^d, can be defined exactly as those in Section 4.1.

For each policy $\pi = (\pi_t^k, k \in K, t \geq 0) \in \Pi_m$ and $k \in K, t \geq 0$, we define the system's state transition rate matrix $Q^k(\pi, t) = (q_{ij}^k(\pi, t))$ and the reward rate column vector $r^k(\pi, t) = (r_i^k(\pi, t))$ as $Q(\pi, t)$ and $r(\pi, t)$, respectively, in Section 4.1, and similarly the instantaneous reward column vector $R^k(\pi, t)$ when the environment state changes. Obviously, all of them depend on π only through π^k. In order to ensure that the model is well defined, we present three conditions in the following.

Condition 6.1: *There exist positive constants θ and δ such that $G_k(\delta) \leq 1 - \theta$ for all $k \in K$.*

This condition is Regular Condition 1 in Chapter 5 for SMDPs. It ensures that there occur only finitely many state transitions of the environment in every finite time interval.

Condition 6.2: *1. For each $\pi \in \Pi_m, k \in K, i, j \in S, q_{ij}^k(\pi, t)$ is continuous almost everywhere (a.e.).*

2. There exists a function $Q(t)$ which is integrable in every finite interval such that $-q_{ii}^k(t, a) \leq Q(t)$ a.e. t for every $k \in K$ and $(i, a) \in \Gamma$.

The condition above is exactly Condition 4.4 for nonstationary CTMDPs. Then, as in Section 4.2, for each policy π and $k \in K$, there exists a unique absolutely continuous $Q^k(\pi, t)$-process $\{P(\pi^k, s, t), 0 \leq s \leq t < \infty\}$.

Condition 6.3: *1. $r^k(t, i, a)$ is Lebesgue measurable in t for every $k \in K$ and $(i, a) \in \Gamma$.*

2. There exists a nonnegative function $r(t)$ and a positive constant M such that:

$$|R^k(i, a)| \leq M, \ | r^k(t, i, a) | \ \leq r(t), \ \text{ a.e. } t \geq 0, \ k \in K, (i, a) \in \Gamma,$$
$$\int_0^\infty e^{-\alpha t} r(t) dt \leq M.$$

This condition ensures that the discounted criterion defined in the following will exist and be finite.

For $t \geq 0$, we denote by $V(t)$ the discounted total reward on $[t, \infty)$, and by $Y(t)$ and $\Delta(t)$ the state and the action taken at time t, respectively. Moreover, we let

$$\begin{aligned} r(n, t) &= r^{J_n}(t - L_n, Y(t), \Delta(t)), \ L_n \leq t < L_{n+1}, \\ R(n) &= R^{J_n}(Y(L_{n+1} - 0), \Delta(L_{n+1} - 0)) \end{aligned}$$

be the the reward rate at $t \in [L_n, L_{n+1})$ and the instantaneous reward at epoch $L_{n+1} - 0$, respectively. Then,

$$\begin{aligned} V(t) &= \int_t^{L_{n+1}} e^{-\alpha(s-t)} r(n,s) ds + \sum_{j=n+1}^{\infty} \int_{L_j}^{L_{j+1}} e^{-\alpha(s-t)} r(j,s) ds \\ &\quad + \sum_{j=n}^{\infty} e^{-\alpha(L_{j+1}-t)} R(j), \quad L_n \le t < L_{+1}, n \ge 0. \end{aligned} \tag{6.3}$$

Now, we define the discounted criterion by

$$U_k(\pi, t_0, t, i) = \bar{G}_k(t - t_0) E_\pi\{V(t) | L_n = t_0, L_{n+1} \ge t, Y(t) = i, J_n = k\},$$

where $\bar{G}_k(t) = 1 - G_k(t)$. The term $\bar{G}_k(t - t_0)$ introduced above is just for simplicity of the following expressions (e.g., Eq. (6.5) later). It would not influence any effect on the problem. The expectation in the above formulae represents the expected discounted total reward in $[t, \infty)$, discounted to t, under the given conditions "$L_n = t_0, L_{n+1} \ge t, Y(t) = i, J_n = k$". It should be noted that n in the above definition can be arbitrary because the environment process is stationary. Let $U_k(\pi, t_0, t)$ be a column vector with the ith component $U_k(\pi, t_0, t, i)$ for $i \in S$.

It follows from Eq. (6.3) that for $L_n \le t < L_{n+1}$,

$$\begin{aligned} |V(t)| &\le \int_{t-L_n}^{\Delta L_n} e^{-\alpha(s+L_n-t)} r(s) ds + \sum_{j=n+1}^{\infty} \int_0^{\Delta L_j} e^{-\alpha(s+L_j-t)} r(s) ds \\ &\quad + \sum_{j=n}^{\infty} e^{-\alpha(L_{j+1}-t)} M \\ &\le \int_0^{\infty} e^{-\alpha(s-t)} r(s) ds + \sum_{j=n+1}^{\infty} e^{-\alpha(L_j - L_{n+1})} \int_0^{\infty} e^{-\alpha s} r(s) ds \\ &\quad + \sum_{j=n+1}^{\infty} e^{-\alpha(L_j - L_{n+1})} M \\ &\le M e^{\alpha t} + 2M \sum_{j=n+1}^{\infty} e^{-\alpha(L_j - L_{n+1})}. \end{aligned}$$

Similarly to Lemma 5.1, we have from Condition 6.1 that

$$\begin{aligned} &\int_0^{\infty} e^{\alpha s} dG_k(s) \le 1 - \theta(1 - e^{-\alpha\delta}) := \beta < 1, \\ &E \sum_{j=n+1}^{\infty} e^{-\alpha(L_j - L_{n+1})} \le \sum_{j=n+1}^{\infty} \beta^{j-n-1} = (1-\beta)^{-1}. \end{aligned}$$

With this and Conditions 6.1–6.3, we can conclude that $U_k(\pi, t_0, t, i)$ is well defined, and

$$|U_k(\pi, t_0, t, i)| \leq Me^{\alpha t} + \frac{2M}{1-\beta} := M(t), \ \forall \pi, k, t, i. \tag{6.4}$$

Moreover, it follows from Eq. (6.3) and the above equation that

$$\begin{aligned}
& U_k(\pi, t_0, t, i) \\
= & \sum_{k'} \int_{t-t_0}^{\infty} dG_{kk'}(u) \\
& \cdot E_\pi\{V(t) | L_n = t_0, L_{n+1} = t_0 + u, Y(t) = i, J_n = k\} \\
= & \sum_{k'} \int_{t-t_0}^{\infty} dG_{kk'}(u) \\
& \cdot \{\int_t^{t_0+u} e^{-\alpha(s-t)} \sum_j P_{ij}(\pi^k, t-t_0, s-t_0) r_j^k(\pi, s-t_0) ds \\
& + e^{-\alpha(t_0+u-t)} \sum_j P_{ij}(\pi^k, t-t_0, u) \\
& \cdot [R_j^k(\pi, u) + \sum_l p_{jl}^k E_\pi(V(t_0+u) \mid (t_0+u, t_0+u, l, k'))]\},
\end{aligned}$$

where $(t_0 + u, t_0 + u, l, k')$ represents the event $\{L_{n+1} = t_0 + u, L_{n+2} \geq t_0 + u, Y(t_0 + u) = l, J_{n+1} = k'\}$. Let $P^k = (p_{ij}^k)$ be a matrix. By exchanging the order of the two integrations above, we have with the form of columns,

$$\begin{aligned}
U_k(\pi, t_0, t) = & \int_t^{\infty} e^{-\alpha(s-t)} P(\pi^k, t-t_0, s-t_0) r^k(\pi, s-t_0) \bar{G}_k(s-t_0) ds \\
& + \sum_{k'} \int_{t-t_0}^{\infty} e^{-\alpha(t_0+u-t)} P(\pi^k, t-t_0, u) \\
& \cdot [R^k(\pi, u) + P^k U_{k'}(\pi, t_0+u, t_0+u)] dG_{kk'}(u) \\
= & \int_{t-t_0}^{\infty} e^{-\alpha(s-(t-t_0))} P(\pi^k, t-t_0, s) r^k(\pi, s) \bar{G}_k(s) ds \\
& + \sum_{k'} \int_{t-t_0}^{\infty} e^{-\alpha(t_0+u-t)} P(\pi^k, t-t_0, u) \\
& \cdot [R^k(\pi, u) + P^k U_{k'}(\pi, t_0+u, t_0+u)] dG_{kk'}(u).
\end{aligned}$$

From the definition of $U_k(\pi)$ and Eq. (6.3), we know that $U_k(\pi, t, t, i)$ is irrespective of t. With this and the above formula we know that $U_k(\pi, t_0, t)$ depends on t_0 and t only through $t - t_0$. We denote it by $U_k(\pi, t - t_0)$. Therefore, we can let $t_0 = 0$ and so we have that

$$U_k(\pi, t, i) = \bar{G}_k(t) E_\pi\{V(t) \mid L_n = 0, L_{n+1} \geq t, Y(t) = i, J_n = k\},$$

$$\begin{aligned} U_k(\pi,t) &= \int_t^\infty e^{-\alpha(s-t)}P(\pi^k,t,s)r^k(\pi,s)\bar{G}_k(s)ds \\ &\quad + \sum_{k'}\int_t^\infty e^{-\alpha(s-t)}P(\pi^k,t,s) \\ &\quad \cdot [R^k(\pi,s)+P^kU_{k'}(\pi,0)]dG_{kk'}(s). \end{aligned} \tag{6.5}$$

Because the environment is stationary, the variable n in the above equation can be arbitrary. We define the optimal value by

$$U_k^*(t)=\sup\{U_k(\pi,t):\pi\in\Pi_m\},\ \ t\geq 0.$$

For $\varepsilon\geq 0$ and $\pi^*\in\Pi_m$, if $U_k(\pi^*,0,i)\geq U_k^*(0,i)-\varepsilon$ for all $i\in S$ and $k\in K$, then we say π^* is ε-optimal. An 0-optimal policy is called optimal. In Subsection 2.2 later, we prove the existence of ε-optimal policies in a stronger sense.

For arbitrary $k,k'\in K$, if $G_{kk'}(\infty)>0$ then $G_{kk'}(t)/G_{kk'}(\infty)$ is a distribution function (d.f.). We assume further that it is mixed. That is, there is a constant $\gamma_{kk'}\in[0,1]$ such that

$$G_{kk'}(t)=\gamma_{kk'}\tilde{G}_{kk',1}(t)+(1-\gamma_{kk'})\tilde{G}_{kk',2}(t), \tag{6.6}$$

where $\tilde{G}_{kk',1}(t)/G_{kk'}(\infty)$ is an absolutely continuous d.f. with the probability density function (p.d.f.) $\tilde{g}_{kk'}(t)/G_{kk'}(\infty)$, and $\tilde{G}_{kk',2}(t)/G_{kk'}(\infty)$ is a discrete type d.f. with the probability law

$$\begin{pmatrix} t_0 & t_1 & \cdots & t_n & \cdots \\ \tilde{p}_{kk',0}/G_{kk'}(\infty) & \tilde{p}_{kk',1}/G_{kk'}(\infty) & \cdots & \tilde{p}_{kk',n}/G_{kk'}(\infty) & \cdots \end{pmatrix}.$$

Because K is countable, we assume above that t_n is irrespective of k,k'. Certainly, some $\tilde{p}_{kk',n}$ may be zero. Without loss of generality, we assume that t_n is increasing strictly in n and tends to infinity. We denote by

$$G_{kk'}(t)=G_{kk',1}(t)+G_{kk',2}(t),$$

where

$$\begin{aligned} &G_{kk',1}(t)=\gamma_{kk'}\tilde{G}_{kk',1}(t), \\ &g_{kk'}(t)=\gamma_{kk'}\tilde{g}_{kk'}(t),\ \ g_k(t)=\sum_{k'}g_{kk'}(t), \\ &G_{kk',2}(t)=(1-\gamma_{kk'})\tilde{G}_{kk',2}(t), \\ &p_{kk',n}=(1-\gamma_{kk'})\tilde{p}_{kk',n},\ \ p_{k,n}=\sum_{k'}p_{kk',n}. \end{aligned}$$

Moreover, we let

$$\hat{r}^k(t,i,a):=r^k(t,i,a)\bar{G}_k(t)+R^k(i,a)g_k(t),\ \ \forall k,t,i,a,$$

and define the column vector $\hat{r}^k(\pi, t)$ similarly to $r^k(\pi, t)$. Certainly,

$$\int_0^\infty e^{-\alpha s} g_k(s) ds \leq \beta, \quad k \in K,$$

where $\beta = 1 - \theta(1 - e^{-\alpha\delta})$. Then, from Eq. (6.5) we have

$$\begin{aligned} U_k(\pi, t) &= \int_t^\infty e^{-\alpha(s-t)} P(\pi^k, t, s) \\ &\quad \cdot \{\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U_{k'}(\pi, 0)\} ds \\ &\quad + \sum_{t_m \geq t} e^{-\alpha(t_m - t)} P(\pi^k, t, t_m) \\ &\quad \cdot [p_{k,m} R^k(\pi, t_m) + \sum_{k'} p_{kk',m} P^k U_{k'}(\pi, 0)]. \end{aligned} \tag{6.7}$$

Based on this equation, we know that if $\gamma_{kk'} \equiv 1$ then

$$\begin{aligned} U_k(\pi, t) &= \int_t^\infty e^{-\alpha(s-t)} P(\pi^k, t, s) \\ &\quad \cdot \{\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U_{k'}(\pi, 0)\} ds. \end{aligned}$$

It depends on $r^k(t, i, a)$ and $R^k(i, a)$ only through $\hat{r}^k(t, i, a)$; that is, the two rewards $r^k(t, i, a)$ and $R^k(i, a)$ can be integrated into one.

1.2 Optimality Equation

We borrow methods and ideas from the nonstationary CTMDPs, studied in Section 4.2, to deal with the optimality equation. First, we prove three lemmas. We define Ω to be a set of functions $v = (v_k(t, i))$ satisfying the following three conditions.

1. For each $k \in K, i \in S$, $v_k(t, i)$, defined on $[0, \infty)$, is absolutely continuous in every interval $(t_n, t_{n+1}]$ for $n \geq 0$; that is, $v_k(t, i)$ is absolutely continuous in closed interval $[t', t_{n+1}]$ for each $t' \in (t_n, t_{n+1}]$, and the limit of $v_k(t, i)$ exists as t tends to $t_n + 0$.
2. $e^{-\alpha t} v_k(t, i)$ tends to zero uniformly in k, i as t tends to infinity.
3. There exists a function $N(t)$ that is integrable in every interval $[t_n, t_{n+1}]$, $n \geq 0$, such that the differential of $v_k(t, i)$ is uniformly bounded above with $N(t)$; that is, $|v_k'(t, i)| \leq N(t)$ a.e. $t \geq 0$ for $k \in K$ and $i \in S$.

It is obvious that $\{U_k(\pi, t)\}$ belongs to the set Ω for each policy $\pi \in \Pi_m$. Therefore, the optimal value $\{U_k^*(t)\}$ also belongs to the set Ω. Hence, Ω is, in fact, the discounted criterion space.

In the following, we refer to t, k, n as $t \geq 0, k \in K$, and $n \geq 0$, respectively, if no other specification is given.

Similarly to Lemma 4.9 in Section 4.2, we have the following lemma based on Eq. (6.7).

Lemma 6.1: *For each policy $\pi = (\pi_t^k) \in \Pi_m$, a constant $\varepsilon \geq 0$ and a function $v = (v_k(t,i))$ in the set Ω,*

1. *$v_k(t) \leq U_k(\pi,t) + (2-\beta)(1-\beta)^{-1}\alpha^{-1}\varepsilon \boldsymbol{e}$ for all $t \geq 0$ and $k \in K$ if*

$$-\frac{d}{dt}v_k(t) \leq \hat{r}^k(\pi,t) + \sum_{k'} g_{kk'}(t)P^k v_{k'}(0) + Q^k(\pi,t)v_k(t)$$
$$-\alpha v_k(t) + \varepsilon \boldsymbol{e}, \text{ a.e. } t,k, \tag{6.8}$$
$$v_k(t_n) = v_k(t_n+0) + p_{k,n}R^k(\pi,t_n) + \sum_{k'} p_{kk',n}P^k U_{k'}(\pi,0), n,k. \tag{6.9}$$

2. *$v_k(t) \geq U_k(\pi,t) - (2-\beta)(1-\beta)^{-1}\alpha^{-1}\varepsilon \boldsymbol{e}$ for all $t \geq 0$ and $k \in K$ if*

$$-\frac{d}{dt}v_k(t) \geq \hat{r}^k(\pi,t) + \sum_{k'} g_{kk'}(t)P^k v_{k'}(0) + Q^k(\pi,t)v_k(t)$$
$$-\alpha v_k(t) - \varepsilon \boldsymbol{e}, \text{ a.e. } t,k,$$
$$v_k(t_n) = v_k(t_n+0) + p_{k,n}R^k(\pi,t_n) + \sum_{k'} p_{kk',n}P^k U_{k'}(\pi,0), n,k.$$

Proof: 1. Pre-multiplying by $e^{-\alpha(s-t)}P(\pi^k,t,s)$ in Eq. (6.8), with t being replaced by s, we can prove similarly to Lemma 4.9 in Section 4.2 that

$$\begin{aligned} & -\frac{\partial}{\partial s}\{e^{-\alpha(s-t)}P(\pi^k,t,s)v_k(s)\} \\ \leq\ & e^{-\alpha(s-t)}P(\pi^k,t,s)\{\hat{r}^k(\pi,s) + \sum_{k'} g_{kk'}(s)P^k v_{k'}(0) + \varepsilon \boldsymbol{e}\} \end{aligned}$$

for a.e. $s \geq t \geq 0$ and $k \in K$. Now, for any given $n \geq 0$ and $t_n < t \leq t_{n+1}$, integrating the above formula in $s \in [t, t_{n+1}]$, we get that

$$\begin{aligned} & v_k(t) - e^{-\alpha(t_{n+1}-t)}P(\pi^k,t,t_{n+1})v_k(t_{n+1}) \\ \leq\ & \int_t^{t_{n+1}} e^{-\alpha(s-t)}P(\pi^k,t,s)[\hat{r}^k(\pi,s) + \sum_{k'} g_{kk'}(s)P^k v_{k'}(0) + \varepsilon \boldsymbol{e}]ds, \end{aligned}$$

for $t \in (t_n, t_{n+1}], n \geq 0$ and $k \in K$. Letting $t \to t_n + 0$, one can get from Eq. (6.9) that

$$\begin{aligned} v_k(t_n) \ \leq\ & e^{-\alpha(t_{n+1}-t_n)}P(\pi^k,t_n,t_{n+1})v_k(t_{n+1}) \\ & + p_{k,n}R^k(\pi,t_n) + \sum_{k'} p_{kk',n}P^k U_{k'}(\pi,0) \\ & + \int_{t_n}^{t_{n+1}} e^{-\alpha(s-t_n)}P(\pi^k,t_n,s) \\ & \cdot [\hat{r}^k(\pi,s) + \sum_{k'} g_{kk'}(s)P^k v_{k'}(0) + \varepsilon \boldsymbol{e}]ds. \end{aligned}$$

We write the above two formulae as the following unified one:

$$\begin{aligned} v_k(t) \;\leq\; & e^{-\alpha(t_{n+1}-t)}P(\pi^k,t,t_{n+1})v_k(t_{n+1}) \\ & + [p_{k,n}R^k(\pi,t_n) + \sum_{k'} p_{kk',n}P^k U_{k'}(\pi,0)]\delta_{t,t_n} \\ & + \int_t^{t_{n+1}} e^{-\alpha(s-t)}P(\pi^k,t,s) \\ & \cdot [\hat{r}^k(\pi,s) + \sum_{k'} g_{kk'}(s)P^k v_{k'}(0) + \varepsilon \boldsymbol{e}]ds, \\ & \qquad t \in [t_n,t_{n+1}],\;\; n \geq 0,\;\; k \in K, \end{aligned} \tag{6.10}$$

where $\delta_{t,t_n} = 1$ when $t = t_n$ and $\delta_{t,t_n} = 0$ otherwise.

Now, Eq. (6.7) can be rewritten in the form of the above equation for $t \in [t_n, t_{n+1}]$. That is,

$$\begin{aligned} & U_k(\pi,t) \\ & = \; e^{-\alpha(t_{n+1}-t)}P(\pi^k,t,t_{n+1})U_k(\pi,t_{n+1}) \\ & \quad + [p_{k,n}R^k(\pi,t_n) + \sum_{k'} p_{kk',n}P^k U_{k'}(\pi,0)]\delta_{t,t_n} \\ & \quad + \int_t^{t_{n+1}} e^{-\alpha(s-t)}P(\pi^k,t,s)[\hat{r}^k(\pi,s) + \sum_{k'} g_{kk'}(s)P^k U_{k'}(\pi,0)]ds, \\ & \qquad\qquad t \in [t_n,t_{n+1}],\;\; n \geq 0,\;\; k \in K. \end{aligned} \tag{6.11}$$

By letting $\Delta_k(t) = v_k(t) - U_k(\pi,t)$, it follows Eq. (6.10) and Eq. (6.11) that

$$\begin{aligned} \Delta_k(t) \;\leq\; & e^{-\alpha(t_{n+1}-t)}P(\pi^k,t,t_{n+1})\Delta_k(t_{n+1}) \\ & + \int_t^{t_{n+1}} e^{-\alpha(s-t)}P(\pi^k,t,s)[\sum_{k'} g_{kk'}(s)P^k\Delta_{k'}(0) + \varepsilon\boldsymbol{e}]ds, \end{aligned}$$

for $t \in [t_n, t_{n+1}]$ and n, k. This implies by the induction method that

$$\begin{aligned} \Delta_k(t_n) \;\leq\; & e^{-\alpha(t_N-t_n)}P(\pi^k,t_n,t_N)\Delta_k(t_N) \\ & + \int_{t_n}^{t_N} e^{-\alpha(s-t_n)}P(\pi^k,t_n,s)[\sum_{k'} g_{kk'}(s)P^k\Delta_{k'}(0) + \varepsilon\boldsymbol{e}]ds \end{aligned}$$

for $N > n \geq 0$. Letting $N \to +\infty$, from the definition of Ω, we get for $n \geq 0$

$$\Delta_k(t_n) \leq \int_{t_n}^{\infty} e^{-\alpha(s-t_n)}P(\pi^k,t_n,s)[\sum_{k'} g_{kk'}(s)P^k\Delta_{k'}(0) + \varepsilon\boldsymbol{e}]ds.$$

We denote $\Delta := \sup_{k,i} \Delta_k(0,i)$. From the above formula we have

$$\Delta \leq \int_0^{\infty} e^{-\alpha s}[g_k(s)\Delta + \varepsilon]ds \leq \beta\Delta + \alpha^{-1}\varepsilon.$$

Thus $\Delta \le \varepsilon^* := (1-\beta)^{-1}\alpha^{-1}\varepsilon$. With this we have that

$$\Delta_k(t_n) \le \int_{t_n}^{\infty} e^{-\alpha(s-t_n)}[g_k(s)\varepsilon^* + \varepsilon]ds \cdot \boldsymbol{e}, \quad n, k.$$

Therefore, for $t \in [t_n, t_{n+1}]$ and n,

$$\begin{aligned}\Delta_k(t) &\le e^{-\alpha(t_{n+1}-t)} \int_{t_{n+1}}^{\infty} e^{-\alpha(s-t_{n+1})}[g_k(s)\varepsilon^* + \varepsilon]ds \cdot \boldsymbol{e} \\ &\quad + \int_t^{t_{n+1}} e^{-\alpha(s-t)}[g_k(s)\varepsilon^* + \varepsilon]ds \cdot \boldsymbol{e} \\ &= \int_t^{\infty} e^{-\alpha(s-t)}[g_k(s)\varepsilon^* + \varepsilon]ds \cdot \boldsymbol{e} \\ &\le (\varepsilon^* + \alpha^{-1}\varepsilon)\boldsymbol{e} = (2-\beta)(1-\beta)^{-1}\alpha^{-1}\varepsilon\boldsymbol{e}, \quad k \in K.\end{aligned}$$

Hence, $v_k(t) \le U_k(\pi, t) + (2-\beta)(1-\beta)^{-1}\alpha^{-1}\varepsilon\boldsymbol{e}$ for all $t \ge 0$ and $k \in K$.

2. This can be proved in a similar way. □

Based on the above lemma, we have equations to characterize $U_k(\pi, t)$, which is given in the following lemma.

Lemma 6.2: *For each policy $\pi = (\pi_t^k) \in \Pi_m$, $U_k(\pi, t)$ is the unique solution in Ω of the following equations.*

$$\begin{aligned}-\frac{d}{dt}v_k(t) &= \hat{r}^k(\pi, t) + \sum_{k'} g_{kk'}(t)P^k v_{k'}(0) + Q^k(\pi, t)v_k(t) - \alpha v_k(t), \\ v_k(t_n) &= v_k(t_n + 0) + p_{k,n}R^k(\pi, t_n) + \sum_{k'} p_{kk',n}P^k U_{k'}(\pi, 0)]\end{aligned}$$

for a.e. $t \ge 0$ and k, n.

Proof: It follows from Eq. (6.7) that $U_k(\pi, t)$ satisfies the above two equations. Now if $v = (v_k(t)) \in \Omega$ is a solution of the above two equations, then by Lemma 6.1, $v_k(t) = U_k(\pi, t)$ for all t and k. □

Before giving the third lemma, we further divide each interval $[t_n, t_{n+1}]$. Fixing any given constant $\beta \in (0, 1)$, we can prove as in Lemma 4.8 in Section 4.2 that for each $n \ge 0$ there exist finite numbers $t_n = t_{n,0} < t_{n,1} < t_{n,2} < \cdots < t_{n,m_n} = t_{n+1}$ such that

$$\int_{t_{n,m}}^{t_{n,m+1}} [2Q(t) + \alpha]dt \le \beta, \quad m = 0, 1, \ldots, m_n - 1.$$

For each n, m, we let $M_{n,m}$ be a set of $v = (v_k(t, i))$, which is uniformly bounded and measurable in $t \in [t_{n,m}, t_{n,m+1}]$. Then, the following lemma can be proved as Lemma 4.10 in Section 4.2.

Lemma 6.3: *For each $m = 0, 1, \ldots, m_n - 1$ and $n \geq 0$, given bounded vectors $\{v_k(0), k \in K\}, \{x_{k,n,m}, k \in K\}$, and $\{y_{k,n,m}, k \in K\}$, there exists $\{U_{k,n,m}(t,i)\} \in M_{n,m}$ such that for all k, n, m, i,*

1. *$U_{k,n,m}(t,i)$ is absolutely continuous in $t \in (t_{n,m}, t_{n,m+1}]$.*

2. *$U_{k,n,m}(t_{n,m+1}) = y_{k,n,m}, U_{k,n,m}(t_{n,m}) = U_{k,n,m}(t_{n,m} + 0) + x_{k,n,m}$.*

3. *The following equation holds for a.e. $t \in (t_{n,m}, t_{n,m+1}]$:*

$$\begin{aligned} -\frac{d}{dt}U_{k,n,m}(t,i) &= \sup_{a \in A(i)} \{\hat{r}^k(t,i,a) + \sum_j q_{ij}^k(t,a) U_{k,n,m}(t,j)\} \\ &\quad - \alpha U_{k,n,m}(t,i) + \sum_{k'} g_{kk'}(t) \sum_j p_{ij}^k v_{k'}(0,j). \end{aligned}$$

Now, we can prove the optimality equation. This is completed in two steps. First, in the following lemma, we show that $U_k^*(t)$ satisfies the following equations (6.12) and (6.13), where the right-hand side includes the optimal value $U_{k'}^*(0)$. Second, we show in Theorem 6.1 below that $U_{k'}^*(0)$ can be replaced with $U_{k'}(0)$.

Lemma 6.4: *$U_k^*(t)$ is the unique solution in Ω of the following equations:*

$$\begin{aligned} -\frac{d}{dt}U_k(t) &= \sup_{f \in F}\{\hat{r}^k(f,t) + Q^k(f,t)U_k(t)\} \\ &\quad - \alpha U_k(t) + \sum_{k'} g_{kk'}(t) P^k U_{k'}(0), \quad \text{a.e. } t, k, \end{aligned} \tag{6.12}$$

$$U_k(t_n) = U_k(t_n + 0) + p_{k,n}R^k + \sum_{k'} p_{kk',n} P^k U_{k'}^*(0), \quad n, k, \tag{6.13}$$

where $R^k = (R^k(i), i \in S)$ is a column vector with its ith element being $R^k(i) = \sup\{R^k(i,a) : a \in A(i)\}$.

Proof: By Lemma 6.3, we know that for each $n \geq 0$ and $m = 0, 1, \ldots, m_n - 1$ there exists a $\{U_{k,n,m}(t)\} \in M_{n,m}$ such that conclusions 1–3 in Lemma 6.3 are satisfied with $x_{k,n,m} = \delta_{m,0}[p_{k,n}R^k + \sum_{k'} p_{kk',n}P^k U_{k'}^*(0)]$ and $v_k(0) = U_k^*(0), y_{k,n,m} = U_k^*(t_{n,m+1})$. Then, we let

$$U_k(t) = U_{k,n,m}(t), \quad \text{if } t \in [t_{n,m}, t_{n,m+1}), \quad m = 0, 1, \cdots, m_n - 1, \quad n \geq 0.$$

We first prove that $U_k(t) = U_k^*(t)$. By Lemma 6.3, we know that for each $\pi = (\pi_t^k) \in \Pi_m$,

$$\begin{aligned} -\frac{d}{dt}U_{k,n,m}(t) &\geq \hat{r}^k(\pi,t) + Q^k(\pi,t)U_{k,n,m}(t) \\ &\quad - \alpha U_{k,n,m}(t) + \sum_{k'} g_{kk'}(t) P^k U_{k'}^*(0), \end{aligned}$$

for a.e. $t \in [t_{n,m}, t_{n,m+1}]$, $m = 0, 1, \ldots, m_n - 1$ and n, k. With this, it can be proved similarly to Lemma 6.1 that

$$\begin{aligned}
&U_{k,n,m}(t) \geq e^{-\alpha(t_{n,m+1}-t)} P(\pi^k, t, t_{n,m+1}) U_k(\pi, t_{n,m+1}) \\
&+ [p_{k,n} R^k(\pi, t_n) + \sum_{k'} p_{kk',n} P^k U_{k'}(\pi, 0)] \delta_{t,t_n} \\
&+ \int_t^{t_{n,m+1}} e^{-\alpha(s-t)} P(\pi^k, t, s) [\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U^*_{k'}(0)] ds, \\
&\quad t \in [t_{n,m}, t_{n,m+1}], \;\; m = 0, 1, \ldots, m_n - 1, \;\; n \geq 0, \;\; k \in K.
\end{aligned}$$

This implies that

$$\begin{aligned}
&U_{k,n,m}(t) \geq \sup_{\pi \in \Pi_m} \{e^{-\alpha(t_{n,m+1}-t)} P(\pi^k, t, t_{n,m+1}) U_k(\pi, t_{n,m+1}) \\
&+ [p_{k,n} R^k(\pi, t_n) + \sum_{k'} p_{kk',n} P^k U_{k'}(\pi, 0)] \delta_{t,t_n} \\
&+ \int_t^{t_{n,m+1}} e^{-\alpha(s-t)} P(\pi^k, t, s) [\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U^*_{k'}(0)] ds\}, \\
&\quad t \in [t_{n,m}, t_{n,m+1}], \;\; m = 0, 1, \ldots, m_n - 1, \;\; n \geq 0, \;\; k \in K. \quad (6.14)
\end{aligned}$$

Note that Eq. (6.11) is still true when t_{n+1} is replaced by $t_{n,m+1}$ and t belongs to $[t_{n,m}, t_{n,m+1})$ for $m = 0, 1, \ldots, m_{n-1}$; that is,

$$\begin{aligned}
&U_k(\pi, t) = e^{-\alpha(t_{n,m+1}-t)} P(\pi^k, t, t_{n,m+1}) U_k(\pi, t_{n,m+1}) \\
&+ [p_{k,n} R^k(\pi, t_n) + \sum_{k'} p_{kk',n} P^k U_{k'}(\pi, 0)] \delta_{t,t_n} \\
&+ \int_t^{t_{n,m+1}} e^{-\alpha(s-t)} P(\pi^k, t, s) [\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U_{k'}(\pi, 0)] ds, \\
&\quad t \in [t_{n,m}, t_{n,m+1}], \;\; m = 0, 1, \ldots, m_n - 1, \;\; n \geq 0, \;\; k \in K.
\end{aligned}$$

This together with Eq. (6.14) implies that

$$U^*_k(t) = \sup_\pi U_k(\pi, t) \leq U_k(t), \;\; t \geq 0, k \in K. \quad (6.15)$$

On the other hand, from 3 of Lemma 6.3 and Lemma 4.11 we know that for each $\varepsilon > 0$, there exists a policy $\pi = (f^k_t) \in \Pi^d_m$ such that

$$\begin{aligned}
-\frac{d}{dt} U_{k,n,m}(t) \;\; &\leq \;\; \hat{r}^k(\pi, t) + Q^k(\pi, t) U_{k,n,m}(t) \\
&\quad - \alpha U_{k,n,m}(t) + \sum_{k'} g_{kk'}(t) P^k U^*_{k'}(0) + \varepsilon \boldsymbol{e},
\end{aligned}$$

for a.e. $t \in [t_{n,m}, t_{n,m+1}], m = 0, 1, \ldots, m_n - 1$ and n, k. Similarly to the above, we can get

$$\begin{aligned}
&U_{k,n,m}(t) \le e^{-\alpha(t_{n,m+1}-t)} P(\pi^k, t, t_{n,m+1}) U_k(t_{n,m+1}) \\
&+ [p_{k,n} R^k + \sum_{k'} p_{kk',n} P^k U^*_{k'}(0)] \delta_{t,t_n} \\
&+ \int_t^{t_{n,m+1}} e^{-\alpha(s-t)} P(\pi^k, t, s) [\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U^*_{k'}(0) + \varepsilon \boldsymbol{e}] ds \\
&\qquad t \in [t_{n,m}, t_{n,m+1}], \;\; m = 0, 1, \ldots, m_n - 1, \;\; n \ge 0, \;\; k \in K.
\end{aligned}$$

Furthermore, the policy π above can be chosen such that $R^k(\pi, t_n) \ge R^k - \varepsilon \boldsymbol{e}$ for all k and n. Thus, we can get from the above formula that

$$\begin{aligned}
&U_k(t) \le e^{-\alpha(t_{n,m+1}-t)} P(\pi^k, t, t_{n,m+1}) U_k(t_{n,m+1}) \\
&+ [p_{k,n} R^k(\pi, t_n) + \sum_{k'} p_{kk',n} P^k U^*_{k'}(0) + p_{k,n} \varepsilon \boldsymbol{e}] \delta_{t,t_n} \\
&+ \int_t^{t_{n,m+1}} e^{-\alpha(s-t)} P(\pi^k, t, s) [\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U^*_{k'}(0) + \varepsilon \boldsymbol{e}] ds, \\
&\qquad t \in [t_{n,m}, t_{n,m+1}], \;\; m = 0, 1, \ldots, m_n - 1, \;\; n \ge 0, \;\; k \in K.
\end{aligned}$$

Getting the expression of $U_k(t_{n,m+1})$ from the above equation and substituting it iteratively into the above equation, we can get that

$$\begin{aligned}
&U_k(t) \le e^{-\alpha(t_{n+1}-t)} P(\pi^k, t, t_{n+1}) U_k(t_{n+1}) \\
&+ [p_{k,n} R^k(\pi, t_n) + \sum_{k'} p_{kk',n} P^k U^*_{k'}(0) + p_{k,n} \varepsilon \boldsymbol{e}] \delta_{t,t_n} \\
&+ \int_t^{t_{n+1}} e^{-\alpha(s-t)} P(\pi^k, t, s) [\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U^*_{k'}(0) + \varepsilon \boldsymbol{e}] ds, \\
&\qquad t \in [t_n, t_{n+1}], \;\; n \ge 0, \;\; k \in K. \qquad (6.16)
\end{aligned}$$

Again, substituting iteratively the expressions of $U_k(t_{n+1})$ into the above equation, one can get that for $k \in K$,

$$\begin{aligned}
U_k(t) \le &\int_t^{\infty} e^{-\alpha(s-t)} P(\pi^k, t, s) [\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U^*_{k'}(0) + \varepsilon \boldsymbol{e}] ds \\
&+ \sum_{t_j \ge t} e^{-\alpha(t_j - t)} P(\pi^k, t, t_j) [p_{k,j} R^k(\pi, t_j) + \sum_{k'} p_{kk',n} P^k U^*_{k'}(0) + p_{k,j} \varepsilon \boldsymbol{e}].
\end{aligned}$$

By this and Eq. (6.7), we have for $k \in K$,

$$0 \quad \le U_k(t) - U_k(\pi, t)$$

$$\begin{aligned} \leq & \sum_{t_j \geq t} e^{-\alpha(t_j - t)} P(\pi^k, t, t_j)[\sum_{k'} p_{kk',j} P^k[U^*_{k'}(0) - U_{k'}(\pi, 0)] + p_{k,j}\varepsilon \boldsymbol{e}] \\ & + \int_t^\infty e^{-\alpha(s-t)} P(\pi^k, t, s)[\sum_{k'} g_{kk'}(s) P^k[U^*_{k'}(0) - U_k(\pi, 0)] + \varepsilon \boldsymbol{e}] ds. \end{aligned} \tag{6.17}$$

If we let $\Delta = \sup_{k,i}[U_k(0,i) - U_k(\pi, 0, i)]$, then

$$\begin{aligned} 0 & \leq U_k(0) - U_k(\pi, 0) \\ & \leq \sum_{j=0}^\infty e^{-\alpha t_j} p_{k,j}(\Delta + \varepsilon)\boldsymbol{e} + \int_0^\infty e^{-\alpha s}[g_k(s)\Delta + \varepsilon] ds \cdot \boldsymbol{e} \\ & \leq \int_0^\infty e^{-\alpha s} dG_k(s) \cdot \Delta \boldsymbol{e} + \varepsilon \boldsymbol{e} + \alpha^{-1}\varepsilon \boldsymbol{e} \\ & \leq \beta \Delta e + (1 + \alpha^{-1})\varepsilon \boldsymbol{e}. \end{aligned}$$

Thus, $\Delta \leq \beta\Delta + (1 + \alpha^{-1})\varepsilon$, and so $\Delta \leq \varepsilon_1 := (1 - \beta)^{-1}(1 + \alpha^{-1})\varepsilon$. By this and Eq. (6.17), we can prove that

$$U_k(t) \leq U_k(\pi, t) + \varepsilon_1 \boldsymbol{e}, \ \ t \geq 0, \ \ k \in K. \tag{6.18}$$

Substituting it into Eq. (6.16), we obtain the result by noting (6.15) that

$$\begin{aligned} U_k(t) & \leq e^{-\alpha(t_{n+1} - t)} P(\pi^k, t, t_{n+1})[U_k(\pi, t_{n+1}) + \varepsilon_1 \boldsymbol{e}] \\ & \quad + [p_{k,n} R^k(\pi, t_n) + \sum_{k'} p_{kk',n} P^k U_{k'}(\pi, 0) + p_{k,n}\varepsilon \boldsymbol{e} + \varepsilon_1 \boldsymbol{e}]\delta_{t,t_n} \\ & \quad + \int_t^{t_{n+1}} e^{-\alpha(s-t)} P(\pi^k, t, s) \\ & \quad \cdot [\hat{r}^k(\pi, s) + \sum_{k'} g_{kk'}(s) P^k U_{k'}(\pi, 0) + \varepsilon_1 \boldsymbol{e} + \varepsilon \boldsymbol{e}] ds \\ & \leq U_k(\pi, t) + \varepsilon_2 \boldsymbol{e} \leq U^*_k(t) + \varepsilon_2 \boldsymbol{e}, \ \ t \in [t_n, t_{n+1}], \ \ n \geq 0, \ \ k \in K, \end{aligned}$$

where ε_2 is a function of ε and tends to zero when $\varepsilon \to 0^+$. Then letting $\varepsilon \to 0^+$ in the above formula one can get $U_k(t) \leq U^*_k(t)$ for all $t \geq 0$ and $k \in K$. Thus, due to (6.15),

$$U_k(t) = U^*_k(t), \ \ t \geq 0, \ \ k \in K. \tag{6.19}$$

From the definition of $U_k(t)$ we have from the above equation that for $m = 1, 2, \ldots, m_n - 1$ and $n \geq 0$,

$$\begin{aligned} \lim_{t \to t_{n,m}-} U_k(t) & = \lim_{t \to t_{n,m}-} U_{k,n,m-1}(t) = U^*_k(t_{n,m}), \\ \lim_{t \to t_{n,m}+} U_k(t) & = \lim_{t \to t_{n,m}+} U_{k,n,m}(t) = U_{k,n,m}(t_{n,m} + 0) \\ & = U_{k,n,m}(t_{n,m}) = U_k(t_{n,m}) = U^*_k(t_{n,m}). \end{aligned}$$

So, $U_k(t)$, and therefore $U_k^*(t)$, is continuous in $t_{n,m}$ for $m = 1, 2, \ldots, m_n - 1$ and $n \geq 0$.

Now, $U_k(t)$ and $U_k^*(t)$ are absolutely continuous in $(t_{n,m}, t_{n,m+1}]$ and continuous in $t_{n,m+1}$ for each $m = 1, 2, \ldots, m_n - 1$. So, they are absolutely continuous in $(t_n, t_{n+1}]$ for each $n \geq 0$. Obviously, the limits of $U_k(t)$ and $U_k^*(t)$ exist as $t \to t_{n+1} - 0$ for all $n \geq 0$.

2 and 3 in the definition of Ω (given at the beginning of this subsection) are certainly satisfied with $U_k^*(t)$. Hence, $\{U_k(t)\} = \{U_k^*(t)\} \in \Omega$.

For the uniqueness of solutions of Eqs. (6.12) and (6.13), suppose that $\{\bar{U}_k(t)\} \in \Omega$ is a solution. Then, exactly as one proves Eq. (6.19), we can prove $\bar{U}_k(t) = U_k^*(t)$ for $k \in K$ and $t \geq 0$. So, $\{U_k^*(t)\}$ is the unique solution in Ω of Eqs. (6.12) and (6.13). □

In the following theorem, U_k^* in Eq. (6.13) is replaced by U_k.

Theorem 6.1: *We have the following conclusions.*

1. *$U_k^*(t)$ is the unique solution in Ω of the following optimality equation,*

$$-\frac{d}{dt}U_k(t) = \sup_{f \in F}\{\hat{r}^k(f,t) + Q^k(f,t)U_k(t)\} - \alpha U_k(t) + \sum_{k'} g_{kk'}(t)P^k U_{k'}(0), \text{ a.e. } t, k, i. \tag{6.20}$$

$$U_k(t_n) = U_k(t_n + 0) + p_{k,n}R^k + \sum_{k'} p_{kk',n}P^k U_{k'}(0), \; n, k. \tag{6.21}$$

2. *$U_k^*(t) = \sup\{U_k(\pi, t) : \pi \in \Pi_m^d\}$ for all t and k.*

3. *For any constant $\varepsilon \geq 0$ and any policy $\pi \in \Pi_m$, if π attains ε-supremum in Eq. (6.20); that is,*

$$\begin{aligned} &\sup_{f \in F}\{\hat{r}^k(f,t) + Q^k(f,t)U_k^*(t)\} \\ \leq \;& \hat{r}^k(\pi^k, t) + Q^k(\pi, t)U_k^*(t) + \varepsilon \boldsymbol{e}, \quad \text{a.e. } t, k, i, \end{aligned}$$

and $R^k(\pi, t_{n+1}) \geq R^k - \varepsilon \boldsymbol{e}$ for all n, k, then $U_k(\pi, t) \geq U_k^*(t) - (1 - \beta)^{-1}(1 + \alpha^{-1})\varepsilon \boldsymbol{e}$ for all t, k.

Proof: 1. Due to Lemma 6.4, one knows that $U_k^*(t)$ is a solution in Ω of Eqs. (6.20) and (6.21). Thus, it suffices to prove the uniqueness of solutions in Ω. Suppose that $\{U_k(t)\} \in \Omega$ is a solution of Eqs. (6.20) and (6.21). Then, for each policy $\pi \in \Pi_m$,

$$-\frac{d}{dt}U_k(t) \geq \hat{r}^k(\pi, t) + \sum_{k'} g_{kk'}(t)P^k U_{k'}(0) + Q^k(\pi, t)U_k(t) - \alpha U_k(t)$$

for a.e. t and each k. From this one can prove similarly to Lemma 6.1 that for each $\pi \in \Pi_m$,

$$\begin{aligned} U_k(t) &\geq e^{-\alpha(t_{n+1}-t)}P(\pi^k,t,t_{n+1})U_k(t_{n+1}) \\ &\quad + [p_{k,n}R^k(\pi,t_n) + \sum_{k'} p_{kk',n}P^k U_{k'}(0)]\delta_{t,t_n} \\ &\quad + \int_t^{t_{n+1}} e^{-\alpha(s-t)}P(\pi^k,t,s)[\hat{r}^k(\pi,s) + \sum_{k'} g_{kk'}(s)P^k U_{k'}(0)]ds, \end{aligned}$$

for $t \in [t_n, t_{n+1}]$ and n, k. For each given policy $\pi \in \Pi_m$, let $\Delta_k(t) = U_k(t) - U_k(\pi,t)$. Then, from Eq. (6.11) and the above inequality, we have

$$\begin{aligned} \Delta_k(t) &\geq e^{-\alpha(t_{n+1}-t)}P(\pi^k,t,t_{n+1})\Delta_k(t_{n+1}) + \sum_{k'} p_{kk',n}P^k\Delta_{k'}(0)\delta_{t,t_n} \\ &\quad + \int_t^{t_{n+1}} e^{-\alpha(s-t)}P(\pi^k,t,s)\sum_{k'} g_{kk'}(s)P^k\Delta_{k'}(0)ds, \\ &\qquad t \in [t_n,t_{n+1}], \;\; n \geq 0, \;\; k \in K. \end{aligned} \tag{6.22}$$

By using the induction method based on the above inequality, one can prove similarly to Lemma 6.1 that

$$\begin{aligned} \Delta_k(t_n) &\geq \sum_{m=n}^{\infty} e^{-\alpha(t_m-t_n)}P(\pi^k,t_n,t_m)\sum_{k'} p_{kk',m}P^k\Delta_{k'}(0) \\ &\quad + \int_{t_n}^{\infty} e^{-\alpha(s-t_n)}P(\pi^k,t_n,s)\sum_{k'} g_{kk'}(s)P^k\Delta_{k'}(0)ds \end{aligned} \tag{6.23}$$

for $n \geq 0$ and $k \in K$. Let $\Delta = \inf_{k,i}\Delta_k(0,i)$. Then

$$\begin{aligned} \Delta &\geq \sum_{m=0}^{\infty} e^{-\alpha t_m}p_{k,m}\Delta + \int_0^{\infty} e^{-\alpha s}g_k(s)\Delta ds \\ &= \int_0^{\infty} e^{-\alpha s}dG_k(s)\cdot\Delta. \end{aligned}$$

If $\Delta \leq 0$, then $\Delta \geq \beta\Delta$. Because $\beta \in (0,1)$, we know $\Delta = 0$. So, $\Delta \geq 0$. Due to Eq. (6.23), $\Delta_k(t_n) \geq 0$ for all $n \geq 0$. Moreover, $\Delta_k(t) \geq 0$ from Eq. (6.22) for all $k \in K$ and $t \geq 0$. That is, $U_k(t) \geq U_k(\pi,t)$ for all $k \in K$ and $t \geq 0$. Due to the arbitrariness of π, we get $U_k(t) \geq U_k^*(t)$ for all $k \in K$ and $t \geq 0$.

On the other hand, for each $\varepsilon \geq 0$ and policy $\pi \in \Pi_m^d$ that satisfies the condition given in 3, it can be proved similarly to Eq. (6.18) that

$$U_k^*(t) \geq U_k(\pi,t) \geq U_k(t) - (1-\beta)^{-1}(1+\alpha^{-1})\varepsilon.$$

Moreover, due to the arbitrariness of ε, we know that $U_k^*(t) \geq U_k(t)$ for all $k \in K$ and $t \geq 0$.

Thus, $U_k(t) = U_k^*(t)$ for all $k \in K$ and $t \geq 0$. □

In Theorem 6.1, Conclusion 3 shows that any policy achieving ε-supremum in the optimality equation is $(1-\beta)^{-1}(1+\alpha^{-1})\varepsilon$-optimal. Such policies can be deterministic Markov due to Lemma 4.11. So, the optimality can be achieved in a smaller policy set Π_m^d.

1.3 Approximation by Weak Convergence

In the optimality equation (6.20), the time variable t is involved. Hence, the equation is complex and may be difficult to be solved. A method to deal with this problem is approximation. In this subsection, we discuss the error bounds of $U_k(\pi, t)$ and $U_k^*(t)$ when the distribution $G_k(t)$ is approximated by $F_k(t)$.

We assume that CTMDP$_k$ in Eq. (6.2) is stationary; that is, $q_{ij}^k(t,a) = q_{ij}^k(a)$ and $r^k(t,i,a) = r^k(i,a)$ are independent of t for all $(i,a) \in \Gamma$ and $j \in S$. Moreover, we assume that $\gamma_{kk'} = 1$; that is, the environment state transition is of continuous type.

First, we introduce some notations. For each policy $\pi \in \Pi_m$, we denote

$$\begin{aligned}
U_k^{(1)}(\pi,t) &= \int_t^\infty e^{-\alpha(s-t)} P(\pi^k,t,s) r^k(\pi,s)[1-G_k(s)]ds, \\
U_k^{(2)}(\pi,t) &= \int_t^\infty e^{-\alpha(s-t)} P(\pi^k,t,s) R^k(\pi,s) g_k(s) ds, \\
U_k^{(3)}(\pi,t) &= \int_t^\infty e^{-\alpha(s-t)} P(\pi^k,t,s) \sum_{k'} g_{kk'}(s) P^k U_{k'}(\pi,0) ds.
\end{aligned}$$

From Eq. (6.7), we know that the above $U_k^{(1)}(\pi,t), U_k^{(2)}(\pi,t)$, and $U_k^{(3)}(\pi,t)$ are three different parts in $U_k(\pi,t)$; that is,

$$U_k(\pi,t) = U_k^{(1)}(\pi,t) + U_k^{(2)}(\pi,t) + U_k^{(3)}(\pi,t), \quad k \in K, \quad t \geq 0.$$

For any two functions $G(t)$ and $F(t)$, defined on $[0,\infty)$, we define

$$\lambda(G,F) = \sup_{t \geq 0} |G(t) - F(t)|$$

as a measure for the distance between G and F.

Suppose $\{F_{kk'}(t), k, k' \in K\}$ is another kernel of the environment process, and both $\{F_{kk'}(t), k, k' \in K\}$ and the original kernel $\{G_{kk'}(t), k, k' \in K\}$ are absolutely continuous and satisfy Condition 6.1 with the same constants θ and δ. Then, it is easy to see that

$$\lambda(G_k, F_k) \leq \sum_{k'} \lambda(G_{kk'}, F_{kk'}).$$

Denote by $V_k^{(m)}(\pi,t)$ the value of $U_k^{(m)}(\pi,t)$, $m=1,2,3$, when $\{G_{kk'}(t)\}$ is replaced by $\{F_{kk'}(t)\}$, and $V_k(\pi,t), V_k^*(t)$ similarly.

Theorem 6.2: *Suppose that $R^k(i,a)=R^k(i)$ is independent of the action a, and constants M_r, M_R, and ω are bounds of $r^k(i,a)$, $R^k(i)$, and $q_{ij}^k(a)$, respectively. Then, for each policy $\pi\in\Pi_m$ and environment state $k\in K$,*

$$\sup_{t,i}|V_k(\pi,t,i)-U_k(\pi,t,i)|\le(2-\beta)(1-\beta)^{-1}d^*\sum_{k'}\lambda(G_{kk'},F_{kk'}),$$

$$\sup_{t,i}|V_k^*(t,i)-U_k^*(t,i)|\le(2-\beta)(1-\beta)^{-1}d^*\sum_{k'}\lambda(G_{kk'},F_{kk'}),$$

where $d^*=\alpha^{-1}(3+2\alpha^{-1}\omega)M_r+2(1+\alpha^{-1}\omega)(2-\beta)(1-\beta)^{-1}M_R$.

Proof: It suffices to prove the first inequality above. First,

$$\begin{aligned}
&|V_k^{(1)}(\pi,t)-U_k^{(1)}(\pi,t)|\\
\le\ &\left|\int_t^\infty e^{-\alpha(s-t)}P(\pi^k,t,s)r^k(\pi,s)[G_k(s)-F_k(s)]ds\right|\\
\le\ &\int_t^\infty e^{-\alpha(s-t)}M_r\lambda(G_k,F_k)ds\cdot\boldsymbol{e}\\
\le\ &\alpha^{-1}M_r\lambda(G_k,F_k)\cdot\boldsymbol{e}.
\end{aligned}$$

Second,

$$\begin{aligned}
&U_k^{(2)}(\pi,t)=\int_t^\infty e^{-\alpha(s-t)}P(\pi^k,t,s)R^k dG_k(s)\\
=\ &-G_k(t)R^k-\int_t^\infty G_k(s)e^{-\alpha(s-t)}P(\pi^k,t,s)[Q^k(\pi,s)-\alpha I]R^k ds.
\end{aligned}$$

Then,

$$\begin{aligned}
&|V_k^{(2)}(\pi,t)-U_k^{(2)}(\pi,t)|\\
\le\ &\lambda(G_k,F_k)M_R\cdot\boldsymbol{e}+\int_t^\infty e^{-\alpha(s-t)}(\alpha+2\omega)M_R ds\cdot\lambda(G_k,F_k)\boldsymbol{e}\\
\le\ &2(1+\alpha^{-1}\omega)M_R\lambda(G_k,F_k)\boldsymbol{e}.
\end{aligned}$$

Third,

$$\begin{aligned}
&V_k^{(3)}(\pi,t)-U_k^{(3)}(\pi,t)\\
=\ &\int_t^\infty e^{-\alpha(s-t)}P(\pi^k,t,s)P^k\sum_{k'}f_{kk'}(s)[V_{k'}(\pi,0)-U_{k'}(\pi,0)]ds\\
&+\int_t^\infty e^{-\alpha(s-t)}P(\pi^k,t,s)P^k\sum_{k'}[f_{kk'}(s)-g_{kk'}(s)]U_{k'}(\pi,0)ds\\
:=\ &I_1+I_2.
\end{aligned}$$

Let $\Delta := \sup_{k,i} |V_k(\pi, 0, i) - U_k(\pi, 0, i)|$. Then,

$$|I_1| \leq \int_t^\infty e^{-\alpha(s-t)} \sum_{k'} f_{kk'}(s) ds \cdot \Delta \cdot \boldsymbol{e} = \int_t^\infty e^{-\alpha(s-t)} dF_k(s) \cdot \Delta \cdot \boldsymbol{e}.$$

Moreover,

$$\begin{aligned} I_2 &= \int_t^\infty e^{-\alpha(s-t)} P(\pi^k, t, s) P^k d_s \{ \sum_{k'} [F_{kk'}(s) - G_{kk'}(s)] \} U_{k'}(\pi, 0) \\ &= P^k \sum_{k'} [G_{kk'}(t) - F_{kk'}(t)] U_{k'}(\pi, 0) \\ &\quad + \int_t^\infty e^{-\alpha(s-t)} P(\pi^k, t, s) [Q^k(\pi, s) - \alpha I] \\ &\quad \cdot P^k \sum_{k'} [G_{kk'}(s) - F_{kk'}(s)] ds U_{k'}(\pi, 0). \end{aligned}$$

So,

$$\begin{aligned} |I_2| &\leq \sum_{k'} \lambda(G_{kk'}, F_{kk'}) M_U \boldsymbol{e} \\ &\quad + \int_t^\infty e^{-\alpha(s-t)} (\alpha + 2\omega) M_U ds \cdot \sum_{k'} \lambda(G_{kk'}, F_{kk'}) \\ &\leq 2(1 + \alpha^{-1}\omega) M_U \sum_{k'} \lambda(G_{kk'}, F_{kk'}) \boldsymbol{e}, \end{aligned}$$

where M_U is the bound of $U_k(\pi, t)$ and may be taken as $M_U = \alpha^{-1} M_r + (1 - \beta)^{-1} M_R$.

Hence, for each t, k, i,

$$|V_k(\pi, t, i) - U_k(\pi, t, i)| \leq \int_t^\infty e^{-\alpha(s-t)} dF_k(s) \cdot \Delta + d^* \sum_{k'} \lambda(G_{kk'}, F_{kk'}).$$

By letting $t = 0$ above we get that $\Delta \leq \beta\Delta + d^* \sum_{k'} \lambda(G_{kk'}, F_{kk'})$, and so

$$\Delta \leq (1 - \beta)^{-1} d^* \sum_{k'} \lambda(G_{kk'}, F_{kk'}).$$

This completes the proof. □

Now, suppose that absolutely continuous d.f.s $G_{kk',n}(t)$ converges weakly to $G_{kk'}(t)$ as n tends to infinity for each k, k'. Then, because $G_{kk',n}(t)$ and $G_{kk'}(t)$ are all absolutely continuous, $G_{kk',n}(t)$ converges uniformly to $G_{kk'}(t)$; that is,

$$\lim_{n \to \infty} \lambda(G_{kk',n}, G_{kk'}) = 0, \quad k, k' \in K.$$

For $n \geq 0$, we define $U_{k,n}(\pi, t)$ as $U_k(\pi, t)$ except that $\{G_{kk'}(t)\}$ is replaced by $\{G_{kk',n}(t)\}$. So, due to Theorem 6.2, $U_{k,n}(\pi, t, i)$ converges to $U_k(\pi, t, i)$ uniformly in π, t, k, i as n tends to infinity.

From probability theory, we know that any d.f. can be approximated by a phase type d.f. (see, e.g., Neuts [98]). Hence, in the following two subsections, we study the case where the environments are Markov and phase type, respectively.

1.4 Markov Environment

In this subsection, we assume also that the CTMPD$_k$ is stationary and $\gamma_{kk'} = 1$ for all $k, k' \in K$. We consider the case where the environment is Markov. That is, there exists a bounded and reserve transition rate family, denoted by $T = (T_{kk'})_{k,k' \in K}$, such that

$$G_k(t) = 1 - e^{T_{kk}t}, \;\; G_{kk'}(t) = \psi_{kk'} G_k(t), \;\; \psi_{kk'} = \frac{T_{kk'}}{T_{kk}}(1 - \delta_{kk'}). \quad (6.24)$$

Then, by letting $r(k, i, a) = r^k(i, a) - T_{kk} R^k(i, a)$ we have

$$\hat{r}^k(t, i, a) = r(k, i, a) e^{T_{kk}t}. \quad (6.25)$$

Substituting it into Eq. (6.7), one can get that for each policy $\pi \in \Pi_m$ and $t \geq 0, k \in K$,

$$\begin{aligned} U_k(\pi, t) &= e^{T_{kk}t} \int_t^\infty e^{(T_{kk} - \alpha)(s-t)} P(\pi^k, t, s) \\ &\quad \cdot [r(k, \pi, s) + \sum_{k' \neq k} T_{kk'} P^k U_{k'}(\pi, 0)] ds, \end{aligned} \quad (6.26)$$

where the vector $r(k, \pi, s)$ is defined similarly to $r^k(\pi, s)$. For any Markov stationary policy $\pi \in \Pi_s$, $r(k, \pi, s) = r(k, \pi, 0)$ is independent of s, and is denoted by $r(k, \pi)$.

In the following theorem, both the criterion $U_k(\pi, t)$ and the optimal value, and therefore the optimality equation, are simplified.

Theorem 6.3: *Suppose that the environment is Markov.*

1. *For any Markov stationary policy $\pi = (\pi_0^k) \in \Pi_s$ and environment state $k \in K$, $e^{-T_{kk}t} U_k(\pi, t)$ is independent of t, denoted by $U_k(\pi)$, and*

$$U_k(\pi) = \int_0^\infty e^{(T_{kk} - \alpha)s} P(\pi_0^k, s)[r(k, \pi_0^k) + \sum_{k' \neq k} T_{kk'} P^k U_{k'}(\pi)] ds. \quad (6.27)$$

2. *$e^{-T_{kk}t} U_k^*(t)$ is independent of t, denoted by U_k^*, and is the unique bounded solution of the following optimality equation,*

$$(\alpha - T_{kk}) U_k(i) = \sup_{a \in A(i)} \{r(k, i, a) + \sum_j q_{ij}^k(a) U_k(j)\}$$

$$+\sum_{k'\neq k} T_{kk'} \sum_j p_{ij}^k U_{k'}(j), \quad i \in S, \quad k \in K. \tag{6.28}$$

Moreover, $U_k^* = \sup\{U_k(\pi, 0) : \pi = (f^k) \in \Pi_s^d\}$ for all $k \in K$; that is, the optimal value can be achieved among stationary policies.

3. *For $\varepsilon \geq 0$, if $\pi = (f^k) \in \Pi_s^d$ attains the ε-supremum in Eq. (6.28), then π is $(1-\beta)^{-1}(1+\alpha^{-1})\varepsilon$-optimal.*

Proof: 1. For a stationary policy π, $P(\pi^k, t, s) = P(\pi^k, 0, s-t)$ and $r(k, \pi, s) = r(k, \pi)$ for all s and t with $0 \leq t \leq s$. Then, the results follow immediately from Eq. (6.26).

2. Substituting Eqs. (6.24) and (6.25) into the optimality equation (6.20), due to

$$g_{kk'}(t) = \psi_{kk'} g_k(t) = -T_{kk'}(1 - \delta_{kk'})e^{T_{kk}t},$$

we get that the optimality equation (6.20) is equivalent to the following equation,

$$\begin{aligned} -\frac{d}{dt}U_k(t,i) &= \sup_{a\in A(i)} \{r(k,i,a)e^{T_{kk}t} + \sum_j q_{ij}^k(a) U_k(t,j)\} \\ &\quad + \sum_{k'\neq k} T_{kk'} e^{T_{kk}t} \sum_j p_{ij}^k U_{k'}(0,j) - \alpha U_k(t,i), \quad \text{a.e. } t, k, i. \end{aligned}$$

This is again equivalent to the following equation,

$$\begin{aligned} -\frac{d}{dt}[U_k(t,i)e^{-T_{kk}t}] &= \sup_{a\in A(i)} \{r(k,i,a) + \sum_j q_{ij}^k(a) U_k(t,j) e^{-T_{kk}t}\} \\ + \sum_{k'\neq k} T_{kk'} \sum_j p_{ij}^k U_{k'}(0,j) &+ (T_{kk} - \alpha) U_k(t,i) e^{-T_{kk}t}, \text{ a.e. } t, k, i. \end{aligned} \tag{6.29}$$

It follows Theorem 6.1 that $\{e^{-T_{kk}t}U_k^*(t)\}$ is the unique solution of the above equation in $\Omega' := \{(e^{-T_{kk}t}v_k(t)) : (v_k(t)) \in \Omega\}$. Moreover, it is easy to see that for any $t_0 \geq 0$, $\{e^{-T_{kk}(t+t_0)}U_k^*(t+t_0)\}$ also belongs to the set Ω' and is also a solution of Eq. (6.29). By the uniqueness of the solutions of Eq. (6.29) and the arbitrariness of t_0, we know that $e^{-T_{kk}t}U_k^*(t)$ is independent of t. We denote it by $U_k^* = e^{-T_{kk}t}U_k^*(t) = U_k^*(0)$. Then, the right-hand side of Eq. (6.29) is zero. Therefore, this equation is exactly the equation (6.28), and U_k^* is the unique bounded solution of Eq. (6.28).

3. It follows 1, 2 and Theorem 6.1. □

Based on Theorem 6.3, the CTMDP model in a Markov environment can be transformed into the following DTMDP model,

$$\{S', A'(i'), r'(i', a), p(j'|i', a), V_{\beta_*}\}. \tag{6.30}$$

Here, the state space is $S' = \{(k,i) : k \in K, i \in S\}$, where state (k,i) means that the inner state is i and the environment state is k. The action set available at state (k,i) is $A'(k,i) = A(i)$, which is irrespective of the environment state k. The reward at state (k,i) when using action a is $r'(k,i,a) = r(k,i,a)/(\alpha + \omega - T_{kk})$, where ω is a upper bound of $-q_{ii}^k(a)$. The state transition probability from state (k,i) to state (k',j) under action a is

$$p(k',j|k,i,a) = \begin{cases} \frac{\beta_*^{-1} T_{kk'} p_{ij}^k}{\alpha+\omega-T_{kk}}, & k' \neq k \\ \frac{\beta^{-1}[q_{ij}^k(a)+\omega\delta_{ij}]}{\alpha+\omega-T_{kk}}, & k' = k, \end{cases}$$

where T^* is a upper bound of $-T_{kk}$, and the discount factor is $\beta_* = (\omega + T^*)/(\alpha + \omega + T^*)$. The discounted criterion is denoted by $V_{\beta_*}(\pi, k, i)$.

Comparing the DTMDP model given in Eq. (6.30) with the CTMDP model in a semi-Markov environment given in Eq. (6.1), we know that they have the same Markov stationary policy set Π_s and the same stationary policy set Π_s^d.

Theorem 6.4: *The CTMDP model in a Markov environment is equivalent to the DTMDP model given in Eq. (6.30) in the following manner.*

1. *For each Markov stationary policy $\pi = (\pi_0^k) \in \Pi_s$, $U_k(\pi, i) = V_\beta(\pi, k, i)$ for all $(k,i) \in S$.*
2. *Both optimality equations of the two models are Eq. (6.28).*

Proof: 1. For $\pi = (\pi_0^k) \in \Pi_s$, it follows from Eq. (6.27) that for $k \in K$,

$$U_k(\pi) = [(\alpha - T_{kk})I - Q^k(\pi_0^k)]^{-1}\{r(k, \pi_0^k) + \sum_{k' \neq k} T_{kk'} P^k U_{k'}(\pi)\}.$$

Premultiplying the above formula by $[(\alpha - T_{kk})I - Q^k(\pi_0^k)]$, one can get that $\{U_k(\pi)\}$ is the unique bounded solution of the following equation,

$$(\alpha - T_{kk})U_k = r(k, \pi_0^k) + \sum_{k' \neq k} T_{kk'} P^k U_{k'} + Q^k(\pi_0^k)U_k, \;\; k \in K.$$

Dividing both sides of the above equation by $\alpha + \omega - T_{kk}$ and rearranging it, we get that

$$U_k(i) = \sum_{a \in A(i)} \pi_0^k(a|i)\{r'(k,i,a) + \beta_* \sum_{k',j} p(k',j|k,i,a)U_{k'}(j)\}, \;\; (k,i) \in S'.$$

But from Theorem 2.4, the unique bounded solution of the above equation is $V_{\beta_*}(\pi, k, i)$. Hence, $U_k(\pi, i) = V_\beta(\pi, k, i)$ for all $(k,i) \in S$.

2. From Theorems 2.4 and 2.6, $V_{\beta_*}^*(k,i) = \sup\{V_\beta(\pi, k, i) : \pi \in \Pi_m\} = \sup\{V_\beta(\pi, k, i) : \pi \in \Pi_s^d\}$ is the unique bounded solution of the following

optimality equation,

$$V_\beta(k,i) = \sup_{a\in A(i)} \{r(k,i,a) + \beta_* \sum_{k',j} p(k',j|k,i,a) V_{\beta_*}(k',j)\}, \quad (k,i) \in S'.$$

As that in 1 above, the proceeding equation is equivalent to Eq. (6.28). □

From Theorem 6.4, one can directly generalize most results in DTMDPs into CTMDPs in a Markov environment for the discounted criterion. It should be noted that CTMDPs in a Markov environment differ from the standard CTMDPs, where there are no instantaneous rewards $R^k(i,a)$ and no instantaneous state transition (p_{ij}^k).

1.5 Phase Type Environment

In this subsection, we assume that the distribution function $G_k(t)$ are all phase type; that is, for each $k \in K$, there exists a Markov infinitesimal generator $(m_k+1)\times(m_k+1)$-matrix

$$Q_k = \begin{pmatrix} T_k & T_k^0 \\ 0 & 0 \end{pmatrix},$$

where $T_k = (T_{mm'}^k)$ is an $m_k \times m_k$ matrix and $T_k^0 = (T_{k,m}^0)$ is an m_k column vector that satisfies ($\boldsymbol{e}_k$ is a unity column vector):

$$T_{mm}^k < 0, \quad T_{mm'}^k \geq 0, \quad m,m' = 1,2,\ldots,m_k, \quad m \neq m',$$

and $T_{k,m}^0 = -\sum_{m'} T_{mm'}^k$ (i.e., $T_k \boldsymbol{e}_k + T_k^0 = 0$). Moreover, $(\gamma_k, \gamma_{k,m_k+1})$ is the initial state probability vector with $\gamma_k = (\gamma_{k,1}, \gamma_{k,2}, \ldots, \gamma_{k,m_k})$.

Having the above Q_k and $(\gamma_k, \gamma_{k,m_k+1})$, the kernel of the environment with phase type can be described by

$$G_{kk'}(t) = \psi_{kk'} G_k(t), \quad \psi_{kk'} \geq 0, \quad \sum_{k'} \psi_{kk'} = 1, \tag{6.31}$$

$$G_k(t) = 1 - \gamma_k \boldsymbol{e}_k e^{T_k t}, \quad g_k(t) = \gamma_k T_k^0 e^{T_k t}.$$

For convenience, we denote hereafter by $T_m^k = -T_{m,m}^k$ for $m = 1,2,\ldots,m_k$ and by $M = \{1,2,\ldots,m_k\}$ and $M' = \{1,2,\ldots,m_k+1\}$ two sets.

In order to ensure that $G_k(t)$ is a distribution function, we assume that the matrix T_k is nonsingular (see Neuts [98]). Thus, states $1,2,\ldots,m_k$ are all transient. We further assume that $\gamma_{k,m_k+1} = 0$. Now, we augment the environment state by introducing the phase in it. We say that the (augmented) environment state is (k,m) if the environment state is k and the phase is m for $k \in K$ and $m = 1,2,\ldots,m_k$. So, the augmented environment is Markov. The distribution function and probability distribution function of the time at (k,m) are, respectively,

$$G_{k,m}(t) = 1 - e^{-T_m^k t}, \quad g_{k,m}(t) = T_m^k e^{-T_m^k t} \tag{6.32}$$

for $m = 1, 2, \ldots m_k$ and $k \in K$. It should be noted that $G_{k,m}(t)$ differs from $G_{kk'}(t)$.

In order to compute the transition probabilities of the environment, we introduce a dummy state (k, m_k+1), at which the sojourn time is zero. Thus, the transition probability of the augmented environment state is (with $T^k_{m,m_k+1} = T^0_{k,m}$)

$$\psi(k', m'|k, m) = \begin{cases} \frac{T^k_{m,m'}}{T^k_m}\delta_{k,k'}, & m \in M, m' \neq m \\ \psi_{k,k'}\gamma_{k',m'}, & m = m_k + 1, m' \in M'. \end{cases} \tag{6.33}$$

Obviously, the augmented environment process satisfies the property: there occur only finitely many transitions in any finite time interval with probability one. Thus, the instantaneous transition probability of inner states at epochs of transitions of the augmented environment states is

$$p^{k,m}_{ij} = \begin{cases} \delta_{ij}, & m \in M \\ p^k_{ij}, & m = m_k + 1, \end{cases} \tag{6.34}$$

or equivalently, $P^{k,m} = I$ for $k \in M$ and $P^{k,m_k+1} = P^k$.

Now, the model of CTMDPs in the (Markov) augmented environment is

$$\{(K', G), S, A(i), q^{k,m}, r^{k,m}, p^{k,m}, R^{k,m}, U, (k, m) \in K'\}. \tag{6.35}$$

Here, the state set of the environment is $K' = \{(k, m) : k \in K, m \in M'\}$, $G, \psi, p^{k,m}$ are defined, respectively, in Eqs. (6.32), (6.33), (6.34), and

$$q^{k,m}_{ij}(a) = q^k_{ij}(a), \;\; r^{k,m}(i, a) = r^k(i, a), \;\; R^{k,m}(i, a) = R^k(i, a)\delta_{m,m_k+1}.$$

The policy set is denoted by $\Pi_m(p)$, where the notation "(p)" is used to distinguish the policy set Π_m of the original model (6.1). The policy sets of $\Pi^m_d(p)$, $\Pi_s(p)$, and $\Pi^d_s(p)$ are similar. The criterion $U_{k,m}(\pi, t)$ is defined as the expected discounted total reward similarly to Eq. (6.5).

Although the model Eq. (6.35) is well defined, nevertheless for

$$G_{k,m+1}(t) = 1, \;\; t \geq 0, k \in K,$$

Condition 6.1 does not hold. So, the results in the previous subsections cannot be used directly here. But fortunately the method used in the previous subsections is adequate here and the same results can be proved similarly. So, the detailed proof is omitted in the following.

By noting that $P^{k,m} = I$ for $m \in M$, it can be proved similarly to Eq. (6.7) that

$$U_{k,m}(\pi, t) \;\; = \;\; \int_t^\infty e^{-\alpha(s-t)} P(\pi^{k,m}, t, s)$$

$$
\begin{aligned}
&\cdot \{r^{k,m}(\pi,s)\bar{G}_{k,m}(s) + g_{k,m}(s)R^{k,m}(\pi,s) \\
&+ g_{k,m}(s)\sum_{k',m'}\psi(k',m'|k,m)U_{k',m'}(\pi,0)\}ds, \quad m \in M, \\
U_{k,m_k+1}(\pi,t) &= \delta_{t,0}[R^k(\pi) + \sum_{k',m'}\psi(k',m'|k,m)P^k U_{k',m'}(\pi,0)],
\end{aligned}
$$

where $R_i^k(\pi) = \sum_{a\in A(i)} R^k(i,a)\pi_0^{k,m_k+1}(a|i)$ for $i \in S$. Substituting the latter into the former, and by the definition of the elements in Eq. (6.35), one can get that for $m \in M$,

$$
\begin{aligned}
U_{k,m}(\pi,t) &= \int_t^\infty e^{-\alpha(s-t)}P(\pi^{k,m},t,s) \\
&\quad \cdot \{r^{k,m}(\pi,s)e^{-T_m^k s} + T_m^k e^{-T_m^k s}R^{k,m}(\pi,s) \\
&\quad + e^{-T_m^k s}\sum_{m'=1,m'\neq m}^{m_k} T_{m,m'}^k U_{k,m'}(\pi,0) \\
&\quad + e^{-T_m^k s}T_{k,m}^0 U_{k,m_k+1}(\pi,0)\}ds \\
&= \int_t^\infty e^{-\alpha(s-t)}P(\pi^{k,m},t,s)e^{-T_m^k s} \\
&\quad \cdot \{r^{k,m}(\pi,s) + T_{k,m}^0 R^k(\pi) + \sum_{m'=1,m'\neq m}^{m_k} T_{m,m'}^k U_{k,m'}(\pi,0) \\
&\quad + T_{k,m}^0 \sum_{k'}\sum_{m'=1}^{m_k}\psi_{kk'}\gamma_{k'm'}P^k U_{k',m'}(\pi,0)\}ds. \qquad (6.36)
\end{aligned}
$$

It is easy to prove that $U_{k,m}(\pi,t)$ is uniformly bounded. Define Ω_p as the set of $v = (v_{k,m}(t,i) : k \in K, m = 1,2,\ldots,m_k, i \in S, t \geq 0)$ that satisfies the following two conditions.

1. For each k,m,i, $v_{k,m}(t,i)$ is absolutely continuous in $t \in [0,\infty)$.
2. $v_{k,m}(t,i)$ is uniformly bounded in k,m,t,i.

Obviously, for each policy $\pi \in \Pi_m(p)$, its objective function $\{U_{k,m}(\pi,t)\}$ belongs to the set Ω_p.

The following lemma can be proved exactly as Lemma 6.1.

Lemma 6.5: *For any policy* $\pi = (\pi_t^{k,m}) \in \Pi_m(p)$, *constant* $\varepsilon \geq 0$, *and* $v = (v_{k,m}(t,i)) \in \Omega_p$,
1. *If for* $k \in K, m \in M$, *and a.e.* $t \geq 0$,

$$
\begin{aligned}
e^{T_m^k t}[-\frac{d}{dt}v_{k,m}(t)] &\leq r^{k,m}(\pi,t) + T_{k,m}^0 R^k(\pi) + Q^{k,m}(\pi,t)v_{k,m}(t) \\
&\quad - \alpha v_{k,m}(t) + \sum_{m'=1,m'\neq m}^{m_k} T_{m,m'}^k v_{k,m'}(0)
\end{aligned}
$$

$$+ T^0_{k,m} \sum_{k'} \sum_{m'=1}^{m_k} \psi_{kk'} \gamma_{k'm'} P^k v_{k',m'}(0) + \varepsilon \boldsymbol{e},$$

then $v_{k,m}(t) \le U_{k,m}(\pi, t) + (2-\beta)(1-\beta)^{-1}\alpha^{-1}\varepsilon \boldsymbol{e}$ *for all* t *and* k.

2. *If* $k \in K, m \in M$, *and a.e.* $t \ge 0$,

$$\begin{aligned} e^{T^k_m t}[-\frac{d}{dt} v_{k,m}(t)] \;\ge\; & r^{k,m}(\pi, t) + T^0_{k,m} R^k(\pi) + Q^{k,m}(\pi, t) v_{k,m}(t) \\ & - \alpha v_{k,m}(t) + \sum_{m'=1, m' \ne m}^{m_k} T^k_{m,m'} v_{k,m'}(0) \\ & + T^0_{k,m} \sum_{k'} \sum_{m'=1}^{m_k} \psi_{kk'} \gamma_{k'm'} P^k v_{k',m'}(0) - \varepsilon \boldsymbol{e}, \end{aligned}$$

then $v_{k,m}(t) \ge U_{k,m}(\pi, t) - (2-\beta)(1-\beta)^{-1}\alpha^{-1}\varepsilon \boldsymbol{e}$ *for all* t *and* k.

3. $U_{k,m}(\pi, t)$ *is the unique solution in* Ω_p *of the following equation,*

$$\begin{aligned} e^{T^k_m t}[-\frac{d}{dt} v_{k,m}(t)] \;=\; & r^{k,m}(\pi, t) + T^0_{k,m} R^k(\pi) + Q^{k,m}(\pi, t) v_{k,m}(t) \\ & - \alpha v_{k,m}(t) + \sum_{m'=1, m' \ne m}^{m_k} T^k_{m,m'} v_{k,m'}(0) \\ & + T^0_{k,m} \sum_{k'} \sum_{m'=1}^{m_k} \psi_{kk'} \gamma_{k'm'} P^k v_{k',m'}(0), \\ & \text{a.e. } t \ge 0, \quad k \in K, \quad m \in M. \end{aligned}$$

From the above discussion, the following theorem can be obtained similarly to Theorem 6.3 about the Markov environment.

Theorem 6.5:

1. *For any Markov stationary policy* $\pi = (\pi_0^{k,m}) \in \Pi_s(p)$, $e^{T^k_m t} U_{k,m}(\pi, t)$ *is independent of* t, *denoted by* $U_{k,m}(\pi)$, *and for* $m \in M$ *and* $k \in K$,

$$\begin{aligned} U_{k,m}(\pi) \;=\; & \int_0^\infty e^{-(T^k_m + \alpha)s} P(\pi^{k,m}, 0, s) \\ & \cdot \{ r^{k,m}(\pi) + T^0_{k,m} R^k(\pi) + \sum_{m'=1, m' \ne m}^{m_k} T^k_{m,m'} U_{k,m'}(\pi, 0) \\ & + T^0_{k,m} \sum_{k'} \sum_{m'=1}^{m_k} \psi_{kk'} \gamma_{k'm'} P^k U_{k',m'}(\pi, 0) \} ds. \end{aligned}$$

2. *$e^{T_m^k t}U_{k,m}^*(t)$ is independent of t, denoted by $U_{k,m}^*$, and is the unique bounded solution of the following optimality equation,*

$$\begin{aligned}(\alpha + T_m^k)U_{k,m}(i) &= \sup_{a\in A(i)}\{r^k(i,a) + \sum_j q_{ij}^k(a)U_{k,m}(j)\} \\ &\quad + T_{k,m}^0 R^k(i) + \sum_{m'=1,m'\neq m}^{m_k} T_{m,m'}^k U_{k,m'}(i) \\ &\quad + T_{k,m}^0 \sum_{k'}\sum_{m'=1}^{m_k} \psi_{kk'}\gamma_{k',m'}\sum_j p_{ij}^k U_{k',m'}(j) \quad (6.37)\end{aligned}$$

for m, k, i, and $U_{k,m}^* = \sup\{U_{k,m}(\pi) : \pi \in \Pi_s^d(p)\}$ for all k and m.

3. *For $\varepsilon \geq 0$, if $\pi \in \Pi_s^d(p)$ attains the ε-supremum in the right-hand side of Eq. (6.37), then π is $(2-\beta)(1-\beta)^{-1}\alpha^{-1}\varepsilon$ optimal.*

As in the previous sections for the Markov environment, based on the above theorem, the CTMDP model in a phase-type environment can be also transformed into an equivalent DTMDP model with state (k, m, i). Please do this as an exercise.

The following remark gives several special cases of the CTMDP model in a semi-Markov environment.

Remark 6.2: *When $K = \{0, 1, 2, \ldots\}$ and $G_{k,k+1}(t) = G_k(t)$, the model (6.1) becomes the shock CTMDP model, discussed in [59]. When $G_k(t) \equiv 0$ for all $t < \infty$ and $k \in K$, the model becomes the usual CTMDP without considering environments. Moreover, when $p_{ij}^k \equiv 0$ the model becomes CTMDP with random horizon L_1.*

At the end of this section, we consider an example.

Example 6.1 Optimal service rate control of a queueing system M/M/1 in a semi-Markov environment.

Consider a queueing system M/M/1 in a semi-Markov environment $\{(J_n, L_n), n \geq 0\}$ which is the same as that given in Eq. (6.1). Suppose that $G_{kk'}(t) = \psi_{kk'}G_k(t)$ and $G_k(t)$ is absolutely continuous with the p.d.f. $g_k(t)$. When the environment state is $k \in K$, the customers arrive at the system according to a Poisson process with rate λ_k, and there is only one customer at each arrival. The service time of each customer is geometrically distributed with parameter μ. Assume that μ is chosen from a countable set $A \subset [0, \infty)$ at any time and all random variables are mutually independent.

We say that the system is in state i if there are i customers in the system. If μ is chosen while state is i and the environment state is k, then the cost rate is $c^k(i, \mu)$. The instantaneous cost, denoted by $R^k(i, \mu)$, occurs when μ is chosen while the state is i and the environment state is k. When $i = 0, c^k(i, \mu) = 0$

and $R^k(i,\mu) = 0$ for all μ. Thus, the problem can be described by the CTMDP model in a semi-Markov environment given in Eq. (6.1) with

$$\begin{aligned}
&S = \{0,1,2,\ldots\}, \;\; A(i) = A, \;\; i \in S,\\
&q_{ij}^k(\mu) = \begin{cases} \mu & j = i-1, \;\; i \geq 1\\ \lambda_k & j = i+1, \;\; i \geq 0, \end{cases}\\
&q_{ii}^k(\mu) = -\sum_{j \neq i} q_{ij}^k(\mu),\\
&r^k(i,\mu) = -c^k(i,\mu), \;\; p_{ij}^k = \delta_{ij}.
\end{aligned}$$

For this problem, the optimality equation is

$$\begin{aligned}
-\frac{d}{dt}U_k(t,i) &= \sup_{\mu \in A}\{-c^k(i,\mu)\bar{G}_k(t) - R^k(i,\mu)g_k(t)\\
&\quad + \mu U_k(t,i-1) - (\lambda_k + \mu + \alpha)U_k(t,i)\}\\
&\quad + g_k(t)\sum_{k'} \psi_{kk'}U_{k'}(0,i) + \lambda_k U_k(t,i+1), \;\; i > 0,\\
-\frac{d}{dt}U_k(t,0) &= -c^k(0)\bar{G}_k(t) - R^k(0)g_k(t) + g_k(t)\sum_{k'} \psi_{kk'}U_{k'}(0,0)\\
&\quad + \lambda_k U_k(t,1) - (\lambda_k + \alpha)U_k(t,0), \;\; i = 0.
\end{aligned}$$

From the above formulae we may obtain some special policies that are optimal or ε-optimal using the results in this section.

The underlying model, M/M/1 in a semi-Markov environment, was discussed in Neuts [98], where he discussed other queueing systems in Markov or semi-Markov environments.

2. Semi-Markov Decision Processes in Semi-Markov Environments

In this section, we study a model of semi-Markov decision processes in a semi-Markov environment (SMDPs-SE for short). Similarly to the previous section, we present the model, prove the optimality equation, and discuss the case when the environment is Markov.

2.1 Model

The model considered here is:

$$\{(K,G),(SMDP^k, p^k, R^k, k \in K), V\}, \tag{6.38}$$

where

(a) The system's environment $\{(J_n, L_n), n \geq 0\}$ is the same as that given in Eq. (6.1) in the previous section.

(b) When the environment's state is k, the system can be described by

$$SMDP^k := \{S, A(i), q^k, T^k, r^k\}. \tag{6.39}$$

Here, the state space S and the action set $A(i)$ are countable. The state transition probability $q_{ij}^k(a)$, the probability distributions $T^k(\cdot \mid i, a, j)$ of the holding time, and the reward $r^k(i, a, j, t)$ are the same as $p_{ij}(a)$, $T(\cdot \mid i, a, j)$ and $r(i, a, j, t)$, respectively, in the SMDP model given by Eq. (5.1). Also, we call the states in S the *inner states* of the system. The SMDP model describing the system in the time interval $[L_n, L_{n+1})$ is denoted by SMDP(n), which is exactly SMDP$_k$ when $J_n = k$.

(c) For $k \in K$ and $(i, a) \in \Gamma$, let $p_{ij}^k(a) := P\{$the inner state at L_{n+1} is $j \mid$ the inner state and the action taken at $L_{n+1} - 0$ are i and a, respectively, $J_n = k\}$. Then, $p_{ij}^k(a)$ is the instantaneous inner state transition probability caused by the change of the environment state. We assume that if L_{n+1} is also the state transition epoch of SMDP(n) then only the state transition caused by the environment is considered; that is, the latter transition is preemptive. $1 - \sum_j p_{ij}^k(a)$ may be greater than zero, and can be interpreted as the probability that the system is terminated by the environment state transition at L_{n+1}. $R^k(i, a)$ is the reward received of the system at $L_{n+1} - 0$ with the inner state being i and action being a when the environment state changes.

(d) $\alpha > 0$ is the discount rate, and V is the discounted criterion defined later.

For $k \in K, t \in E := [0, \infty), i \in S$, let $x = (k, t, i)$ denote that for some $n \geq 0$ the inner state is i at $L_n + t$ and $J_n = k, t < \Delta L_n$. For convenience, x is also called a state, the set of which is denoted by X. It is apparent that one can consider only those $x = (k, t, i)$ with $G_k(t) < 1$. A history is $h_m = (x_0, a_0, s_0, x_1, a_1, s_1, \ldots, x_m)$, where $x_n = (k_n, t_n, i_n)$ is the state of the system after its nth state transition, and a_n and s_n denote, respectively, the action taken and the holding time at state x_n. As a history, h_m should satisfy the following rule.

Rule. *For $i = 0, 1, 2, \ldots, m-1, a_n \in A(i_n)$, and (1) $t_{n+1} = 0$, or (2) $k_{n+1} = k_n$ and $t_{n+1} = t_n + s_n$.*

For $m \geq 0$, the set of all histories h_m is denoted by H_m. A policy π is a sequence $(\pi_0, \pi_1, \ldots)$ such that for each $m \geq 0$ and $h_m = (x_0, a_0, s_0, x_1, a_1, s_1, \ldots, x_m)$, $\pi_m(\cdot|h_m)$ is a probability in $A(i_m)$. We denote by Π the set of all policies. For $\pi \in \Pi$, if $\pi_m(\cdot \mid h_m) = \pi_m(\cdot \mid x_m)$, π is called a stochastic Markov policy. Moreover, if $\pi_m(\cdot|x) = \pi_0(\cdot|x)$, we then say that π is a stochastic stationary policy. We denote by Π_m (or Π_s) the set of all stochastic Markov (or stationary) policies.

Define the decision function set $F = \{f$: for each $k \in K$ and $i \in S$, $f(k,t,i)$ belongs to the action set $A(i)$ and is Lebesgue measurable in $t\}$. As in Chapter 2, we simply denote the stationary policy $(f, f, \cdots)$ by f.

For each $\pi \in \Pi$, the probability space $(\Omega, \mathcal{F}, P_\pi)$ under policy π can be constructed in an obvious way. In fact, the system under any policy $\pi \in \Pi_m$ is a piecewise semi-Markov reward process; that is, in each $[L_n, L_{n+1})$, the system is a semi-Markov process with reward.

For $x = (k,t,i) \in X$, suppose that the system is initially in x and $J_0 = k, L_1 > t$. Then, we denote by S_n^m, Δ_n^m, t_n^m the state, the action chosen, and the holding time, respectively, after the nth state transition in $[L_m, +\infty)$ for SMDP(m) ($n, m \geq 0$). Let

$$T_0^m = t\delta_{m0}, \;\; T_n^m = T_{n-1}^m + t_{n-1}^m, \;\; n > 0, m \geq 0.$$

T_n^m is the epoch of the nth state transition in $[L_m, +\infty)$ for SMDP(m). We also denote by X_m the mth state of the type (k,t,i) in $[t, +\infty)$. By the stationariness of the semi-Markov environment, $X_0 = x = (k,t,i)$ can be interpreted as that the system is in state i at $L_0 + t$ provided that $J_0 = k$ and $L_1 > t$. Denote by N_m the number of inner state transitions in $[L_m + T_0^m, L_{m+1})$ (not including the state transition at L_{m+1} caused by the environment). It should be noted that the event $\{N_m \geq n\}$ is exactly the event $\{T_n^m < L_{m+1}\}$. For $m, n \geq 0$, we let

$$\begin{aligned} r(m,n) &= r^{J_m}(S_n^m, \Delta_n^m, S_{n+1}^m, t_n^m), \\ r(m) &= r^{J_m}(S_{N_m}^m, \Delta_{N_m}^m, S_{N_m+1}^m, \Delta_m - T_{N_m}^m), \\ R(m) &= R^{J_m}(S_{N_m}^m, \Delta_{N_m}^m) \end{aligned}$$

be rewards received, respectively, in the nth state transition, in the last state transition, and at $L_{m+1} - 0$ in the interval $[L_m, L_{m+1})$. Then, we define

$$V_m = \sum_{n=0}^{N_m - 1} e^{-\alpha T_n^m} r(m,n) + e^{-\alpha T_{N_m}^m} r(m) + e^{-\alpha \Delta L_m} R(m)$$

to be the total reward discounted to t in the interval $[t, L_1)$ if $m = 0$ or discounted to L_m in the interval $[L_m, L_{m+1})$ if $m > 0$. With this, we now define the discounted criterion by

$$V(\pi, x) = \bar{G}_k(t) E_{\pi,x} \sum_{m=0}^{\infty} e^{-\alpha(L_m - t)} V_m, \;\; x = (k,t,i) \in X, \pi \in \Pi. \quad (6.40)$$

$V(\pi, x)$ is the expected discounted total reward in $[t, \infty)$ under policy π when the state at t is $x = (k,t,i)$. Just as in the previous section, the term $\bar{G}_k(t)$ in the above definition is only for notational simplicity.

We need some conditions to ensure the existence of $V(\pi, x)$. The first one is the regular condition on both the environment process and on the SMDP$_k$.

Condition 6.4 (Regular Condition): *There exist constants $\theta \in (0,1)$ and $\delta > 0$ such that*

$$G_k(\delta) \le 1-\theta, \quad k \in K,$$
$$\sum_j q_{ij}^k(a) T^k(\delta \mid i,a,j) \le 1-\theta, \quad (i,a) \in \Gamma, \quad k \in K.$$

From Ross [114], we can prove similarly that under each policy, the process is regular. That is, in every finite time interval there happen only a finite number of state transitions. Moreover, for given $\alpha > 0$ and all $(i,a) \in \Gamma, k \in K$, we have similarly to Lemma 5.1 that

$$\sum_j q_{ij}^k(a) \int_0^\infty e^{-\alpha t} T^k(dt \mid i,a,j) \le 1-\theta[1-e^{-\alpha\delta}] := \beta,$$
$$\int_0^\infty e^{-\alpha t} G_k(dt) \le \beta.$$

The second condition requires the uniformly bounded rewards.

Condition 6.5: *There exists a nonnegative constant M such that $|r^k(t,i,a)| \le M$ and $|R^k(i,a)| \le M$ for all k,t,i,a.*

With Condition 6.4 and Condition 6.5, it is apparent that the $V(\pi,(k,t,i))$ is well defined, bounded uniformly in π,k,t,i, and measurable in t. In the following, when we say that $V(x)$ is measurable if $V(x)$ is measurable in t.

We define the optimal value by

$$V^*(x) = \sup\{V(\pi,x) : \pi \in \Pi\}, \quad x \in X.$$

It is sure that $V^*(x)$ is measurable. ε-optimal policies and optimal policies are defined as usual.

At the end of this subsection, it should be noted that SMDP-SE generalizes SMDP. In fact, SMDP-SE are also the usual SMDPs with state $x = (k,t,i)$. But the Regular Condition for these SMDPs may not hold. We consider the following example,

$$G_k(s) = \min(s/2,1), \quad T^k(s|i,a,j) = \chi(s \ge 1), \quad s \ge 0, k,i,a,j.$$

Obviously, the variable t in (k,t,i) can be limited in $[0,2)$ and for such t, it can be proved that the d.f. of the holding time in (k,t,i) is

$$F(k,t,i)(s) = \begin{cases} \frac{s}{2-t} & 0 \le s \le \min(1,2-t) \\ 1 & s > \min(1,2-t). \end{cases}$$

It is apparent that there exist no positive constants θ and δ such that $F(k,t,i)(\delta) \leq 1-\theta$ for all k,t,i. The expected discount factor of one state transition is

$$\begin{aligned}\beta_\alpha(k,t,i) &= \int_0^\infty e^{-\alpha s} d_s F(k,t,i)(s) \\ &= \frac{1-e^{-\alpha \min(1,2-t)}}{\alpha(2-t)} \to 1, \quad t \to 2^-.\end{aligned}$$

So, for SMDPs-SE, the Regular Condition may not hold and the expected discount factor of one state transition may not be less uniformly than one. But in the following, we still prove the standard results for the SMDP-SE model.

2.2 Optimality Equation

First, we should simplify the form of the reward functions.

For $x=(k,t,i) \in X, a \in A(i)$, and $k' \in K$, we define

$$\begin{aligned} r(x,a) = \sum_j q_{ij}^k(a) \int_0^\infty T^k(ds|i,a,j)\{\bar{G}_k(t+s) r^k(i,a,j,s) \\ + \int_{t^+}^{t+s} dG_k(u)[r^k(i,a,j,u-t) + e^{-\alpha(u-t)} R^k(i,a)]\}, \end{aligned} \tag{6.41}$$

$$\beta(x,a,k') = \sum_j q_{ij}^k(a) \int_0^\infty T^k(ds|i,a,j) \int_{t^+}^{t+s} e^{-\alpha(u-t)} dG_{kk'}(u). \tag{6.42}$$

Surely, $r(x,a)$ is the expected discounted total reward in one decision period when the system enters the state x and action a is chosen, and $\beta(x,a,k')$ is the one-step expected discount factor with state x and action a and the next environment state k'.

The following lemma shows that the two reward functions $r^k(i,a,j,u)$ and $R^k(i,a)$ contribute to the criterion $V(\pi,x)$ only through $r(x,a)$. Hence, the same $r(x,a)$ will result in the same $V(\pi,x)$.

Lemma 6.6: *For any policy $\pi \in \Pi$ and $x=(k,t,i) \in X$,*

$$\begin{aligned} V(\pi,x) = \sum_{a \in A(i)} \pi_0(da \mid x)\{r(x,a) \\ + \sum_{i_1} q_{ii_1}^k(a) \int_0^\infty T^k(ds \mid i,a,i_1) \cdot [e^{-\alpha s} V(\pi^{x,a,s}, (k,t+s,i_1)) \\ + \int_{t+}^{t+s} \sum_{k'} dG_{kk'}(u) e^{-\alpha(u-t)} \sum_j p_{ij}^k(a) V(\pi^{x,a,u-t}, (k',0,j))]\}, \end{aligned} \tag{6.43}$$

where the policy $\pi^{x,a,s} = (\pi_0', \pi_1', \ldots) \in \Pi$ is defined by $\pi_n'(\cdot \mid h_n) = \pi_{n+1}(\cdot \mid x,a,s,h_n)$ for $n \geq 0$.

Proof: For convenience, we let an event be $EV = \{X_0 = x, \Delta_0^0 = a, S_1^0 = i_1, t_0^0 = s, L_1 = u, J_1 = k'\}$. Then, for each state $x = (k, t, i) \in X$, we have that

$$
\begin{aligned}
V(\pi, x) &= \bar{G}_k(t) E_{\pi,x}\{e^{\alpha t} V_0 + e^{-\alpha(L_1 - t)} \sum_{m=1}^{\infty} e^{-\alpha(L_m - L_1)} V_m\} \\
&= \sum_{k'} \int_{t+}^{\infty} dG_{kk'}(u) \sum_{a \in A(i)} \pi_0(da \mid x) \sum_{i_1} q_{ii_1}^k(a) \int_0^{\infty} T^k(ds \mid i, a, i_1) \\
&\quad \cdot E_{\pi,x}\{e^{\alpha t} V_0 + e^{-\alpha(u-t)} \sum_{m=1}^{\infty} e^{-\alpha(L_m - L_1)} V_m | EV\} \\
&= \sum_{a \in A(i)} \pi_0(da \mid x) \sum_{i_1} q_{ii_1}^k(a) \int_0^{\infty} T^k(ds \mid i, a, i_1) \\
&\quad \cdot \Bigg\{ \sum_{k'} \int_{t+}^{t+s} dG_{kk'}(u) [r^k(i, a, i_1, u - t) + e^{-\alpha(u-t)} R^k(i, a) \\
&\quad + e^{-\alpha(u-t)} \sum_j p_{ij}^k(a) E_{\pi}(\sum_{m=1}^{\infty} e^{-\alpha(L_m - L_1)} V_m | EV, S_0^1 = j)] \\
&\quad + \sum_{k'} \int_{(t+s)+}^{\infty} dG_{kk'}(u) [r^k(i, a, i_1, s) + E_{\pi}(\sum_{n=1}^{N_0 - 1} e^{-\alpha(T_n^0 - t)} r(0, m) \\
&\quad + e^{-\alpha(T_{N_0}^0 - t)} r(0) + e^{-\alpha(u-t)} (R(0) + \sum_{m=1}^{\infty} e^{-\alpha(L_m - L_1)} V_m) | EV)] \Bigg\} \\
&= \sum_{a \in A(i)} \pi_0(da \mid x) \sum_{i_1} q_{ii_1}^k(a) \int_0^{\infty} T^k(ds \mid i, a, i_1) \\
&\quad \cdot \{ \sum_{k'} \int_{t+}^{t+s} dG_{kk'}(u) [r^k(i, a, i_1, u - t) + e^{-\alpha(u-t)} R^k(i, a) \\
&\quad + e^{-\alpha(u-t)} \sum_j p_{ij}^k(a) V(\pi^{x,a,u-t}, (k', 0, j))] \\
&\quad + \sum_{k'} \int_{(t+s)+}^{\infty} dG_{kk'}(u) [r^k(i, a, i_1, s) \\
&\quad + e^{-\alpha s} E_{\pi^{x,a,s}, (k, t+s, i_1)} \sum_{m=0}^{\infty} e^{-\alpha(L_m - t - s)} V_m] \}.
\end{aligned}
$$

This together with the definition of $r(x, a)$ implies the lemma. □

For a stochastic stationary policy $\pi \in \Pi_s$, we have a simpler expression than Eq. (6.43). This is shown in the following theorem.

Theorem 6.6: *For a stochastic stationary $\pi \in \Pi_s, V(\pi, x)$ is the unique bounded measurable solution of the following equation,*

$$V(x) = \sum_{a \in A(i)} \pi_0(da|x)\{r(x,a) + \sum_{k'} \beta(x,a,k') \sum_j p_{ij}^k(a) V(k',0,j) + \sum_j q_{ij}^k(a) \int_0^\infty e^{-\alpha s} T^k(ds|i,a,j) V(k,t+s,j)\}, \ x \in X. \quad (6.44)$$

Proof: It follows Lemma 6.6 that $V(\pi, x)$ satisfies Eq. (6.44). Hence, $V(\pi, x)$ is a bounded and measurable solution of Eq. (6.44). For the uniqueness, suppose that V is another bounded measurable solution of Eq. (6.44). Let for convenience

$$DV(x,a) = \sum_{k'} \beta(x,a,k') \sum_j p_{ij}^k(a) V(k',0,j).$$

Then, from Eq. (6.44), we can rewrite $V(x)$ as follows:

$$\begin{aligned} V(x) &= \bar{G}_k(t) E_{\pi,x}\{r(X_0, \Delta_0) + DV(X_0, \Delta_0)\} \\ &\quad + E_{\pi,x}\{\chi(N_0 \geq 1) e^{-\alpha(T_1^0 - t)} V(X_1)\}. \end{aligned}$$

From the equation above, we can show by the induction method that

$$\begin{aligned} V(x) &= \bar{G}_k(t) \sum_{n=0}^{N} E_{\pi,x}\{\chi(N_0 \geq n) e^{-\alpha(T_n^0 - t)} \\ &\quad \cdot [r(X_n, \Delta_n) + DV(X_n, \Delta_n)]\} \\ &\quad + E_{\pi,x}\{\chi(N_0 \geq N+1) e^{-\alpha(T_{N+1}^0 - t)} V(X_{N+1})\}, \ \ N \geq 0. \end{aligned}$$

By letting $N \to \infty$ in the above formulae, it follows the definition of $r(x,a)$ and $\beta(x,a,k')$ and we have that

$$\begin{aligned} V(x) &= \bar{G}_k(t) \sum_{n=0}^{\infty} E_{\pi,x}\{\chi(N_0 \geq n) e^{-\alpha(T_n^0 - t)} \\ &\quad \cdot [r(X_n, \Delta_n) + DV(X_n, \Delta_n)]\} \\ &= \bar{G}_k(t) \sum_{n=0}^{\infty} E_{\pi,x}\{\chi(N_0 \geq n) e^{-\alpha(T_n^0 - t)} r(X_n, \Delta_n)\} \\ &\quad + \bar{G}_k(t) \sum_{n=0}^{\infty} E_{\pi,x}\{\chi(N_0 = n) e^{-\alpha(T_n^0 - t)} e^{-\alpha(L_1 - T_n^0)} V(J_1, 0, S_0^1)\} \\ &= \bar{G}_k(t) \sum_{n=0}^{\infty} E_{\pi,x}\{\chi(N_0 \geq n) e^{-\alpha(T_n^0 - t)} r(X_n, \Delta_n)\} \\ &\quad + \bar{G}_k(t) E_{\pi,x}\{e^{-\alpha(L_1 - t)} V(J_1, 0, S_0^1)\}. \end{aligned} \quad (6.45)$$

We let $Q(x) := V(\pi, x) - V(x)$ for $x \in X$. By noting that $V(\pi, x)$ also satisfies Eq. (6.45), we have that

$$Q(x) = \bar{G}_k(t) E_{\pi,x}\{e^{-\alpha(L_1 - t)} Q(J_1, 0, S_0^1)\}. \tag{6.46}$$

From the above equation, Condition 6.4, and the boundedness of $Q(x)$, we can get that

$$|Q(x)| \leq \int_t^\infty e^{-\alpha(u-t)} dG_k(u) \cdot \sup_{k,i} |Q(k, 0, i)|, \quad x = (k, t, i) \in X.$$

Then, by taking $x = (k, 0, i)$ above we get

$$\sup_{k,i} |Q(k, 0, i)| \leq \beta \sup_{k,i} |Q(k, 0, i)|.$$

Due to $\beta < 1$, we know that $\sup_{k,i} |Q(k, 0, i)| = 0$; that is, $Q(k, 0, i) = 0$ for all k and i. Hence, $Q(x) = 0$ for all $x \in X$; that is, $V(x) = V(\pi, x)$ for all $x \in X$. This shows the uniqueness of solutions. □

With Eq. (6.45), the following corollary can be proved by the induction method.

Corollary 6.1: *We have the following expressions for each policy* $\pi \in \Pi$ *and state* $x \in X$*:*

$$V(\pi, x) = \sum_{m=0}^{\infty} V_m(\pi, x),$$

$$V_m(\pi, x) = \bar{G}_k(t) \sum_{n=0}^{\infty} E_{\pi,x}\{\chi(N_m \geq n) e^{-\alpha(L_m + T_n^m - t)} r(X_n, \Delta_n)\}. \tag{6.47}$$

Now, we introduce another SMDP-SE model as follows.

$$\{(K, G), S, A(i), q^k, T^k, r'^k, p^k, k \in K, V'\}, \tag{6.48}$$

where the elements are the same as those given in Eq. (6.38) except that $R^k(i, a) = 0$ and $r'^k(t, i, a) = r(x, a)/\bar{G}_k(t)$ which is received when the system enters into the state $x = (k, t, i)$ with the action a being chosen. The discounted criterion V' is then defined as follows.

$$\begin{aligned} V'_m &= \sum_{m=0}^{N_m} e^{-\alpha T_n^m} r'^{J_m}(T_n^m, S_n^m, \Delta_n^m), \quad m \geq 0, \\ V'(\pi, x) &= \bar{G}_k(t) E_{\pi,x} \sum_{m=0}^{\infty} e^{-\alpha(L_m - t)} V'_m. \end{aligned}$$

This is similar to that in Eq. (6.40) for the original SMPD-SE model (6.38).

The SMDP-SE model (6.48) differs from the SMDP-SE model (6.38) in two aspects: (1) there is no reward received when the environment changes its state, and (2) the reward received when the system's inner state changes depends on t, the lasting duration of the environment at the same state.

The following theorem shows that these two models are equivalent.

Theorem 6.7: *For each $\pi \in \Pi$ and $x \in X$, $V'(\pi, x) = V(\pi, x)$.*

Proof: It follows from Eq. (6.47) that both $V'(\pi, x)$ and $V(\pi, x)$ depend only on $r(x, a)$. But $r(x, a)$ for the two models are identical. Therefore, the theorem is true. □

From the theorem above, we still write $V'(\pi, x)$ as $V(\pi, x)$.

For a bounded measurable function V on X, state $x = (k, t, i) \in X$, and action $a \in A(i)$, we introduce $T_a V(x)$ as follows.

$$\begin{aligned} T_a V(x) &= r(x,a) + \sum_{k'} \beta(x,a,k') \sum_j p_{ij}^k(a) V(k',0,j) \\ &\quad + \sum_j q_{ij}^k(a) \int_0^\infty e^{-\alpha s} T^k(ds \mid i,a,j) V(k, t+s, j). \end{aligned}$$

Then, we define

$$T_f V(x) = T_{f(x)} V(x), \quad f \in F, \quad TV(x) = \sup_{f \in F} T_f V(x).$$

It is easy to see that $T_f V$ and TV are also bounded and measurable. Using the term $T_f V$, Theorem 6.6 above says that $V(f)$ is the unique bounded and measurable solution of $T_f V = V$. As in the previous chapters, $T_a V, T_f V$, and TV are introduced just for notational simplicity. Based on Theorem 6.6 we know that the optimality equation for SMDP-SE models is as follows.

$$\begin{aligned} V(x) &= \sup_{a \in A(i)} \{ r(x,a) + \sum_{k'} \beta(x,a,k') \sum_j p_{ij}^k(a) V(k',0,j) \\ &\quad + \sum_j q_{ij}^k(a) \int_0^\infty e^{-\alpha s} T^k(ds|i,a,j) V(k, t+s, j) \}, \\ &\qquad\qquad x \in X. \end{aligned} \tag{6.49}$$

By using the notation T, the optimality equation above can be rewritten as a simple form: $V = TV$.

Theorem 6.8: *V^* is the unique bounded and measurable solution of the optimality equation (6.49). Moreover, for any $\varepsilon > 0$, if $TV^* \leq T_f V^* + \varepsilon$ then f is $(2-\beta)(1-\beta)^{-2}\varepsilon$-optimal.*

Proof: By Lemma 6.6 it is easy to show that $V^* \leq TV^*$. So, it follows from Lemma 4.11 that for any given $\varepsilon > 0$, there is f such that:

$$V^*(x) \leq TV^*(x) \leq T_f V^*(x) + \varepsilon, \;\; x \in X.$$

Let $Q(x) = V^*(x) - V(f,x)$. Then, it can be proved from the definition of T_f and Eq. (6.44) that for $x = (k,t,i) \in X$,

$$\begin{aligned} Q(x) &\leq T_f V^*(x) - V(f,x) + \varepsilon \\ &= \sum_{k'} \beta(x,f,k') \sum_j p_{ij}^k(f) Q(k',0,j) \\ &\quad + \sum_j q_{ij}^k(f) \int_0^\infty e^{-\alpha s} T^k(ds|i,f,j) Q(k,t+s,j) + \varepsilon. \end{aligned}$$

As in the proof of Theorem 6.6, we introduce a notation

$$DQ(x,a) = \sum_{k'} \beta(x,a,k') \sum_j p_{ij}^k(a) Q(k',0,j).$$

Then, the previous expression for $Q(x)$ can be rewritten as

$$\begin{aligned} Q(x) &\leq \bar{G}_k(t) E_{f,x}\{DQ(X_0,\Delta_0) + \varepsilon\} \\ &\quad + \bar{G}_k(t) E_{f,x}\{\chi(N_0 \geq 1) e^{-\alpha(T_1^0 - t)} Q(X_1)\}. \end{aligned}$$

By the induction method, we can prove from the above formulae that

$$\begin{aligned} Q(x) &\leq \bar{G}_k(t) \sum_{n=0}^\infty E_{f,x}\{\chi(N_0 \geq n) e^{-\alpha(T_n^0 - t)} \varepsilon\} \\ &\quad + \bar{G}_k(t) E_{f,x} e^{-\alpha(L_1 - t)} Q(J_1, 0, S_0^1) \\ &\leq \bar{G}_k(t) E_{f,x}\{e^{-\alpha(L_1 - t)} Q(J_1,0,S_0^1)\} + (1-\beta)^{-1}\varepsilon, \;\; x \in X. \end{aligned}$$

Let $\Delta = \sup_{k,i} Q(k,0,i)$. Then Δ is finite and from the above formula we have that

$$Q(k,0,i) \leq E_{f,x} e^{-\alpha L_1} \Delta + (1-\beta)^{-1}\varepsilon \leq \beta\Delta + (1-\beta)^{-1}\varepsilon.$$

By taking the supremum in the above formula over k and i, we can get that $\Delta \leq \beta\Delta + (1-\beta)^{-1}\varepsilon$. Hence, $\Delta \leq (1-\beta)^{-2}\varepsilon$, and so

$$\begin{aligned} Q(x) &\leq \bar{G}_k(t) E_{f,x}\{e^{-\alpha(L_1 - t)}\Delta\} + (1-\beta)^{-1}\varepsilon, \\ &\leq \Delta + (1-\beta)^{-1}\varepsilon, \\ &\leq (1-\beta)^{-2}\varepsilon + (1-\beta)^{-1}\varepsilon := d\varepsilon, \;\; x \in X, \end{aligned}$$

where $d = (2-\beta)(1-\beta)^{-2}\varepsilon$. This implies that $V^*(x) \le V(f,x) + d\varepsilon$ for all $x \in X$, and so f is a $(2-\beta)(1-\beta)^{-2}\varepsilon$-optimal policy. Moreover,

$$\begin{aligned} V^*(x) &\ge V(f,x) = T_f V(f,x) \ge T_f[V^*(x) - d\varepsilon] \\ &\ge T_f V^*(x) - (1+\beta)d\varepsilon \\ &\ge TV^*(x) - \varepsilon - (1+\beta)d\varepsilon. \end{aligned}$$

Letting $\varepsilon \to 0$, one gets $V^* \ge TV^*$. So, $V^* = TV^*$.

Next, suppose that V is a bounded and measurable solution of the optimality equation. Let $Q^*(x) = V^*(x) - V(x)$ for $x \in X$. Then, we can get from the optimality equation (6.49) that

$$\begin{aligned} |Q^*(x)| \le \sup_{a \in A(i)} \{ &|DQ^*(x,a)| \\ &+ \sum_j q_{ij}^k(a) \int_0^\infty e^{-\alpha s} T^k(ds|i,a,j)|Q^*(k,t+s,j)|\}. \end{aligned}$$

For any $\varepsilon > 0$, there is $f \in F$ attaining the ε-supremum in the above for $Q(x)$. So, the similar procedure as above can result in $Q^*(x) = 0$ for all $x \in X$. That is, $V(x) = V^*(x)$ for all $x \in X$. This shows the uniqueness of solutions of the optimality equation. □

In this subsection, we first get an expression for the criterion $V(\pi, x)$, given in Lemma 6.6, which corresponds to Condition 2.2 for DTMDPs. Then, we simplify the reward functions into $r(x,a)$ in Theorem 6.7. Based on this we show in Theorem 6.8 other standard results, that is, the validity of the optimality equation and the optimality of a policy achieving the supremum of the optimality equation. In the next subsection, we study the SMDP-SE model when the environment is Markov.

2.3 Markov Environment

Suppose that there is a bounded and reserve matrix $T = (T_{kk'})$ such that Eq. (6.24) is true. We define for $k, k' \in K$ and $(i,a) \in \Gamma$ that

$$\begin{aligned} r(k,i,a) = \sum_j q_{ij}^k(a) \int_0^\infty T^k(ds|i,a,j) \Big\{ &e^{T_{kk}s} r^k(i,a,j,s) \\ &- T_{kk} \int_0^s e^{T_{kk}u}[r^k(i,a,j,u) + e^{-\alpha u} R^k(i,a)]du \Big\}, \end{aligned} \tag{6.50}$$

$$\begin{aligned} \beta(k,i,a,k') = (1-\delta_{kk'})T_{kk'} \sum_j q_{ij}^k(a) \\ \cdot \int_0^\infty T^k(ds|i,a,j) \int_0^s e^{(T_{kk}-\alpha)u} du. \end{aligned} \tag{6.51}$$

Then, from the definitions of $r(x,a)$ and $\beta(x,a,k')$ given in Eqs. (6.41) and (6.42), respectively, it is easy to see that

$$r(x,a) = e^{T_{kk}t} r(k,i,a), \quad \beta(x,a,k') = e^{T_{kk}t}\beta(k,i,a,k'). \tag{6.52}$$

Based on these equations, we have similar properties for the discounted criterion $V(f,(k,t,i))$ and the optimal value $V^*(k,t,i)$, as shown in the following theorem, where $\alpha(k,i,a,j) = \int_0^\infty e^{(T_{kk}-\alpha)s} T^k(ds|i,a,j)$.

Theorem 6.9: *We have for all f,k,t,i,*

$$V(f,(k,t,i)) = e^{T_{kk}t} V(f,(k,0,i)), \quad V^*(k,t,i) = e^{T_{kk}t} V^*(k,0,i).$$

Moreover, the optimality equation becomes

$$\begin{aligned} V(k,i) = \sup_{a\in A(i)} \{ & r(k,i,a) + \sum_j q_{ij}^k(a)\alpha(k,i,a,j)V(k,j) \\ & + \sum_{k'} \beta(k,i,a,k') \sum_j p_{ij}^k(a) V(k',j)\}, \quad k\in K, \quad i\in S, \end{aligned} \tag{6.53}$$

which has the unique bounded solution $V^*(k,0,i)$.

Proof: Substituting Eq. (6.52) into the optimality equation (6.49), we can get that for $x=(k,t,i)\in X$,

$$\begin{aligned} V(k,t,i) = \sup_{a\in A(i)} & e^{T_{kk}t}\{ r(k,i,a) + \sum_{k'} \beta(k,i,a,k') \sum_j p_{ij}^k(a) V(k',0,j) \\ & + \sum_j q_{ij}^k(a) \int_0^\infty e^{-\alpha s} T^k(ds|i,a,j) e^{-T_{kk}t} V(k,t+s,j)\}. \end{aligned}$$

We denote $\hat{V}^*(k,t,i) = e^{T_{kk}t} V^*(k,t,i)$. Then, by multiplying both sides of the above equation by $e^{-T_{kk}t}$, we can know that $\hat{V}^*$ is the unique bounded and measurable solution of the following equation, for $x=(k,t,i)\in X$,

$$\begin{aligned} \hat{V}(k,t,i) = \sup_{a\in A(i)} & \{ r(k,i,a) + \sum_{k'} \beta(k,i,a,k') \sum_j p_{ij}^k(a) \hat{V}(k',0,j) \\ & + \sum_j q_{ij}^k(a) \int_0^\infty e^{(T_{kk}-\alpha)s} T^k(ds|i,a,j) \hat{V}(k,t+s,j)\}. \end{aligned}$$

Taking any $t_0 \geq 0$, it is easy to see that $\{\hat{V}^*(k,t+t_0,i) : k,t,i\}$ is also a bounded and measurable solution of the above equation. Due to the uniqueness of the solution of the above equation, we know that $\hat{V}^*(k,t,i)$ is irrespective of t. That is,

$$e^{-T_{kk}t} V^*(k,t,i) = V^*(k,0,i), \quad \forall k,t,i.$$

Because $\hat{V}^*(k,0,i) = V^*(k,0,i)$, we know that $V^*(k,0,i)$ is the unique bounded solution of Eq. (6.53).

For a stationary policy f, it can be proved similarly that $V(f,(k,t,i)) = e^{T_{kk}t}V(f,(k,0,i))$. □

The phase-type environment and the weak approximation problem can be studied similarly to those in the previous section. We do not discuss them here.

3. Mixed Markov Decision Processes in Semi-Markov Environments

A mixed MDP in a semi-Markov environment (MMDP-SE for short) can describe such systems as they are influenced by their environments. The environment can be modeled by a stationary semi-Markov process, whereas in a given environment state, the system itself can be modeled by a CTMDP or by a SMDP model according to which environment state it is. This model differs from those discussed in the previous two sections where the MDP model describing the system may change at the environment state transition. We first formulate the model precisely and present several conditions. Then, we combine the ideas and methods in the previous two sections to show the validity of the optimality equation and the existence of ε-optimal policies. Furthermore, we discuss the Markov environment.

3.1 Model

The model of the mixed Markov decision process studied here is:

$$\{(K,G),(MDP_k,k\in K),(p^k,k\in K),V\}. \tag{6.54}$$

The elements above are as follows.

(a) The system's environment$\{(J_n,L_n),n\geq 0\}$ is the same as those in the previous sections. Moreover, it is assumed that K is divided into two disjoint subsets K_1 and K_2; that is, $K = K_1 \bigcup K_2$ with $K_1 \bigcap K_2 = \emptyset$, and for $k \in K_2$, $G_{kk'}(t)$ is absolutely continuous with p.d.f. $g_{kk'}(t)$ and $g_k(t) = \sum_{k'} g_{kk'}(t)$.

(b) In the time interval $[L_n, L_{n+1})$ with the environment state $J_n = k$, if $k \in K_1$ then the system can be described by a SMDP model:

$$\text{SMDP}^k = \{S, A(i), q_{ij}^k(a), T^k(\cdot|i,a,j), r^k(t,i,a)\}. \tag{6.55}$$

Here, each element is the same as in the SMDP model (5.1) except that $A(i)$ is countable and $r^k(t,i,a)$ is the reward received by the system at time $L_n + t$ with state i being just reached and action a being taken while $J_n = k$. Hence, the SMDP^k here is nonstationary.

If the environment state $J_n = k \in K_2$, then the system in $[L_n, L_{n+1})$ can be described by a CTMDP model:

$$\text{CTMDP}^k = \{S, (A(i), i \in S), q^k, r^k\}. \tag{6.56}$$

Here, each element is the same as in the CTMDP model (4.1) except that $\lambda^k(i) := \sup\{-q^k_{ii}(a) | a \in A(i)\} < \infty$ for all $i \in S$ and $r^k(t, i, a)$ is the reward rate received by the system at time $L_n + t$ when the system state is i and action a is chosen while $J_n = k$. Here, the state transition rate $q^k_{ij}(a)$ is stationary and the reward rate $r^k(t, i, a)$ is nonstationary.

We call the elements in S the inner states. The MDP model describing the system in the time interval $[L_n, L_{n+1})$ is also denoted by MDP(n). Then, the type of MDP(n) depends on $J_n = k$. It is SMDPs when $k \in K_1$ and CTMDP when $k \in K_2$.

(c) $p^k_{ij}(a) := P\{$the inner state at L_{n+1} is $j \mid$ the inner state and the action taken at $L_{n+1} - 0$ are i and a, respectively, $J_n = k\}$, which is the same as that in SMDP-SE model discussed in the last section.

(d) V is the discounted criterion with the discount rate $\alpha > 0$. It is defined later.

Remark 6.3: *Our model (6.54) can be generalized as the following aspects.*

1. *The reward function can be more complex as discussed in the previous sections. But as shown there, it suffices to consider the form of $r^k(t, i, a)$ given in Eq. (6.55) and Eq. (6.56).*
2. *A more complex case about the kernel $G_{kk'}(t)$ for $k \in K_2$ is that discussed in Section 1. This complex case can also be considered here. But certainly, this only makes the formulae below more complex.*
3. *Because DTMDP is a special case of SMDP, we have not considered DTMDP in the above mixed MDP model.*

For $k \in K, t \in E = [0, \infty), i \in S$, let $x = (k, t, i)$ have the same meaning as that in Section 2.

A history for SMDP$_k$ with $k \in K_1$ is $h_m = ((k, t_0, i_0), a_0, s_0, (k, t_1, i_1), a_1, s_1, \ldots, (k, t_m, i_m))$ with the same environment state k. It has the same meaning as that in the SMDP-SE discussed in the last section. Let H^k_m be the set of such histories h_m. For SMDP$_k$ with $k \in K_1$, a policy is a sequence $\pi^k = (\pi^k_0, \pi^k_1, \ldots)$, the same as that in Chapter 5 for SMDP, the set of which is denoted by Π^k.

For CTMDP$_k$ with $k \in K_2$, a policy is $\pi^k = (\pi^k_t, t \geq 0) \in \Pi^k$, the same as that in Chapter 4 for CTMDPs.

We define the decision function set for MDP$_k$ by $F^k = \{f | f$ is a function from $[0,\infty) \times S$ to $\bigcup_i A(i)$ such that $f(t,i) \in A(i)$ and is measurable in t for each $i\}$.

Now, a policy for the mixed MDP is $\pi = (\pi^k, k \in K)$ with $\pi^k \in \Pi^k$ for $k \in K$. The policy set is denoted by Π. We define the decision function set by $F = \{f = (f^k, k \in K) | f^k \in F^k$ for $k \in K\}$. An element of F is also called a stationary policy. Let Π_s be the set of all stochastic stationary policies, in which π^k is stochastic stationary for each $k \in K$. From the discussion in the previous sections, we know that the (ε-)optimal policies can be restricted in the set of stochastic stationary policies. Hence, we restrict the policies to Π_s. Relaxing this restriction will only make the notations more complex.

For $k \in K_2$, MDP$_k$ is a CTMDP. For each $\pi^k = (\pi^k_t) \in \Pi^k$, we define a matrix $Q^k(\pi,t) = (q^k_{ij}(\pi,t))$ and a column vector $r^k(\pi,t) = (r^k_i(\pi,t))$ the same as those defined in Chapter 4.

To ensure that the model is regular, we summarize the conditions in Section 1 for CTMDP^k with $k \in K_1$, those in Section 2 for SMDP^k with $k \in K_2$, and that for the environment process. This is given as follows.

Condition 6.6:

1. *For each $k \in K_2$, $\pi^k \in \Pi^k$, and $i,j \in S$, $q^k_{ij}(\pi,t)$ is continuous a.e.*

2. *There exist constants $\theta \in (0,1)$ and $\delta > 0$ such that*

$$\begin{aligned} G_k(\delta) &\le 1-\theta, \qquad k \in K, \\ \sum_j q^k_{ij}(a) T^k(\delta | i,a,j) &\le 1-\theta, \qquad (i,a) \in \Gamma, k \in K_1. \end{aligned}$$

From Sections 1, 2, we know that under the above condition the process under each policy is regular. Moreover, for each $(i,a) \in \Gamma$ we have that

$$\begin{aligned} &\sum_j q^k_{ij}(a) \int_0^\infty e^{-\alpha t} T^k(dt | i,a,j) \le \beta, \qquad k \in K_1, \\ &\int_0^\infty e^{-\alpha t} G_k(dt) \le \beta, \qquad k \in K, \end{aligned}$$

where $\beta := 1 - \theta[1 - e^{-\alpha\delta}] < 1$.

For each policy $\pi \in \Pi$, the probability space under policy π can be constructed in an obvious way.

For any given state $x = (k,t,i) \in X$, suppose that the system is initially in state x with $J_0 = k$ and $L_1 > t$. In the following, we suppose that $m \ge 0$ is a given integer and we consider the system in the interval $[L_m, L_{m+1})$.

If $J_m \in K_1$, then we denote by S^m_n, Δ^m_n, t^m_n the inner state, the action chosen, and the holding time in S^m_n, respectively, after the nth state transition in $[L_m, +\infty)$ for MDP(m) ($n \ge 0$). Let $T^m_0 = t\delta_{m0}$ and $T^m_n = T^m_{n-1} + t^m_{n-1}$ for $n > 0$. T^m_n is the epoch of the nth state transition in $[L_m, +\infty)$ for MDP(m).

We also denote by X_m the mth state of the type (k,t,i). Denote by N_m the number of inner state transitions in $[L_m + T_0^m, L_{m+1})$ (not including the state transition caused by the environment). It should be noted that the event $\{N_m \geq n\}$ is equivalent to the event $\{T_n^m < L_{m+1}\}$. For convenience, let $r(m,n) = r^{J_m}(T_n^m, S_n^m, \Delta_n^m)$ be the reward received from the system for its nth state transition in $[L_m, \infty)$. It should be noted that we need less notations here than for the SMDPs-SE because only one reward is considered here.

If $J_m \in K_2$, then we denote by $Y(t)$ and $\Delta(t)$ the state and the action chosen at time t, respectively. Let $r(m,t) = r^{J_m}(t - L_m, Y(t), \Delta(t))$ be the reward rate at time $t \in [L_m, L_{m+1})$.

Having the above preparation, we now define for any integer $m \geq 0$,

$$V_m = \begin{cases} \sum_{n=0}^{N_m} e^{-\alpha T_n^m} r(m,n) & \text{if } J_m \in K_1 \\ \int_{L_m}^{L_{m+1}} e^{-\alpha(s-L_m)} r(m,s) & \text{if } J_m \in K_2, \end{cases}$$

the discounted total reward in $[L_m, L_{m+1})$ discounted to L_m. Moreover, for any given policy $\pi \in \Pi$ and state $x = (k,t,i) \in X$ we define

$$V(\pi, x) = \bar{G}_k(t) E_{\pi,x}\{\sum_{m=0}^{\infty} e^{-\alpha(L_m - t)} V_m\}, \tag{6.57}$$

the expected discounted total reward in $[t, \infty)$ discounted to t under policy π from the initial state x. Because α is fixed, we omit it in the notation of $V(\pi, x)$.

To ensure the existence of $V(\pi, x)$, we give the following condition.

Condition 6.7: *$r^k(t,i,a)$ is uniformly bounded in $k \in K, t \geq 0, i \in S$, and $a \in A(i)$, and Lebesgue measurable in t.*

This condition ensures that $V(\pi, x)$ exists and is uniformly bounded. Let $V^*(x) = \sup\{V(\pi, x) | \pi \in \Pi\}$ be the optimal value. (ε-)optimal policies can be defined as usual.

3.2 Optimality Equation

For $x = (k,t,i) \in X$ and $a \in A(i)$, we define $r(x,a) = \bar{G}_k(t) r^k(t,i,a)$. Moreover, we let for $k \in K_1$,

$$\beta(x,a,k') = \sum_j q_{ij}^k(a) \int_0^{\infty} T^k(ds|i,a,j) \int_{t+}^{t+s} e^{-\alpha(u-t)} dG_{kk'}(u), \tag{6.58}$$

and for $k \in K_2$,

$$r(\pi, x) = \sum_{a \in A(i)} \pi_t^k(a|i) r^k(t,i,a), \quad x = (k,t,i), \pi \in \Pi^k.$$

The following theorem characterizes the discounted criterion $V(\pi, x)$. It combines the similar results given in Theorem 6.6 for SMDPs-SE and Lemma 6.2 for CTMDPs-SE.

Theorem 6.10: *For any stochastic stationary policy $\pi \in \Pi_s$ and state $x = (k, t, i) \in X$,*

$$\begin{aligned} V(\pi, x) &= \sum_{a \in A(i)} \pi_0(a|x)\{r(x, a) \\ &\quad + \sum_{k'} \beta(x, a, k') \sum_j p_{ij}^k(a) V(\pi, (k', 0, j)) \\ &\quad + \sum_j q_{ij}^k(a) \int_0^\infty T^k(ds|i, a, j) \\ &\quad \cdot e^{-\alpha s} V(\pi, (k, t + s, j))\}, \quad k \in K_1, \end{aligned} \tag{6.59}$$

$$\begin{aligned} -\frac{d}{dt} V(\pi, (k, t, i)) &= r(\pi, (k, t, i)) + \sum_j q_{ij}^k(\pi, t) V(\pi, (k, t, j)) \\ &\quad - \alpha V(\pi, (k, t, i)) \\ &\quad + \sum_{k'} g_{kk'}(t) \sum_j p_{ij}^k(\pi^k, t) V(\pi, (k', 0, j)), \\ &\qquad \text{a.e. } t \geq 0, k \in K_2, \end{aligned} \tag{6.60}$$

where $p_{ij}^k(\pi^k, t) = \sum_{a \in A(i)} p_{ij}^k(a) \pi_t^k(a|i)$ for $i, j \in S$ and $k \in K_2$.

Proof: For convenience, we denote $V' = \sum_{m=1}^\infty e^{-\alpha(L_m - L_1)} V_m$. Moreover, for $k \in K_1$ and $x = (k, t, i)$, we denote an event $EV = \{X_0 = x, \Delta_0^0 = a, S_1^0 = i_1, t_0^0 = s, L_1 = u, J_1 = k'\}$. Then, from Eq. (6.57) it can be proved as in Lemma 6.6 that

$$\begin{aligned} &V(\pi, x) \\ &= \sum_{a \in A(i)} \pi_0(a|x) \sum_{i_1} q_{ii_1}^k(a) \int_0^\infty T^k(ds|i, a, i_1) \\ &\quad \cdot \sum_{k'} \int_{t+}^\infty dG_{kk'}(u) E_\pi\{e^{\alpha t} V_0 + e^{-\alpha(u-t)} V' | EV\} \\ &= \sum_{a \in A(i)} \pi_0(a|x) \sum_{i_1} q_{ii_1}^k(a) \int_0^\infty T^k(ds|i, a, i_1)\{r^k(t, i, a) \bar{G}_k(t) \\ &\quad + \sum_{k'} \int_{t+}^{t+s} dG_{kk'}(u) e^{-\alpha(u-t)} \sum_j p_{ij}^k(a) E_\pi(V' | EV, S_0^1 = j) \\ &\quad + \sum_{k'} \int_{(t+s)+}^\infty dG_{kk'}(u) E_\pi[e^{\alpha t} \sum_{n=1}^{N_0} e^{-\alpha T_n^0} r(0, n) + e^{-\alpha(u-t)} V' | EV]\} \end{aligned}$$

$$= \sum_{a\in A(i)} \pi_0(a|x)\{r(x,a) + \sum_{k'} \beta(x,a,k') \sum_j p_{ij}^k(a) V(\pi,(k',0,j))$$
$$+ \sum_j q_{ij}^k(a) \int_0^\infty T^k(ds|i,a,j) e^{-\alpha s} V(\pi,(k,t+s,j))\}.$$

Hence, Eq. (6.59) is true.

For $k \in K_2$, let an event be $EV = \{X_0 = x, L_1 = u, J_1 = k'\}$. We have from Eq. (6.57) that

$$\begin{aligned} V(\pi,k,t,i) &= \int_t^\infty \sum_{k'} dG_{kk'}(u) E_\pi\{\int_t^u e^{-\alpha(s-t)} r(0,s)ds + e^{-\alpha(u-t)} V'|EV\} \\ &= \int_t^\infty \sum_{k'} dG_{kk'}(u)\{\int_t^u e^{-\alpha(s-t)} \sum_{i_1} P_{ii_1}(\pi^k,t,s) r(\pi,k,s,i_1)ds \\ &\quad + \sum_{i_1} P_{ii_1}(\pi^k,t,u) \sum_j p_{i_1 j}^k(\pi^k,u) e^{-\alpha(u-t)} V(\pi,(k',0,j))\}. \end{aligned}$$

We define a vector function $V(\pi,k,t)$ by its ith element being $V(\pi,(k,t,i))$ and $r(\pi,k,s)$ by its ith element being $r(\pi,(k,s,i))$. Then the above equation can be rewritten as the following vector's form,

$$V(\pi,k,t) = \int_t^\infty e^{-\alpha(s-t)} P(\pi^k,t,s) \bar{G}_k(s) r(\pi,k,s)ds$$
$$+ \sum_{k'} \int_t^\infty e^{-\alpha(s-t)} P(\pi^k,t,s) p^k(\pi^k,s) V(\pi,k',0) dG_{kk'}(s). \quad (6.61)$$

Thus, Eq. (6.60) follows by differentiating Eq. (6.61). □

Now, we consider the space of the criterion. For $k \in K_1$, we define Ω^k as the set of bounded and measurable real vector functions $x(t) = (x(t,i), i \in S)$ on $[0,\infty)$. For $k \in K_2$, we define Ω^k as the set of real vector functions $x(t) = (x(t,i), i \in S)$ on $[0,\infty)$ satisfying the following two conditions.

i. $x(t,i)$ is absolutely continuous in t for each $i \in S$.

ii. Both $x(t,i)$ and its derivative $x'(t,i)$ are uniformly bounded in $t \geq 0$ and $i \in S$.

Then, we define $\Omega = \{(x^k(t), k \in K) | x^k(t) \in \Omega^k \text{ for } k \in K\}$. It is easy to see from Theorem 6.10 that $V(\pi) \in \Omega$ for each policy $\pi \in \Pi$. So, Ω is a real criterion space.

As in Section 1, we prove the validity of the optimality equation in two steps. For simplicity, we define operators T_f^* for $f \in F$ and T^* in Ω as follows. For

$V \in \Omega$, $x = (k,t,i) \in X$ and $a \in A(i)$, we define for $k \in K_1$,

$$\begin{aligned} T_a^* V(x) &= r(x,a) + \sum_{k'} \beta(x,a,k') \sum_j p_{ij}^k(a) V^*(k',0,j) \\ &\quad + \sum_j q_{ij}^k(a) \int_0^\infty e^{-\alpha s} T^k(ds|i,a,j) V(k,t+s,j), \end{aligned}$$

and for $k \in K_2$,

$$\begin{aligned} T_a^* V(x) &= r(x,a) + \sum_{k'} g_{kk'}(t) \sum_j p_{ij}^k(a) V^*(k',0,j) \\ &\quad + \sum_j q_{ij}^k(a) V(k,t,j) - \alpha V(k,t,i), \end{aligned}$$

and then

$$T_f^* V(x) = T_{f(x)}^* V(x), \;\; T^* V(x) = \sup_{f \in F} T_f^* V(x), \;\; x \in X.$$

It is easy to see that all $T_f^* V (f \in F)$ and $T^* V$ belong to the space Ω when V belongs to Ω. It should be noted that both right-hand sides in $T_a^* V(x)$ for $k \in K_1$ and $k \in K_2$ include $V^*(k',0,j)$, not $V(k',0,j)$. Hence, $V = T^* V$ is not the optimality equation.

We define sets $X_l = \{x = (k,t,i) \in X \mid k \in K_l\}$ for $l = 1,2$. Then, X_1 and X_2 are disjoint and $X = X_1 \cup X_2$.

Theorem 6.11: *The optimal value V^* is the unique solution of the following equations (i.e., $V = T^* V$) in Ω.*

$$T^* V(k,t,i) = V(k,t,i), \;\; x = (k,t,i) \in X_1,$$
$$T^* V(k,t,i) = -\frac{d}{dt} V(k,t,i), \;\; x = (k,t,i) \in X_2.$$

Proof: First, we show that V^* is a solution of $V = T^* V$ in Ω. For $x \in X_1$, it is easy to see from Theorem 6.10 that $V^*(x) \le T^* V^*(x)$. For $x \in X_2$, one knows from Lemma 4.10 that there is a unique $V_*^k \in \Omega_k$ such that

$$V_*^k(t,i) = T^* V_*^k(t,i), \;\; \text{a.e. } t, \;\; k \in K_2, \;\; i \in S. \tag{6.62}$$

We define $V_*(k,t,i) = V_*^k(t,i)$ for $x = (k,t,i) \in X_2$. Thus, from Lemma 4.11 we know that for any given $\varepsilon > 0$, there is a decision function f attaining the ε-supremum in $T^* V^*(x)$ for $x \in X_1$ or in $T^* V_*(x)$ for $x \in X_2$. That is, $T^* V^*(x) \le T_f^* V^*(x) + \varepsilon$ for $x \in X_1$ and $T^* V_*(x) \le T_f^* V_*(x) + \varepsilon$ for $x \in X_2$. Then,

$$V^*(x) \le T^* V^*(x) \le T_f^* V^*(x) + \varepsilon, \qquad x \in X_1, \tag{6.63}$$

$$-\frac{d}{dt} V_*(x) = T^* V_*(x) \le T_f^* V_*(x) + \varepsilon, \qquad x \in X_2. \tag{6.64}$$

For convenience, let $Q(x) = V^*(x) - V(f,x)$ for $x \in X_1$ and $Q(x) = V_*(x) - V(f,x)$ for $x \in X_2$, and $Q^*(x) = V^*(x) - V(f,x)$ for $x \in X$. Then, for $x \in X_1$, one gets from Theorem 6.10 and Eq. (6.63) that

$$\begin{aligned} Q(x) &\le T_f^* V^*(x) + \varepsilon - V(f,x) \\ &= \sum_{k'} \beta(x, f(x), k') \sum_j p_{ij}^k(f(x)) Q^*(k', 0, j) + \varepsilon \\ &\quad + \sum_j q_{ij}^k(f(x)) \int_0^\infty e^{-\alpha s} T^k(ds|i, f(x), j) Q(k, t+s, j) \\ &= E_{f,x}\{\sum_{k'} \beta(X_0, f(X_0), k') \sum_j p_{S_0^0 j}^k(f(X_0)) Q^*(k', 0, j) + \varepsilon\} \\ &\quad + E_{f,x}\{\chi(N_0 \ge 1) e^{-\alpha(T_1^0 - t)} Q(X_1)\}. \end{aligned}$$

Thus, it can be proved by the induction method that

$$\begin{aligned} Q(x) &\le \sum_{n=0}^N E_{f,x}\{\chi(N_0 \ge n) e^{-\alpha(T_n^0 - t)} [\sum_{k'} \beta(X_n, f(X_n), k') \\ &\quad \cdot \sum_j p_{S_n^0 j}^k(f(X_n)) Q^*(k', 0, j) + \varepsilon]\} \\ &\quad + E_{f,x}\{\chi(N_0 \ge N+1) e^{-\alpha(T_{N+1}^0 - t)} Q(X_{N+1})\}, \quad x \in X_1. \end{aligned}$$

Letting $N \to \infty$, we get from the definition of $\beta(x, a, k')$ that

$$\begin{aligned} Q(x) &\le \sum_{n=0}^\infty E_{f,x}\{\chi(N_0 \ge n) e^{-\alpha(T_n^0 - t)} [\sum_{k'} \beta(X_n, f(X_n), k') \\ &\quad \cdot \sum_j p_{S_n^0 j}^k(f(X_n)) Q^*(k', 0, j) + \varepsilon]\} \\ &= \sum_{n=0}^\infty E_{f,x}\{\chi(N_0 \ge n) e^{-\alpha(T_n^0 - t)} [\sum_{k'} \sum_j q_{S_n^0 j}^k(f(X_n)) \\ &\quad \cdot \int_0^\infty T^k(ds|S_n^0, f(X_n), j) \\ &\quad \cdot \int_{T_n^0 + 0}^{T_n^0 + s} e^{-\alpha(u - T_n^0)} dG_{kk'}(u) Q^*(k', 0, S_0^1) + \varepsilon]\} \\ &\le \sum_{n=0}^\infty E_{f,x}\{\chi(N_0 \ge n) [\int_{T_n^0 + 0}^{T_n^0 + s} e^{-\alpha(u-t)} dG_k(u) D^* + e^{-\alpha(T_n^0 - t)} \varepsilon]\} \\ &\le \int_{t+}^\infty e^{-\alpha(u-t)} dG_k(u) D^* + (1-\beta)^{-1} \varepsilon, \qquad x \in X_1, \end{aligned} \tag{6.65}$$

where $D^* = \sup_{k,i} Q^*(k, 0, i)$.

For $k \in K_2$, it follows from Theorem 6.10 and Eq. (6.64) that

$$\begin{aligned} -\frac{d}{dt}Q(k,t,i) &= -\frac{d}{dt}V_*(k,t,i) + \frac{d}{dt}V(f,(k,t,i)) \\ &\le T_f^* V_*(k,t,i) + \frac{d}{dt}V(f,(k,t,i)) + \varepsilon \\ &= \sum_{k'} g_{kk'}(t) \sum_j p_{ij}^k(f(x)) Q^*(k',0,j) + \varepsilon \\ &\quad + \sum_j q_{ij}^k(f(x)) Q(k,t,j) - \alpha Q(k,t,i), \ \text{ a.e. } t, x \in X_2. \end{aligned}$$

This equation can be rewritten as the following matrix-column form,

$$\begin{aligned} -\frac{d}{dt}Q(k,t) &\le \sum_{k'} g_{kk'}(t) p^k(f,t) Q^*(k',0) + \varepsilon e \\ &\quad + Q^k(f,t) Q(k,t) - \alpha Q(k,t), \quad \text{a.e. } t, \ \ k \in K_2, \end{aligned}$$

where $Q(k,t)$ is a vector with its ith component being $Q(k,t,i)$, $Q^*(k,0)$ similar, and $p^k(f,t) = (p_{ij}^k(f(k,t,i)))$ is a matrix. For $s \le t$, premultiplying the above formula by $e^{-\alpha(t-s)}P(f^k,s,t)$ and rearranging it, we can get for $k \in K_2$,

$$\begin{aligned} & -\frac{d}{dt}\{e^{-\alpha(t-s)} P(f^k,s,t) Q(k,t)\} \\ \le\ & e^{-\alpha(t-s)} \sum_{k'} g_{kk'}(t) P(f^k,s,t) p^k(f,t) Q^*(k',0) + e^{-\alpha(t-s)}\varepsilon, \ \text{ a.e. } t. \end{aligned}$$

Integrating it in $t \in [s,\infty)$, we obtain

$$\begin{aligned} Q(k,s) &\le \int_s^\infty e^{-\alpha(t-s)} \sum_{k'} g_{kk'}(t) P(f^k,s,t) p^k(f,t) Q^*(k',0) dt \\ &\quad + \alpha^{-1}\varepsilon, \ \ s \ge 0, k \in K_2. \end{aligned} \tag{6.66}$$

For $k \in K_2$ and $\pi \in \Pi$, we have from Eq. (6.62) that

$$\begin{aligned} -\frac{d}{dt}V_*(k,t,i) &\ge r_i^k(\pi,t) + \sum_j q_{ij}^k(\pi^k,t) V_*(k,t,j) - \alpha V_*(k,t,i) \\ &\quad + \sum_{k'} g_{kk'}(t) \sum_j p_{ij}^k(\pi^k,t) V_*(k',0,j), \ \text{ a.e. } t \ge 0. \end{aligned}$$

It can be proved similarly to Eq. (6.66) that for each $\pi \in \Pi$,

$$\begin{aligned} V_*(k,t) &\ge \int_s^\infty e^{-\alpha(t-s)} P(\pi^k,s,t) r^k(\pi,t) dt \\ &\quad + \int_s^\infty e^{-\alpha(t-s)} \sum_{k'} g_{kk'}(t) P(\pi^k,s,t) p^k(\pi^k,t) V_*(k',0) dt \\ &\ge V(\pi,(k,t)), \ \ k \in K_2. \end{aligned} \tag{6.67}$$

Then, due to the arbitrariness of π, $V_*(x) \geq V^*(x)$ for $x \in X_2$. Therefore, $Q(x) \geq Q^*(x) \geq 0$ for all x. We define $D = \sup_x Q(x)$. Then from Eqs. (6.65) and (6.66),

$$Q(k,0,i) \leq \beta D^* + \sigma\varepsilon, \qquad k \in K, \;\; i \in S,$$

where $\sigma = \max(\alpha^{-1}, (1-\beta)^{-1})$. So, $D^* \leq \beta D^* + \sigma\varepsilon$ and thus $D^* \leq (1-\beta)^{-1}\sigma\varepsilon$. Again by Eqs. (6.65) and (6.66), we have

$$Q(x) \leq D^* + \sigma\varepsilon \leq (2-\beta)(1-\beta)^{-1}\sigma\varepsilon := \beta(\varepsilon),$$

which implies that $D \leq \beta(\varepsilon)$. So

$$\begin{aligned} V^*(x) &\leq V(f,x) + \beta(\varepsilon), \;\; x \in X_1, \\ V_*(x) &\leq V(f,x) + \beta(\varepsilon), \;\; x \in X_2. \end{aligned} \tag{6.68}$$

This results in $V_*(x) \leq V^*(x)$ for all $x \in X_2$ from Eq. (6.67) and the arbitrariness of ε. So, $V_*(x) = V^*(x)$ for $x \in X_2$. By Eq. (6.62), it is obvious that $V^*(x) = T^*V^*(x)$ for $x \in X_2$. Moreover, we get $V^*(x) = \sup\{V(f,x) | f \in F\}$ for all $x \in X_2$.

On the other hand, for $x \in X_1$, it follows Theorem 6.10 that

$$\begin{aligned} V^*(x) &\geq V(f,x) \\ &= r(x,f) + \sum_{k'} \beta(x, f(x), k') \sum_j p_{ij}^k(f(x)) V(f,(k',0,j)) \\ &\quad + \sum_j q_{ij}^k(f(x)) \int_0^\infty e^{-\alpha s} T^k(ds|i, f(x), j) V(f,(k,t+s,j)) \\ &\geq T_f^* V^*(x) - \sum_{k'} \beta(x, f(x), k')\beta(\varepsilon) \\ &\quad - \sum_j q_{ij}^k(f(x)) \int_0^\infty e^{-\alpha s} T^k(ds|i, f(x), j)\beta(\varepsilon) \\ &\geq T^*V^*(x) - (2+\beta)\beta(\varepsilon), \;\; x \in X_1. \end{aligned}$$

Letting $\varepsilon \to 0^+$, it follows from Eq. (6.63) that $V^*(x) = T^*V^*(x)$ for $x \in X_1$.

Overall, V^* is a solution of the equation $V = T^*V$. It is easy to see that $V^* \in \Omega$.

Next, suppose that V is a solution of $V = T^*V$ in Ω. Then, following the above procedure we get $V = V^*$. □

Now, we can prove the optimality equation.

Theorem 6.12: *V^* is the unique solution of the following optimality equation,*

$$V(x) = \sup_{a \in A(i)} \{ r(x,a) + \sum_{k'} \beta(x,a,k') \sum_j p_{ij}^k(a) V(k',0,j)$$

$$+\sum_j q_{ij}^k(a)\int_0^\infty e^{-\alpha s}T^k(ds|i,a,j)V(k,t+s,j)\},$$
$$x\in X_1, \tag{6.69}$$

$$-\frac{d}{dt}V(k,t,i) = \sup_{a\in A(i)}\{r(x,a)+\sum_{k'}g_{kk'}(t)\sum_j p_{ij}^k(a)V(k',0,j)$$
$$+\sum_j q_{ij}^k(a)V(k,t,j)-\alpha V(k,t,i)\},\ \ x\in X_2. \tag{6.70}$$

Proof: From Theorem 6.11 we know that $V^*(x)$ is a solution of Eqs. (6.69) and (6.70). The uniqueness can be proved as in Theorem 6.11. □

The following corollary follows the proof of Theorem 6.11.

Corollary 6.2: *$V^*(x)=\sup\{V(f,x)|f\in F\}$ for all $x\in X$; that is, the optimality can be achieved in the set of stationary policies. Moreover, for each $\varepsilon\geq 0$, if a decision function f attains the ε-supremum of Eqs. (6.69) and (6.70) (such f must exist whenever ε is positive), then f is $(2-\beta)(1-\beta)^{-1}\sigma\varepsilon$-optimal.*

We have gotten the standard results for the model. In the next subsection, we study the Makrov environment.

3.3 Markov Environment

Suppose there is a transition rate family $T=(T_{kk'})_{k,k'\in K}$, which is conservative and bounded, such that for all $k,k'\in K$,

$$G_{kk'}(t)=\psi_{kk'}G_k(t),\qquad G_k(t)=1-e^{-T_k t},$$
$$\psi_{kk'}=(1-\delta_{kk'})T_{kk'}/T_k,\qquad T_k=-T_{kk}. \tag{6.71}$$

We also assume that both the SMDP^k (6.55) and the CTMDP^k (6.56) are stationary. That is, $r^k(t,i,a)=r^k(i,a)$ is independent of t for all k,i,a. Then, it can be proved easily that for $x=(k,t,i)$ and $a\in A(i)$,

$$r(x,a) = \bar{G}_k(t)r^k(t,i,a)=e^{-T_k t}r(k,i,a), \tag{6.72}$$
$$\beta(x,a,k') = e^{-T_k t}\beta(k,i,a,k'),\qquad k\in K_1, \tag{6.73}$$

where

$$r(k,i,a) = \sum_j q_{ij}^k(a)\int_0^\infty T^k(ds|i,a,j)\{e^{T_{kk}s}r^k(i,a,j,s)$$
$$-T_{kk}\int_0^s e^{T_{kk}u}[r^k(i,a,j,u)+e^{-\alpha u}R^k(i,a)]du\},\ \ k\in K_1,$$

the same as Eq. (6.50), and

$$r(k,i,a)=r(k,i,a)e^{T_{kk}t},\ \ k\in K_2,$$

the same as Eq. (6.25), whereas

$$\begin{aligned}\beta(k,i,a,k') &= (1-\delta_{kk'})T_{kk'}\sum_j q_{ij}^k(a) \\ &\quad \cdot \int_0^\infty T^k(ds|i,a,j)\int_0^s e^{(T_{kk}-\alpha)u}du,\end{aligned}$$

the same as Eq. (6.51). Let $\alpha(k,i,a,j) = \int_0^\infty e^{(T_{kk}-\alpha)s}T^k(ds|i,a,j)$ for $k \in K_1$.

Based on the criterion space Ω, we define another space

$$\Omega' = \{(e^{T_k t}x^k(t,i)) \mid (x^k(t,\cdot)) \in \Omega_k \text{ for each k}\}.$$

The following theorem combines the results in Theorems 6.3 and 6.9.

Theorem 6.13: *When the environment is Markov, $e^{T_k t}V^*(k,t,i) = V^*(k,0,i)$ is independent of t, denoted by $V^*(k,i)$, and $\{V^*(k,i)\}$ is the unique solution in Ω' of the following equations.*

$$\begin{aligned}V(k,i) = \sup_{a\in A(i)}\{&r(k,i,a) + \sum_{k'}\beta(k,i,a,k')\sum_j p_{ij}^k(a)V(k',j) \\ &+ \sum_j q_{ij}^k(a)\alpha(k,i,a,j)V(k,j)\}, \quad k \in K_1, \quad i \in S, \end{aligned} \tag{6.74}$$

$$\begin{aligned}(T_k+\alpha)V(k,i) = \sup_{a\in A(i)}\{&r(k,i,a) + \sum_{k'}T_{kk'}\sum_j p_{ij}^k(a)V(k',j) \\ &+ \sum_j q_{ij}^k(a)V(k,j)\}, \quad k \in K_2, \quad i \in S. \end{aligned} \tag{6.75}$$

Moreover, for $\varepsilon \geq 0$, if $f = (f^k)$ attains the ε-supremum in Eqs. (6.74) and (6.75), then f is $(2-\beta)(1-\beta)^{-1}\sigma\varepsilon$-optimal.

Proof: We prove the theorem similarly to Theorems 6.3 and 6.9. Substituting Eqs. (6.72) and (6.73) into Eqs. (6.69) and (6.70), we can obtain from Theorem 6.12 that $\{e^{T_k t}V^*(k,t,i)\}$ is a solution in Ω' of the following equations.

$$\begin{aligned}V(k,t,i) &= \sup_{a\in A(i)}\{r(k,i,a) + \sum_{k'}\beta(k,i,a,k')\sum_j p_{ij}^k(a)V(k',0,j) \\ &\quad + \sum_j q_{ij}^k(a)\int_0^\infty e^{-(T_k+\alpha)s}T^k(ds|i,a,j)V(k,t+s,j)\}, \quad x \in X_1,\end{aligned}$$

$$\begin{aligned}-\frac{d}{dt}V(k,t,i) &= \sup_{a\in A(i)}\{r(k,i,a) + \sum_{k'}T_{kk'}\sum_j p_{ij}^k(a)V(k',0,j) \\ &\quad + \sum_j q_{ij}^k(a)V(k,t,j) - \alpha V(k,t,i)\}, \quad x \in X_2.\end{aligned}$$

It is easy to see that for each $t_0 \geq 0, \{e^{T_k(t+t_0)}V^*(k,t+t_0,i)\} \in \Omega$ is also a solution of the above equations. By the uniqueness of solutions, we know that $\{e^{T_k t}V^*(k,t,i)\}$ is independent of t. With this, the remainder is apparent. □

Based on the theorem above, similarly to the previous sections, we can transform the mixed MDP in the Markov environment into the following equivalent DTMDP model,

$$\{S', A'(k,i), r'((k,i),a), q'_{(k,i),(k,j)}(a), V_\beta\}. \tag{6.76}$$

For this model, the state space is defined as $S' = \{(k,i) : k \in K, i \in S\}$. The action set available at state (k,i) is $A'(k,i) = A(i)$, which is irrespective of k, the reward function is given by $r'(k,i,a) = r(k,i,a)$ for $k \in K_1$ and $r'(k,i,a) = r(k,i,a)/(T_k + \lambda^k(i) + \alpha)$ for $k \in K_2$, and the state transition probability for $k \in K_1$ is

$$q'_{(k,i)(k',j)}(a) = \begin{cases} \beta^{-1}q^k_{ij}(a)\alpha(k,i,a,j), & k' = k \\ \dfrac{\beta^{-1}\beta(k,i,a,k')p^k_{ij}(a)}{T^k}, & k' \neq k, \end{cases}$$

and for $k \in K_2$ is

$$q'_{(k,i)(k,j)}(a) = \begin{cases} \dfrac{\beta^{-1}[q^k_{ij}(a)+\lambda^k(i)\delta_{ij}]}{T_k+\lambda^k(i)+\alpha}, & k' = k \\ \dfrac{\beta^{-1}T_{kk'}p^k_{ij}(a)}{T_k+\lambda^k(i)+\alpha}, & k' \neq k. \end{cases}$$

V_β is the discounted criterion with the discount factor β.

Theorem 6.14: *The mixed MDP model (6.54) in the Markov environment (6.71) is equivalent to the DTMDP model (6.76) in the following manner: both of their optimality equations are Eqs. (6.74) and (6.75).*

Proof: We know from Theorems 2.2 and 2.4 that the optimal value function $V^*_\beta(k,i) = \sup_\pi V_\beta(\pi,(k,i))$ of the DTMDP model (6.76) is the unique bounded solution of the following equation,

$$V(k,i) = \sup_{a\in A(i)} \{r'(k,i,a) + \beta \sum_{k'} \sum_{j} q'_{(k,i)(k',j)}(a)V(k',j)\}, \quad k \in K, i \in S.$$

By substituting r' and q' into the above equation one can know that it is exactly Eqs. (6.74) and (6.75). □

By Theorem 6.14, we can directly generalize most of the results in DTMDPs into the mixed MDPs in a Markov environment, for example, the varied algorithms to find an ε-optimal stationary policy.

To deal with the general case of Eqs. (6.69) and (6.70), one can approximate $G_{kk'}(t)$ by a phase type distribution function, as discussed in Section 1.

At the end of this section, we consider an example.

Example 6.2: The example is also about the optimal control of a queueing system M/M/1 in a semi-Markov environment, as in Example 6.1. It is described as follows.

1. The environment process is a stationary semi-Markov process, the same as that given in Eq. (6.1).

2. When the environment state is $k \in K$, the customers arrive at the system according to a Poisson process with rate λ_k, and there is only one customer at each arrival.

3. For the server, when the environment state is $k \in K_1$, the server cannot serve the customer (e.g., the server is down). In this time, each arriving customer can be rejected or accepted by the system. When the environment state is $k \in K_2$, the server serves a customer exponentially with parameter μ which can be chosen from a countable set $A \subset [0, \infty)$ at any time.

4. We say that the system is in state i if and only if there are i customers in the system. Suppose that the system is in state i.

5. For $k \in K_1$, there is a holding cost rate $h^k(i)$, and an instantaneous cost d^k occurs when rejecting an arriving customer, and for $k \in K_2$, there is a cost rate $c^k(i, \mu)$ if μ is chosen. We assume that $c^k(0, \mu) = c^k(0)$. In addition, there is an instantaneous cost $R^k(i)$ received before the environment state transition with the inner state i. A special case of the above model is that the server may be down and can be repaired, so $K = \{b, s\}$, where b represents the server is down and s represents the server is in service. Hence, $K_1 = \{b\}$ and $K_2 = \{s\}$.

6. It is assumed that all random variables are mutually independent.

The problem can be modeled by a model of mixed MDPs in a semi-Markov environment. The state space $S = \{0, 1, 2, \ldots\}$. For $k \in K_1$, $A_k(i) = \{c(\text{accepted}), r(\text{rejected})\}$, $q^k_{ij}(c) = \delta_{j,i+1}$, $q^k_{ij}(r) = \delta_{ji}$, $T^k(t|i, a, j) = 1 - e^{-\lambda_k t}$, $p^k_{ij}(a) = \delta_{ij}$ for $a = c, r$, and

$$
\begin{aligned}
r^k(t,i,c) &= \bar{G}_k(t)^{-1} \int_0^\infty \lambda_k e^{-\lambda_k s} ds \{\bar{G}_k(t+s) \int_0^s e^{-\alpha v} dv h^k(i+1) \\
&\quad + \int_{t+}^{t+s} dG_k(u) \\
&\quad \cdot [\int_0^{u-t} e^{-\alpha v} dv h^k(i+1) + e^{-\alpha(u-t)} R^k(i+1)]\}, \\
r^k(t,i,r) &= \bar{G}_k(t)^{-1} \int_0^\infty \lambda_k e^{-\lambda_k s} ds \{\bar{G}_k(t) d_k \\
&\quad + \bar{G}_k(t+s) \int_0^s e^{-\alpha v} dv h^k(i)
\end{aligned}
$$

$$+\int_{t+}^{t+s} dG_k(u)[\int_0^{u-t} e^{-\alpha v} dv h^k(i) + e^{-\alpha(u-t)} R^k(i)]\},$$

$$\beta(x,a,k') = \int_t^{\infty} e^{-\alpha(u-t)}[1-e^{-\lambda_k(u-t)}]dG_{kk'}(u) := \beta_{kk'}(t).$$

For $k \in K_2$, $A_k(i) = A \subset [0,\infty)$,

$$q_{ij}^k(\mu) = \begin{cases} \mu & j = i-1, i \geq 1 \\ -(\lambda_k + \mu) & j = i, i \geq 1 \\ \lambda_k & j = i+1, i \geq 1, \end{cases}$$

$$q_{0j}^k(\lambda) = \begin{cases} -\lambda_k & j = 0 \\ \lambda_k & j = 1, \end{cases}$$

and $r^k(t,i,\mu) = c^k(i,\mu) + R^k(i)g_k(t)/[1-G_k(t)]$, $p_{ij}^k(\mu) = \delta_{ij}$. Substituting the above formulae into Eqs. (6.69) and (6.70), we get the following optimality equation for the optimal control problem of M/M/1 in a semi-Markov environment.

$$V(k,t,i) = \max \left\{ \begin{array}{l} \bar{G}_k(t) r^k(t,i,c) + \lambda_k \int_0^{\infty} e^{-(\alpha+\lambda_k)s} V(k,t+s,i+1) ds \\ \bar{G}_k(t) r^k(t,i,r) + \lambda_k \int_0^{\infty} e^{-(\alpha+\lambda_k)s} V(k,t+s,i) ds \end{array} \right\}$$
$$+ \sum_{k'} \beta_{kk'}(t) V(k',0,i), \quad k \in K_1,$$

and

$$-\frac{d}{dt}V(k,t,i) = \sup_{\mu \in A}\{\bar{G}_k(t) r^k(t,i,\mu) + \sum_j q_{ij}^k(\mu) V(k,t,j)\}$$
$$+ \sum_{k'} g_{kk'}(t) V(k',0,i) - \alpha V(k,t,i), \quad k \in K_2.$$

Solving the above equations, we can get the optimal value and optimal policies.

4. Notes and References

Traditionally, MDPs describe closed systems without considering influences of the environments. But in many practical problems, the influences cannot be ignored. In fact there are many papers considering the influences of the environments on queueing systems, reliability systems, and inventory systems, and so on. For example, one can see Cao [15] and Neuts [98].

About MDPs in stochastic environments, we first studied MDPs under stochastic shocks in Hu [59] and [61]. Then, we studied CTMDPs in a semi-Markov environment in Hu [62] and [63]. Section 1 is from these two papers.

Section 2 is from Hu [67], in which SMDPs in semi-Markov environment were studied where the rewards are unbounded. Based on these studies, mixed MDPs was presented in Hu and Wang [72], which are explored in Section 3. In the above studies, the state spaces are all countable and the criterion is the discounted criterion. Xu and Hu [152] studied SMDPs-SE with the Borel state space and the total reward criterion with positive or negative reward.

Further research should include the following several aspects: (a) algorithms to compute the optimal value and the optimal policies, (b) the average criterion, and (c) particular properties and methods when dealing with various practical problems, for example, the optimal control of queueing systems as discussed in Examples 6.1 and 6.2.

Problems

1. In Example 6.1, we discuss an optimal service rate control of a queueing system $M/M/1$ in a semi-Markov environment. Now, when the environment is Markov, please simplify the optimality equation. Then how much further results can you get from the simplified optimality equation?

2. In Example 6.2, we discuss an optimal control of a queueing system $M/M/1$ in a semi-Markov environment. Now, if the environment is Markov, please simplify the optimality equation. Then how much further results can you get from the simplified optimality equation?

Chapter 7

OPTIMAL CONTROL OF DISCRETE EVENT SYSTEMS: I

Supervisory control for discrete event systems (DESs) belongs essentially to the logic level for control problems in DESs. In this chapter, we study a new optimal control problem in DESs. The performance measure is to maximize the maximal discounted total reward among all possible strings (i.e., paths) of the controlled system. The condition we need for this is only that the performance measure is well defined. By using the method and ideas presented in Chapter 2 for MDPs, we divide the problem into three subcases where the optimal values are, respectively, finite, positive infinite, and negative infinite. We then show the validity of the optimality equation in the case with a finite optimal value. Also, we characterize the optimality equation together with its solutions and characterize the structure of the set of all optimal policies. Based on the above results, we give a link between this performance model with the supervisory control for DESs. Finally, we apply these equations and solutions to a resource allocation system.

1. System Model

Discrete event systems are those systems that are driven by often occurring finite events. DESs correspond to the traditional dynamic systems that change continuously in time. Supervisory control of DESs was presented by Ramadge and Wonham [104], [105], and [151] with two branches: event feedback control and state feedback control. It belongs to the logic level of control for DESs [106].

A discrete event system is often described by strings of occurring events. Each time an event occurs, a new string forms and the next event that can occur is constrained in an event subset. The event set is further divided into two disjoint subsets called, respectively, the controllable event set and the uncontrollable event set. A control input is an event subset but includes the uncontrollable

event set. We let Γ be the set of all control inputs. In event feedback control, a control input should be chosen from Γ based on strings. This implies that the occurrence of events from the control input is prohibited. The system's behavior is described by languages, defined as sets of strings. In accordance with this description, the synthesizing problem, or more specifically, the supervisory control problem for controlling the system can be expressed as whether a given language can be synthesized by a supervisor. As shown in Ramadge and Wonham [104], a given language can be synthesized if and only if it is controllable and closed. The synthesizing problem together with its result is the essential basis for supervisory control theory.

Therefore, the essential task of supervisory control is to constrain the system's behavior in a given region, where the system's behavior is at most times described by a set of strings of occurring events or, at other times, by a state subset. At the same time, the control task in supervisory control is categorized as hard, where strings in the given behaviors are allowed and all other strings are strictly prohibited. However, there are many practical problems related to optimal control that belong to the performance level (i.e., optimizing some performance measures). And the control tasks in these practical problems are soft, where when some behaviors occur, it is better if their control tasks are soft.

In the following, we give the above description formally.

For a finite event set Σ, we denote by Σ^* the set of all finite strings on Σ including the empty string ϵ. For strings s, t, and r, if $s = t\,r$ then we call t a prefix of s and denote it by $t \leq s$. We call sets of Σ^* languages. For a language $L \subset \Sigma^*$, we define its closure, denoted by $\bar{L}$, a set of all prefixes of strings in L. We call L a closed language if $\bar{L} = L$. Let Σ^ω be the set of all infinite strings on Σ. For strings $s \in \Sigma^\omega$ and $t \in \Sigma^*$, if there is $r \in \Sigma^\omega$ such that $s = t\,r$, then we call t a prefix of s and denote it by $t \leq s$. We call sets of Σ^ω infinite languages.

A discrete event system based on automatons is

$$G = \{Q, \Sigma, \delta, q_0\},$$

where Q is a countable state space, Σ is a finite event set, δ is a partial function from $\Sigma \times Q$ to Q, and $q_0 \in Q$ is the initial state. We generalize δ by $\delta(s\sigma, q) = \delta(\sigma, \delta(s, q))$ inductively for $s\sigma \in \Sigma^*$, and when $\delta(s, q)$ is well defined we denote it by $\delta(s, q)!$. Moreover, we let

$$\Sigma(q) = \{\sigma \mid \delta(\sigma, q)!\}$$

be the set of events that may occur at state q and $\Sigma(s) = \{\sigma \mid \delta(s\sigma, q_0)!\}$ be the set of events that may occur after string s. We define the language generated by G from state q by

$$L(G, q) = \{s \in \Sigma^* \mid \delta(s, q)!\}$$

and the infinite language generated by G from state q by

$$\begin{aligned} L^{\omega}(G,q) &= \{s \in \Sigma^{\omega} \mid \delta(t,q)!, \forall t \leq s\} \\ &= \{s \in \Sigma^{\omega} \mid t \in L(G,q), \forall t \leq s\}, \quad q \in Q. \end{aligned}$$

Especially, for $q = q_0$, we call $L(G) := L(G, q_0)$ the language generated by G and $L^{\omega}(G) := L^{\omega}(G, q_0)$ the infinite language generated by G. It is assumed that G is alive; that is, $\Sigma(q)$ is nonempty for each state $q \in Q$. This assumption can be relaxed. In fact, if there are some empty $\Sigma(q)$ then we introduce a fictitious event $\sigma_J \notin \Sigma$ and let $\delta(\sigma_J, q) = q$ whenever $\Sigma(q)$ is empty.

The event set Σ is divided into an uncontrollable event set Σ_u and a controllable event set Σ_c. Σ_u and Σ_c are disjoint. A control input is an event subset γ satisfying $\Sigma_u \subset \gamma \subset \Sigma$. The set of such control inputs is denoted by Γ. A control input γ is chosen at state q (or string s) meaning that the next event should be in the set $\Sigma(q) \cap \gamma$ (or $\Sigma(s) \cap \gamma$).

We define a supervisor as a mapping $\boldsymbol{\pi} : L(G) \to \Gamma$ and a state feedback as a mapping $f : Q \to \Gamma$. Under the supervisor $\boldsymbol{\pi}$, $\boldsymbol{\pi}(s) \in \Gamma$ is chosen as the control input whenever string s occurs, whereas under state feedback f, $f(q) \in \Gamma$ is chosen as the control input whenever state q is visited. A state feedback f is a special supervisor $\boldsymbol{\pi}$ with $\boldsymbol{\pi}(s) = f(\delta(s, q_0))$ for all $s \in L(G)$. We denote the sets of supervisors and state feedbacks by Π and F, respectively.

We introduce several concepts in DESs (see [104], [105], and [151]). For a supervisor $\boldsymbol{\pi}$, we define the language $L(\boldsymbol{\pi}/G)$ generated by the system supervised under $\boldsymbol{\pi}$ inductively by

(a) $\epsilon \in L(\boldsymbol{\pi}/G)$ and
(b) If $s \in L(\boldsymbol{\pi}/G)$ and $\sigma \in \Sigma(s)$ and $\sigma \in \boldsymbol{\pi}(s)$, then $s\sigma \in L(\boldsymbol{\pi}/G)$.

The infinite language generated by the system supervised under $\boldsymbol{\pi}$ is denoted by $L^{\omega}(\boldsymbol{\pi}/G)$. Let $L^{\omega}(\boldsymbol{\pi}/G, q)$ be similar to $L^{\omega}(G, q)$. Moreover, for any $f \in F$, we define $f/G := \{Q, \Sigma, \delta_f, q_0\}$ as the system that is controlled under f. Here, $\delta_f(\sigma, q) = \delta(\sigma, q)$ is well defined if and only if $\sigma \in f(q)$ and $\delta(\sigma, q)!$.

Suppose that there is an extended real-valued reward function $c(q, \sigma) \in [-\infty, +\infty]$ defined in $Q \times \Sigma$ for an event σ occurring at state q for $q \in Q$ and $\sigma \in \Sigma(q)$. If the fictitious event σ_J has been introduced to ensure that G is alive, then we let $c(q, \sigma_J) = 0$ for state q whenever $\Sigma(q)$ is empty. The reward function characterizes the performance measure we introduce for the given DES.

Suppose that $\beta \geq 0$ is a given constant representing the discount factor. Let

$$v_q(t) = \sum_{k=0}^{\infty} \beta^k c(q_k, \sigma_k)$$

be the discounted total reward of occurring string $t = \sigma_0\sigma_1\cdots$ from the state q, where $q_{k+1} = \delta(\sigma_k, q_k)$ with $q_0 = q$ and $k = 0, 1, \ldots$. We simply call $v_q(t)$ *the reward for* t *at* q.

In general, there are infinite possible strings that may be generated by the system (G or f/G), but there is only one string that will be generated. We cannot know which string will be generated before the end of the system. Thus we consider, respectively, the maximal discounted total reward and the minimal discounted total reward of all possible strings that may be generated by the system controlled under f. Formally, we define

$$I(f,q) = \sup_{t\in L^\omega(f/G,q)} v_q(t), \quad q \in Q, \tag{7.1}$$

$$J(f,q) = \inf_{t\in L^\omega(f/G,q)} v_q(t), \quad q \in Q \tag{7.2}$$

as, respectively, the maximal discounted total reward and the minimal discounted total reward of the system supervised under f when state q is reached. Knowing the supremum and the infimum of the discounted total rewards may help us to resolve the system.

We define the optimal value functions, respectively, by

$$I^*(q) = \sup_{f\in F} I(f,q), \quad q \in Q, \tag{7.3}$$

$$J^*(q) = \sup_{f\in F} J(f,q), \quad q \in Q. \tag{7.4}$$

$I^*(q)$ and $J^*(q)$ are, respectively, the best case and the worst case we have for the discounted total reward. We call a state feedback f^* I-optimal at state q if $I(f^*, q) = I^*(q)$ and call f^* I-optimal if it is optimal at all $q \in Q$. J-optimal state feedbacks are defined similarly.

2. Optimality

The optimality in the model is studied by using ideas presented in [74] for Markov decision processes (MDPs). Following [74], we define the following general condition.

Condition 7.1: *$v_q(t)$, the discounted total reward for occurring t at state q is well defined for each state $q \in Q$ and each infinite string $t \in L^\omega(G, q)$.*

We should point out that $v_q(t)$ is well defined as a series where t is infinite. Condition 7.1 will be true, for example, when the reward function $c(\cdot,\cdot)$ is nonnegative, or is nonpositive, or is uniformly bounded and $\beta \in (0, 1)$. Condition 7.1 implies that both the objective functions $I(f, q)$ and $J(f, q)$ are well defined for each state feedback f and state $q \in Q$. Surely, Condition 7.1 is the basis for discussing the optimal control problem. Thus, the condition is assumed throughout this book.

The following lemma is obvious from Condition 7.1, where we view G as a weighted graph with nodes set Q, arc $q \xrightarrow{\sigma} q'$ with weight $c(q, \sigma)$ if and only if $q' = \delta(\sigma, q)$, and we call $v_q(t)$ the discounted total weight of the path t.

Lemma 7.1: *In any path in the weighted graph G, there are no two arcs with, respectively, positive infinite weight and negative infinite weight. Furthermore, any path with positive infinite discounted total weight includes no negative infinite arc.*

The latter result in the lemma above means that for each infinite string $t \in L^\omega(G, q)$ with $q = \delta(q_0, s)$, both $v_{q_0}(s)$ and $v_q(t)$ would not be simultaneously infinity with different symbols.

Now we introduce some concepts. Suppose that $\Sigma'(q) \subset \Sigma(q)$ is an event subset for each $q \in Q$. We call $\Sigma' := \{\Sigma'(q), q \in Q\}$ a constraint on G. When we constrain the DES G by Σ', we mean that the event set $\Sigma(q)$ is replaced by $\Sigma'(q)$ at state q. That is, the event that can occur at state q is among $\Sigma'(q)$. Moreover, for any $q \in Q$ and $r = \sigma_0\sigma_1 \cdots \sigma_k \in \Sigma^*$, let $q_0 = q$. If $\sigma_l \in \Sigma'(q_l)$ with $q_l = \delta(\sigma_{l-1}, q_{l-1})$ for $l = 1, 2, \ldots, k+1$, then we say that string r can occur at the state q through Σ'. The set of such strings r is denoted by $\Sigma'^*(q)$.

Definition 7.1: *1. For two states $q, q' \in Q$, if there is $r \in \Sigma'^*(q)$ such that $q' = \delta(r, q)$ then we say that q' can be reached from q through Σ', or q can reach q' through Σ'. We denote it by $q \xrightarrow{r}_{\Sigma'} q'$ or simply $q \rightarrow_{\Sigma'} q'$. It is clear that $q \xrightarrow{\epsilon}_{\Sigma'} q$ for each state $q \in Q$.*

2. For a predicate $P \subset Q$ and a state $q \in Q$, if there is state $q' \in P$ such that $q \rightarrow_{\Sigma'} q'$ then we say that P can be reached from q through Σ' and denote it by $q \rightarrow_{\Sigma'} P$. $P \rightarrow_{\Sigma'} q$ can be defined similarly.

3. For two predicates $P_1, P_2 \subset Q$, if P_1 can reach some state in P_2 through Σ' then we say that P_2 can be reached from P_1 through Σ', or P_1 can reach P_2 through Σ'. We denote it by $P_1 \rightarrow_{\Sigma'} P_2$.

For a predicate $P \subset Q$, let $P^*_{\Sigma'} = \{q' \mid P \rightarrow_{\Sigma'} q'\}$ be the set of states that can be reached from P through Σ', and $\bar{P}_{\Sigma'} = \{q \mid q \rightarrow_{\Sigma'} P\}$ be the set of states that can reach P through Σ'. It is obvious that $\bar{P}_{\Sigma'} \rightarrow_{\Sigma'} P \rightarrow_{\Sigma'} P^*_{\Sigma'}$. Moreover, $P \subset P^*_{\Sigma'}$ and $P \subset \bar{P}_{\Sigma'}$. For the reverse inclusions, we introduce the following definition.

Definition 7.2: *$P \subset Q$ is said to be a closed predicate under Σ' if any state that can reach P through Σ' is in P (i.e., $\bar{P}_{\Sigma'} = P$), whereas P is said to be an invariant predicate under Σ' if any state that can be reached from P through Σ' is in P (i.e., $P^*_{\Sigma'} = P$).*

It is clear that when there is an event subset $\Sigma_0 \subset \Sigma$ such that $\Sigma'(q) = \Sigma_0$ for all $q \in Q$, the closed predicate under Σ' defined above is identical to the Σ_0-invariant predicate defined in supervisory control literature [9]. Under the

constraint of Σ', to say a predicate P is closed means that the past of P is included in P itself, and saying P is invariant means that P includes its future; that is, the system will remain in P whenever the system begins in P.

2.1 Maximum Discounted Total Reward

We discuss the objective $I(f, q)$ in this subsection. It is easy to see that the state feedback f^* with $f^*(q) = \Sigma$ for each $q \in Q$ is I-optimal. Hence,

$$I^*(q) = I(f^*, q) = \max_{t \in L^\omega(G,q)} v_q(t), \quad q \in Q. \tag{7.5}$$

Then the remaining problems are, in fact, to get the value of $I^*(q)$ and to find strings with the maximal discounted total reward. We call a string $t \in L^\omega(G, q)$ a maximal string from state q if $v_q(t) = I^*(q)$. In the following, we characterize the optimal value I^* and discuss how to find a maximal string. Obviously, this problem is a dynamic programming problem, where the state space is Q, the action set at $q \in Q$ is the event set $\Sigma(q)$, the reward function is $c(q, \sigma)$, and the state transition is $q \xrightarrow{\sigma} \delta(\sigma, q)$. Because the reward function is unbounded and the number of horizons is infinite, we cannot directly use the optimality principle from dynamic programming. We prove the optimality equation by applying the ideas presented in Chapter 2 for Markov decision processes.

For $q \in Q$, we let an event subset

$$\Sigma_1(q) = \{\sigma \in \Sigma(q) \mid c(q, \sigma) > -\infty\}.$$

Lemma 7.2: *I^* satisfies the following I-optimality equation*

$$I(q) = \max_{\sigma \in \Sigma_1(q)} \{c(q, \sigma) + \beta I(\delta(\sigma, q))\}, \quad q \in Q. \tag{7.6}$$

Here, the maximum value is defined to be $-\infty$ when $\Sigma_1(q)$ is empty.

Proof: For each $q \in Q$, if $\Sigma_1(q)$ is empty, then $c(q, \sigma) = -\infty$ for each $\sigma \in \Sigma(q)$. Hence, for each $t \in L^\omega(G, q)$, $v_q(t) = -\infty$, and so $I^*(q) = -\infty$. Therefore, Eq. (7.6) is true for such a state q.

Then we consider any state q with $\Sigma_1(q)$ being nonempty. In this case, for any string $t \in L^\omega(G, q)$, if $t = \sigma t'$ with $c(q, \sigma) = -\infty$ then $v_q(t) = -\infty$ and such a string would not be maximal. Hence,

$$I^*(q) = \max_{\sigma t \in L^\omega(G,q), \sigma \in \Sigma_1(q)} v_q(t).$$

This results in that $\Sigma(q)$ can be sized down to $\Sigma_1(q)$ at state q. Moreover, if there is an event $\sigma \in \Sigma_1(q)$ such that $c(q, \sigma) = +\infty$, then it is easy to see that $I^*(q) = +\infty$ and so the I-optimality Eq. (7.6) is true for state q. Otherwise,

$c(q,\sigma)$ is finite for all $\sigma \in \Sigma_1(q)$. Then from Eq. (7.5) and the definition of $v_q(t)$ we have

$$\begin{aligned}
&I^*(q)\\
&= \max_{\sigma_0\sigma_1\cdots\in L^\omega(G,q),\sigma_0\in\Sigma_1(q)}\{c(q,\sigma_0)+\sum_{k=1}^{\infty}\beta^k c(\delta(\sigma_0\sigma_1\cdots\sigma_{k-1},q),\sigma_k)\}\\
&= \max_{\sigma_0\in\Sigma_1(q)}\ \max_{\sigma_1\sigma_2\cdots\in L^\omega(G,\delta(\sigma_0,q))}\{c(q,\sigma_0)\\
&\quad +\beta\sum_{k=1}^{\infty}\beta^{k-1}c(\delta(\sigma_0\sigma_1\cdots\sigma_{k-1},q),\sigma_k)\}\\
&= \max_{\sigma_0\in\Sigma_1(q)}\{c(q,\sigma_0)\\
&\quad +\beta\max_{\sigma_1\sigma_2\cdots\in L^\omega(G,\delta(\sigma_0,q))}\sum_{k=1}^{\infty}\beta^{k-1}c(\delta(\sigma_0\sigma_1\cdots\sigma_{k-1},q),\sigma_k)\}\\
&= \max_{\sigma\in\Sigma_1(q)}\{c(q,\sigma)+\beta I^*(\delta(\sigma,q))\}.
\end{aligned}$$

Hence, the I-optimality Eq. (7.6) is true. □

We separate the state set Q into several subsets. Let

$$\begin{aligned}
Q^0 &= \{q\in Q \mid I^*(q)\in(-\infty,+\infty)\},\\
Q^+ &= \{q\in Q \mid I^*(q)=+\infty\},\\
Q^- &= \{q\in Q \mid I^*(q)=-\infty\}
\end{aligned}$$

be the state subsets of finite, positive infinite, and negative infinite optimal values, respectively. Furthermore, let

$$Q^{+\infty}=\{q\in Q \mid \text{there is string } t\in L^\omega(G,q) \text{ such that } v_q(t)=+\infty\}.$$

Certainly, $Q^{+\infty}$ is a subset of Q^+. We have the following lemma on Q^+ and Q^-.

Lemma 7.3: *Under $\Sigma_1:=\{\Sigma_1(q),q\in Q\}$, Q^+ is closed and Q^- is invariant.*

Proof: We prove the lemma by the following two steps.

1. For any state $q\in Q$ with $q\to_{\Sigma_1} Q^+$, there is string $r\in\Sigma_1^*(q)$ such that $q':=\delta(r,q)\in Q^+$. Because $v_q(r)>-\infty$ and $I^*(q')=+\infty$, we have

$$\begin{aligned}
I^*(q) &= \max_{x\in L^\omega(G,q)} v_q(x)\\
&\geq \max_{y:ry\in L^\omega(G,q)} v_q(ry)\\
&= \max_{y\in L(G,q')}\{v_q(r)+\beta^{|r|}v_{q'}(y)\}
\end{aligned}$$

$$\begin{aligned} &= v_q(r) + \beta^{|r|} \max_{y \in L^\omega(G,q')} v_{q'}(y) \\ &= v_q(r) + \beta^{|r|} I^*(q') \\ &= \infty, \end{aligned}$$

where $|r|$ is the length of string r. Hence, $I^*(q) = \infty$ and $q \in Q^+$. Thus Q^+ is closed under Σ_1.

2. Suppose that Q^- is not invariant. Then there is state $q' \in Q - Q^-$ with $Q^- \rightarrow_{\Sigma_1} q'$. Surely, there are state $q \in Q^-$ and string $r \in \Sigma_1^*(q)$ such that $q' = \delta(r, q)$. Then from the proof for 1 we have that

$$I^*(q) \geq v_q(r) + \beta^{|r|} I^*(\delta(r, q)) > -\infty.$$

This contradicts $I^*(q) = -\infty$ for $q \in Q^-$. Hence, Q^- is invariant under Σ_1.

This completes the proof. □

From Lemma 7.3, we know that under the constraint Σ_1, each state q' that can reach some state $q \in Q^+$ belongs to Q^+ too, and each state q' that can be reached from some state $q \in Q^-$ belongs to Q^- too.

With Q^0, we let

$$\Sigma_2(q) = \begin{cases} \{\sigma \in \Sigma(q) \mid c(q,\sigma) > -\infty, \delta(\sigma, q) \in Q^0\}, & q \in Q^0 \\ \{\sigma \in \Sigma(q) \mid c(q,\sigma) > -\infty\} = \Sigma_1(q), & \text{otherwise.} \end{cases}$$

Surely, $\Sigma_2(q) \subset \Sigma_1(q)$ for all q.

It is obvious from Lemma 7.3 that under Σ_2, Q^+ is still closed and Q^- is still invariant. Summarizing the above results, we have the following obvious theorem.

Theorem 7.1: *Under either Σ_1 or Σ_2, Q^+ is closed and Q^- is invariant, although under Σ_2, Q^0 is invariant and the I-optimality equation in Q^0 is equivalent to*

$$I(q) = \max_{\sigma \in \Sigma_2(q)} \{c(q,\sigma) + \beta I(\delta(\sigma, q))\}, \quad q \in Q^0. \tag{7.7}$$

With the above theorem, we focus our attention on $\{I^*(q), q \in Q^0\}$. In the following, we further characterize solutions of the I-optimality Eq. (7.7) in Q^0.

Let $\Sigma_2^\omega(q)$ be the infinite language generated from state q through Σ_2. Moreover, we let a set $TQ_0 = \{(t,q) | t \in \Sigma_2^\omega(q)$ and $q \in Q^0$ with $v_q(t) \neq -\infty\}$ for notational simplicity in the following lemma.

Lemma 7.4: *We have the following four statements.*

1. *I^* satisfies the following condition,*

$$\limsup_{n \to \infty} \beta^n I(\delta(t_n, q)) \geq 0, \quad \forall (t,q) \in TQ_0. \tag{7.8}$$

2. *$I \geq I^*$ if I is a solution of the I-optimality Eq. (7.7) and satisfies condition Eq. (7.8).*
3. *$I \leq I^*$ if I is a solution of the I-optimality Eq. (7.7) and satisfies*

$$\liminf_{n\to\infty} \beta^n I(\delta(t_n, q)) \leq 0, \ \ \forall (t, q) \in TQ_0. \tag{7.9}$$

4. *$I = I^*$ if I is a solution of the I-optimality Eq. (7.7) and satisfies*

$$\lim_{n\to\infty} \beta^n I(\delta(t_n, q)) = 0, \ \ \forall (t, q) \in TQ_0. \tag{7.10}$$

Proof: 1. Let $t = \sigma_0\sigma_1 \cdots$ be an infinite string. The result follows the fact of $\beta^n I^*(\delta(t_n, q)) \geq \beta^n \sum_{k=n}^{\infty} c(\delta(t_k, q), \sigma_k)$ which tends to zero because $v_q(t)$ is finite by Condition 7.1 and $q \in Q^0$.

2. Suppose that I satisfies the given conditions. Then for each $q \in Q^0$ and $t \in \Sigma_2^\omega(q)$ with $v_q(t) \neq \infty$, it follows from the I-optimality Eq. (7.7) that

$$I(q) \geq v_q(t_n) + \beta^n I(\delta(t_n, q)), \ \ n \geq 0.$$

By taking $\limsup_{n\to\infty}$ in the above inequality we obtain $I(q) \geq v_q(t)$ due to condition Eq. (7.8). Because $t \in \Sigma_2^\omega(q)$ is arbitrary, we know that $I(q) \geq I^*(q)$.

3. This can be proved similarly to 2.
4. The result follows 2 and 3 above. □

We call I *asymptotic (discounted) nonnegative* if I satisfies condition Eq. (7.8), and similarly *asymptotic (discounted) nonpositive* or *zero* if I satisfies condition Eq. (7.9) or Eq. (7.10), respectively.

From the above lemma, especially result 1, the optimal value I^* is asymptotic nonnegative (condition Eq. (7.9)) and is equivalent to asymptotic zero (condition Eq. (7.10)). So, we have the following obvious theorem that characterizes the solution of the optimality equation.

Theorem 7.2:

1. *I^* is the smallest asymptotic nonnegative solution of the I-optimality Eq. (7.7).*
2. *I^* is the unique asymptotic zero (or nonpositive) solution of the I-optima-lity equation (7.7), if and only if Eq. (7.7) has an asymptotic zero (or nonpositive) solution.*

For the maximal strings, we let

$$\Sigma_3(q) = \{\sigma \mid \sigma \in \Sigma_2(q), I^*(q) = c(q, \sigma) + \beta I^*(\delta(\sigma, q))\}, \ \ q \in Q_0.$$

Because Σ is finite, $\Sigma_3(q)$ is nonempty. We call $\Sigma_3(q)$ the optimal event set at state q. Let $\Sigma_3^\omega(q)$ be the set of infinite strings generated from state q through Σ_3.

Theorem 7.3:

1. *There exist maximal strings from $q \in Q^{+\infty}$, there is no maximal string from $q \in Q^{+} - Q^{+\infty}$, and each string from $q \in Q^{-}$ is maximal.*

2. *For $q \in Q_0$, if I^* is asymptotic zero then any infinite string $t \in \Sigma_3^{\omega}(q)$ with $v_q(t) \neq -\infty$ is a maximal string from state q.*

Proof: 1. The results are obvious. 2. Due to the definition and the given conditions, we know that

$$I^*(q) = v_q(t_n) + \beta^n I^*(\delta(t_n, q)), \;\; n \geq 1.$$

By letting $n \to \infty$ in the equation above, we get that $I^*(q) = v_q(t)$. Hence, such a string t is a maximal string from state q. □

The theorem above characterizes maximal strings.

The following corollary follows Theorem 7.3 immediately.

Corollary 7.1: *Suppose that I^* is asymptotic zero and the state feedback f satisfies $f(q) \subseteq \Sigma_3(q)$ for all $q \in Q_0$. If $I(f, q) > -\infty$ then f is optimal at state q, for $q \in Q_0$.*

This corollary characterizes the structure of the set of optimal state feedback.

2.2 Minimum Discounted Total Reward

For the objective $J(f, q)$, it is easy to see that the state feedback f_u with $f_u(q) \equiv \Sigma_u$ for $q \in Q$ is optimal. We define a system $G_u = \{\Sigma_u, Q, \delta, q_0\}$ to be a subsystem of G by restricting the event set to Σ_u. Suppose that G_u is alive. Let $L(G_u, q)$ and $L^{\omega}(G_u, q)$ be, respectively, the finite and infinite languages generated by G_u with the initial state q. Then

$$J^*(q) = \max_{t \in L^{\omega}(G_u, q)} v_q(t), \;\; q \in Q.$$

Hence, all the results in the above subsection are true for J^* except that the event set Σ should be replaced by Σ_u. We omit the details here.

3. Optimality in Event Feedback Control

In this section, we generalize the model and results discussed in the previous section to the case of event feedback control. The basic difference between the model in this section and that in the previous sections is that here the extended real-valued reward function is $c(s, \sigma) \in [-\infty, +\infty]$ for an event σ occurring after string s for $s \in L(G)$ and $\sigma \in \Sigma(s)$.

For any infinite string $t = \sigma_0 \sigma_1 \sigma_2 \cdots \in \Sigma^{\omega}$, let its prefixes be $t_0 = \epsilon$ and $t_k = \sigma_0 \sigma_1 \cdots \sigma_{k-1}$ for $k \geq 1$. Let $\beta > 0$ be a discount factor. For each finite

string $s \in L(G)$ and each infinite string $t \in \Sigma^\omega$ with $st \in L^\omega(G)$, we define

$$v_s(t) = \sum_{k=0}^{\infty} \beta^k c(st_k, \sigma_k)$$

as the discounted total reward for t occurring after s. For a supervisor $\boldsymbol{\pi} \in \Pi$, we define

$$I(\boldsymbol{\pi}, s) = \sup_{t \in L^\omega(\boldsymbol{\pi}/G, s)} v_s(t), \quad s \in L(G), \tag{7.11}$$

$$J(\boldsymbol{\pi}, s) = \inf_{t \in L^\omega(\boldsymbol{\pi}/G, s)} v_s(t), \quad s \in L(G) \tag{7.12}$$

as, respectively, the maximal discounted total reward and the minimal discounted total reward of the system supervised under $\boldsymbol{\pi}$ when string s has occurred. Knowing the supremum and the infimum of the discounted total rewards may help us to resolve the system. Moreover, we define the optimal value functions, respectively, by

$$I^*(s) = \sup_{\boldsymbol{\pi} \in \Pi} I(\boldsymbol{\pi}, s), \quad s \in L(G), \tag{7.13}$$

$$J^*(s) = \sup_{\boldsymbol{\pi} \in \Pi} J(\boldsymbol{\pi}, s), \quad s \in L(G), \tag{7.14}$$

where $I^*(s)$ and $J^*(s)$ are, respectively, the best case and the worst case we have for the discounted total reward. We call a supervisor $\boldsymbol{\pi}^*$ *I-optimal at string* s if $I(\boldsymbol{\pi}^*, s) = I^*(s)$ and call $\boldsymbol{\pi}^*$ *I-optimal* if it is optimal at all $s \in L(G)$. *J-optimal supervisors* are defined similarly.

We separate the language $L(G)$ into several sublanguages. Let

$$\begin{aligned} L^0(G) &= \{s \in L(G) \mid I^*(s) \in (-\infty, +\infty)\}, \\ L^+(G) &= \{s \in L(G) \mid I^*(s) = +\infty\}, \\ L^-(G) &= \{s \in L(G) \mid I^*(s) = -\infty\} \end{aligned}$$

be the sets of strings with finite, positive infinite, and negative infinite optimal values, respectively. Further let $L^{+\infty}(G) = \{s \in L(G) \mid \text{there is } \boldsymbol{\pi} \text{ such that } I(\boldsymbol{\pi}, s) = +\infty\} = \{s \in L(G) \mid \text{there is } t \in \Sigma^\omega \text{ such that } st \in L^\omega(G) \text{ and } v_s(t) = +\infty\}$. It is clear that

$$\{s \mid c(s, \sigma) = +\infty \text{ for some } \sigma\} \subset L^{+\infty}(G) \subset L^+(G).$$

As for $\Sigma_1(q)$ and $\Sigma_2(q)$, we define

$$\begin{aligned} \Sigma_1(s) &= \{\sigma \in \Sigma(s) \mid c(s, \sigma) > -\infty\}, \\ \Sigma_2(s) &= \begin{cases} \{\sigma \in \Sigma(s) \mid c(s, \sigma) > -\infty, s\sigma \in L^0(G)\}, & s \in L^0(G) \\ \{\sigma \in \Sigma(s) \mid c(s, \sigma) > -\infty\} = \Sigma_1(s), & \text{otherwise.} \end{cases} \end{aligned}$$

Moreover, we introduce the following conditions similar to Eqs. (7.8) through (7.10), where the set $TL^0(G) = \{(t,s)|t \in \Sigma_2^\omega(s) \text{ with } s \in L^0(G) \text{ satisfying } v_s(t) \neq -\infty\}$,

$$\limsup_{n\to\infty} \beta^n I(s \cdot t_n) \geq 0, \ \ \forall (t,s) \in TL^0(G), \tag{7.15}$$

$$\liminf_{n\to\infty} \beta^n I(s \cdot t_n) \leq 0, \ \ \forall (t,s) \in TL^0(G), \tag{7.16}$$

$$\lim_{n\to\infty} \beta^n I(s \cdot t_n) = 0, \ \ \forall (t,s) \in TL^0(G). \tag{7.17}$$

We call I *asymptotic (discounted) nonnegative, nonpositive, zero* if I satisfies Eq. (7.15), Eq. (7.16), and Eq. (7.17), respectively. The closed or invariant languages under Σ_1 or Σ_2 can be defined similarly to the closed or invariant predicates given in Definition 7.2.

We have the following theorem that can be proved similarly to Theorems 7.1 and 7.2.

Theorem 7.4:

1. *Under either Σ_1 or Σ_2, $L^+(G)$ is closed and $L^-(G)$ is invariant, whereas under Σ_2, $L^0(G)$ is invariant and the optimality equation in $L^0(G)$ is equivalent to*

$$I(s) = \max_{\sigma \in \Sigma_2(s)} \{c(s,\sigma) + \beta I(s\sigma)\}, \quad s \in L^0(G). \tag{7.18}$$

2. *I^* is the smallest asymptotic nonnegative solution of optimality equation (7.18).*

3. *I^* is the unique asymptotic zero (or nonpositive) solution of the optimality equation (7.18), if and only if optimality equation (7.18) has an asymptotic zero (or nonpositive) solution.*

Similar results to Theorem 7.3 and Corollary 7.1 can be proved and their details are omitted here.

For J-optimal, we have similar results and so we omit the details except for the following corollary, where the supervisor π_u is defined by $\pi_u(s) = \Sigma_u$ for all s and the supervisor π_m^* is defined by

$$\pi_m^*(s) = \Sigma_u \bigcup \{\sigma \in \Sigma_c | c(s,\sigma) + \beta J^*(s\sigma) = J^*(s)\}, \ \ s \in L^0(G).$$

Corollary 7.2: *For the criterion $J(\pi, s)$, any supervisor π with $\pi_u \leq \pi \leq \pi_m^*$ is J-optimal. Hence, the minimal J-optimal supervisor is π_u and the maximal J-optimal supervisor is π_m^*.*

This corollary characterizes the structure of the set of J-optimal supervisors. The set is well structured because it is, in fact, an interval. It is easy to prove the corollary.

4. Link to Logic Level

In this section, we use the results for J^* obtained in the previous sections to describe and solve uniformly the basic supervisory control problems in event feedback control and state feedback control, by viewing the latter as a stationary version of the former. We show a link between the logic level and the performance level of control in DESs.

The supervisory control problem in event feedback control is for any given language $L \subset L(G)$ to solve

$$\max L(\boldsymbol{\pi}/G), \ \text{ s. t. } \ L(\boldsymbol{\pi}/G) \subset L, \ \boldsymbol{\pi} \in \Gamma^{L(G)}. \tag{7.19}$$

Its solutions are: (a) the maximal closed controllable sublanguage, denoted by $L^{\uparrow}$, of the given language L, and (b) a supervisor $\boldsymbol{\pi}^*$ with $L(\boldsymbol{\pi}^*/G) = L^{\uparrow}$ [104], [151]. Similarly, the supervisory control problem in state feedback control is for any given predicate P to solve

$$\max R(f/G), \ \text{ s. t. } \ R(f/G) \subset P, \ f \in \Gamma^{Q}. \tag{7.20}$$

Its solutions are: (a) the maximal controllable subpredicate, denoted by $P^{\uparrow}$, of the given predicate P, and (b) a state feedback f^* with $R(f^*/G) = P^{\uparrow}$ [105]. Here, $R(f/G)$ is the reachable states set of f/G.

Now we use the optimal control problem with the criterion J^* to describe and solve these two basic problems. Suppose that the reward function is nonnegative and satisfies the following condition,

$$\begin{aligned} & c(s,\sigma) \text{ is bounded in } s \in L \text{ and } \sigma \in \Sigma \\ \text{and} \quad & c(s,\sigma) = \infty \text{ for } s \notin L, \sigma \in \Sigma. \end{aligned} \tag{7.21}$$

It means that the language L is ideal whereas the strings out of L are strictly prohibited. We wish the language generated by the system to be in L. Let the discount factor be $\beta \in (0,1)$. Similarly to Eq. (7.6), we have the following J-optimality equation due to the definition of $c(s,\sigma)$:

$$J^*(s) = \max_{\sigma \in \Sigma_u(s)} \{c(s,\sigma) + \beta J^*(s\sigma)\}, \ \ s \in L,$$

where $\Sigma_u(s) = \Sigma_u \wedge \Sigma(s)$ for $s \in L(G)$.

It is clear that for any string s, if there is a uncontrollable string from s such that their conjunction is a prohibited string, then the optimal value at s, $J^*(s)$, will be positive infinite. Thus we denote by

$$L^* = \{s \mid J^*(s) < \infty\},$$

the set of strings with finite optimal values, because $J^*(s) > -\infty$ for all s.

We introduce the following condition for a language K,

$$K\Sigma_u \cap L(G) \subset K, \tag{7.22}$$

where $K\Sigma_u$ is defined as a language $\{st|s \in K, t \in \Sigma_u\}$. It is obvious that the above condition is similar to

$$\bar{K}\Sigma_u \cap L(G) \subset \bar{K},$$

the definition of a controllable language [104]. The difference between the above two formulae is only whether the closure of K is required. By noting the definition of a control invariant predicate [105], we know that the condition given in Eq. (7.22) is, in fact, the corresponding concept of a control invariant predicate in state feedback control. Thus we say K is a control invariant language if K satisfies Eq. (7.22). It is easy to see that for a closed language K, it is controllable if and only if it is control invariant. Moreover, from the definition, we know that K is control invariant if and only if there is a supervisor $\boldsymbol{\pi}$ such that $s\sigma \in K$ for each $s \in K$ and $\sigma \in \pi(s)$ with $s\sigma \in L(G)$. Such a supervisor $\boldsymbol{\pi}$ is said to be a permissive supervisor of K, the set of which is denoted by $\Pi(K)$, a corresponding concept of $F(P)$, the set of all permissive stationary policies of a control invariant predicate P.

For any language K, we let

$$MC(K) = \{s \in K \mid t \in K \text{ for all } t \leq s\}$$

be the maximal closed sublanguage of K.

The following theorem concerns the controllability languages.

Theorem 7.5: *L^* is the maximal control invariant sublanguage of L and $MC(L^*)$ is the maximal closed controllable sublanguage of L, i.e., $L^{\uparrow} = MC(L^*)$.*

Proof: 1. First, we show that L^* is a sublanguage of L. In fact, if $s \in L^* - L$ then $c(s, \sigma) = \infty$ for all σ. Due to the J-optimality equation, $J^*(s) = \infty$. This contradicts a fact of $s \in L^*$. Thus $L^* \subset L$.

Next, we show that L^* is a control invariant language. For each pair $s \in L^*$ and $\sigma \in \Sigma_u$ with $s\sigma \in L(G)$, if $s\sigma \notin L^*$, then $J^*(s\sigma) = \infty$. From the J-optimality equation we know that

$$J^*(s) \quad \geq \quad c(s, \sigma) + \beta J^*(s\sigma) = \infty,$$

which contradicts a fact of $s \in L^*$. So, $s\sigma \in L^*$. Thus, L^* is control invariant.

It is easy to see that $J^*(s)$ is bounded in $s \in L^*$ because the cost function $c(s, \sigma)$ is bounded in $s \in L$ and $\sigma \in \Sigma$. In fact, we can limit the system in $s \in L^*$. So for the solutions of the J-optimality equation, $\{J^*(s), s \in L^*\}$ is the unique one.

Finally, suppose that $L' \subset L$ is control invariant. Then,

$$s \in L', \sigma \in \Sigma_u(s) \implies s\sigma \in L'.$$

By letting M be a upper bound of $c(s,\sigma)$ in $s \in L$ and $\sigma \in \Sigma$, from the J-optimality equation, we have that

$$\begin{aligned} J^*(s) &= \max_{\sigma\in\Sigma_u(s)} \{c(s,\sigma) + \beta J^*(s\sigma)\} \\ &\leq M + \beta \max_{\sigma\in\Sigma_u(s)} J^*(s\sigma) \\ &\leq M + \beta \max_{s'\in L'} J^*(s'),\ s \in L', \end{aligned}$$

where $s\sigma \in L'$ for $\sigma \in \Sigma_u(s)$. This results in that $\max_{s\in L'}, J^*(s) \leq M + \beta \max_{s\in L'} J^*(s)$ and so

$$J^*(s) \leq (1-\beta)^{-1}M < \infty,\ s \in L'.$$

So $L' \subset L^*$. Thus, L^* is the maximal control invariant sublanguage of L.

2. First, it is easy to see that $MC(L^*)$ is a closed sublanguage of L. We only show that $MC(L^*)$ is also control invariant. In fact, suppose that $s \in MC(L^*)$ and $\sigma \in \Sigma_u$ with $s\sigma \in L(G)$. Due to $s \in L^*$ and Eq. (7.22) we know that $s\sigma \in L^*$. Thus $s\sigma \in MC(L^*)$ because $s \in MC(L^*)$. Therefore $MC(L^*)$ is a closed controllable sublanguage of L.

Any closed controllable sublanguage of L is a control invariant sublanguage of L, so it is also a sublanguage of L^*. This implies that $MC(L^*)$ is the maximal closed controllable sublanguage of L. □

$MC(L^*)$ is a part of L^* that can be realized by the system through strings in L^*. Theorem 7.5 says that L^* is not only the maximal control invariant sublanguage of L, but also the language with finite optimal values for an optimal control problem with the reward function satisfying Eq. (7.20). So the meaning of control invariant languages is stronger than just "control invariance" in the supervisory control [104].

Next, we apply the results for the stationary case to study state feedback control. Suppose that there is a nonnegative function $c(q,\sigma)$ defined on $Q \times \Sigma$ such that $c(s,\sigma) = c(\delta(s,q_0),\sigma)$ for $s \in L(G)$ and $\sigma \in \Sigma$. Moreover, there is a predicate $P \subset Q$ such that

$$\begin{aligned} &c(q,\sigma) \text{ is bounded in } q \in P \text{ and } \sigma \in \Sigma \\ \text{and}\quad &c(q,\sigma) = \infty \text{ for } q \notin P,\ \sigma \in \Sigma. \end{aligned} \tag{7.23}$$

The state subset P is ideal in which the reward is finite. We wish for the system to be in P, whereas the state out of P is strictly prohibited. Also, we assume that the discount factor is $\beta \in (0,1)$.

It is clear that for any state q, if there is a uncontrollable string from q under which the system will visit some prohibited states, then the optimal value at q, $J^*(q)$, will be positive infinite. We denote by

$$P^* = \{q \mid J^*(q) < \infty\}$$

the set of states with finite optimal values. From Theorems 7.3 to 7.5, we have the following theorem on state feedback control.

Theorem 7.6: *P^* is the maximal control invariant subpredicate of P.*

We have shown the stronger meanings of the control invariant languages and predicates above. But in order to compute them, we can only let the reward function satisfy

$$c(s,\sigma) = \begin{cases} 0, & s \in L \\ 1, & \text{otherwise,} \end{cases}$$

or

$$c(q,\sigma) = \begin{cases} 0, & q \in P \\ 1, & \text{otherwise.} \end{cases} \tag{7.24}$$

In these cases, we have, respectively,

$$L^* = \{s \in L \mid J^*(s) = 0\}, \;\; P^* = \{q \in P \mid J^*(q) = 0\}.$$

By applying Corollary 7.2 to the models satisfying the conditions given in Eq. (7.24), we have the following results about the maximal permissive supervisor and the maximal permissive state feedback.

Theorem 7.7: *With the objective of $J(\boldsymbol{\pi}, s)$ and the conditions given in Eq. (7.24), we have*

$$\begin{aligned} \Pi(L) &= \{\boldsymbol{\pi} \in \Pi \mid \boldsymbol{\pi}_u \le \boldsymbol{\pi} \le \boldsymbol{\pi}^*_m\}, \\ F(P) &= \{f \in F \mid f_u \le f \le f^*_m\}, \end{aligned}$$

where $\boldsymbol{\pi}^*_m$ and f^*_m are, respectively, the maximal permissive supervisor and the maximal permissive state feedback. Here, $\boldsymbol{\pi}^*_m$ is given in Corollary 7.2 and similarly,

$$f^*_m(q) = \Sigma_u \bigcup \{\sigma \in \Sigma_c \mid c(q,\sigma) + \beta J^*(\delta(\sigma,q)) = J^*(q)\}, \;\; q \in P.$$

Remember that we require the system G to be alive, otherwise we introduce the fictitious state q_J. Hence, for a general system G, we have, due to $J^*(s), J^*(q) \in \{0,1\}$, that

$$\begin{aligned} \boldsymbol{\pi}^*_m(s) &= \Sigma_u \cup \{\sigma \in \Sigma_c \mid s\sigma \notin L(G), \text{ or } s\sigma \in L^*\}, \\ f^*_m(q) &= \Sigma_u \cup \{\sigma \in \Sigma_c \mid \delta(\sigma,q) \text{ is undefined,} \\ &\qquad\qquad \text{or } \delta(\sigma,q) \in P^*\}. \end{aligned} \tag{7.25}$$

In the following, we consider a numerical example that is based on an example presented in paper [90] to illustrate the state feedback control of a DES. By the results given in Theorem 7.6 and Theorem 7.7, we can obtain easily the maximal control invariant subpredicate and the corresponding maximal permissive state feedback. But we should note that the example presented in [90] is for a partially observable DES without uncontrollable events.

Example 7.1: The state variable is given by $x = (x_1, x_2, x_3)$ where each x_i $(i = 1, 2, 3)$ takes nonnegative integers and the event set is given by $\Sigma = \{\tau, \alpha, \lambda, \theta\}$ with a uncontrollable event set $\Sigma_u = \{\tau\}$ and a controllable event set $\Sigma_c = \{\alpha, \lambda, \theta\}$, and the state transition function is described by

$$\begin{aligned}
&\delta(\alpha, (x_1, x_2, x_3)) = (x_1, x_2, x_3) + (-1, 1, 0), \ \text{ if } x_1 \geq 1,\\
&\delta(\lambda, (x_1, x_2, x_3)) = (x_1, x_2, x_3) + (-1, 1, 1), \ \text{ if } x_1 \geq 1,\\
&\delta(\theta, (x_1, x_2, x_3)) = (x_1, x_2, x_3) + (0, -1, 1), \ \text{ if } x_2 \geq 1,\\
&\delta(\tau, (x_1, x_2, x_3)) = (x_1, x_2, x_3) + (1, 0, -1), \ \text{ if } x_3 \geq 1,
\end{aligned}$$

with the initial state $x^0 = (1, 0, 0)$.

We consider a given predicate $P = \{x \mid x_3 \leq 1\}$. For $x \in P$, if $x_3 = 1$ then $\delta(\tau, x) \in P$, otherwise, τ cannot occur at x. Thus by Eq. (7.25) we have that

$$P^* = P, \ \ A^*(x) = \{\sigma \in \Sigma \mid \delta(\sigma, x) \in P\}, \ x \in Q.$$

Therefore, the maximal control invariant subpredicate of P is itself; that is, $P^* = P$, and the maximal permissive state feedback of P is $f^*(x) = A^*(x)$, whereas the maximal controllable subpredicate of P is $R(f^*/G)$ as follows.

$$\begin{aligned}
&R(f^*/G) = \{(1, 0, 0), (0, 1, 0), (0, 1, 1), (0, 0, 1)\},\\
&f^*(1, 0, 0) = f^*(0, 1, 0) = f^*(0, 0, 1) = \{\tau, \alpha, \lambda, \theta\},\\
&f^*(0, 1, 1) = \{\tau, \alpha, \lambda\}.
\end{aligned}$$

Therefore, the optimal state feedback is only to prohibit the event θ at the state $(0, 1, 1)$.

At the end of this section, we give the following remark to compare the optimal control problem of DESs with MDPs.

Remark 7.1: *1. Both systems of DESs and MDPs have the following two common features. (a) Only one string will be finally generated, but there are many possible strings that the system may generate, even infinite, and (b) the discounted total cost for each string is the same.*

2. The differences between DESs and MDPs are as follows. First, which string will occur is nondeterministic in DESs whereas it is random in MDPs. Stated more clearly, when a history s occurs in DESs, the next event is nondeterministic among $\Sigma(s)$, but when an event occurs,

then the next state is determined, although in MDPs, after an action (corresponding to an event in DESs) is chosen, the next state is still random among the states in Q. Second, in DESs, a range of possible events is given whereas in MDPs, one particular action is chosen. Third, for the objectives, we choose the maximal discounted total cost among all possible strings in DESs whereas in MDPs we take the average (in probability) value of the discounted total cost of all possible strings.

3. We further see the nondeterministic features in both systems. The nondeterministic feature in DESs is that when the system is at some states, there are many possible events that can occur, but it is nondeterministic and is not described in DESs which event will actually occur. It can be said that many (infinitely often) strings are possible but it is nondeterministic which one among them will actually occur. If there exists some probability distribution among the possible occurring events at any state, then the system becomes random, and each possible string will occur according to some probability. Then the optimal control problem becomes a MDPs problem, where we can consider the expected discounted total cost. This is the difference, as well as a link, between the optimal control problems considered here and those considered in MDPs. Based on this link, we used the ideas and methods developed in MDPs to solve the optimal control problem in DESs.

5. Resource Allocation System

Reveliotis and Choi [110] studied the optimality of randomized deadlock avoidance policies for resource allocation systems (RASs) based on one example. In this section, we modify this example and consider it from another viewpoint, in which there are two machines, R_1 and R_2, and two job types, JT_1 and JT_2. Its DES model is given in Figure 7.1.

The state set is

$$S = \{(i,j), i,j = 0,1,2\} \cup \{(1^*,j),(i,2^*),(1^*,2^*) \mid i,j = 1,2\}.$$

In the state variable, the first component i represents that a job i is being processed in machine R_1 and 1^* represents that a job 1 has been processed and is waiting in machine R_1, and the second component j represents that a job j is being processed in machine R_2 and 2^* represents that a job 2 has been processed and is waiting in machine R_2. In state $(1^*,2^*)$, the system is deadlocked and should be resolved artificially; that is, the system should exchange the two blocked jobs in the two machines.

The event set is

$$\Sigma = \{\rho, \lambda_i, \mu_{ij} \mid i,j = 1,2\},$$

where event ρ is to resolve the deadlock, event λ_i represents the arrival of a job i, and event μ_{ij} represents the completion of a job i in machine R_j. Here it is

assumed that only the arrival event can be controlled. Hence the uncontrollable event set is $\Sigma_u = \{\rho, \mu_{ij} \mid i, j = 1, 2\}$, and the controllable event set is $\Sigma_c = \{\lambda_1, \lambda_2\}$.

Reveliotis and Choi [110] considered randomness in the example where λ_i and μ_{ij} are rates of exponential distributions for respective processing times and introduced control probabilities ω_1 and ω_2 for respective transition $(1, 0) \rightarrow (1, 2)$ and $(0, 2) \rightarrow (1, 2)$, and a control ρ at state $(1, 2)$ by swapping the two deadlocked jobs. Reveliotis and Choi discussed the optimal values of ω_1 and ω_2 to maximize the long-run average throughput of the system.

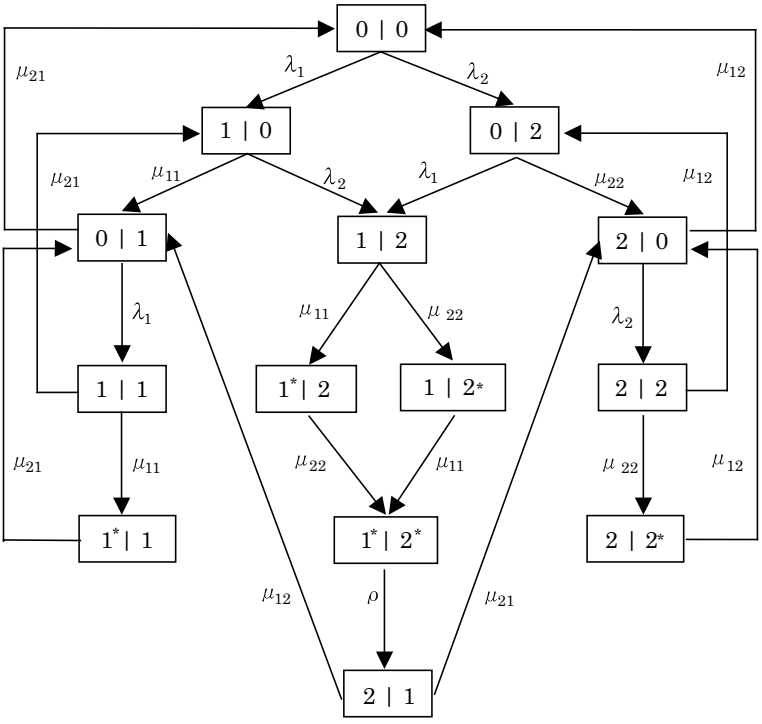

Figure 7.1. A resource allocation system: the DES model.

In order to avoid the deadlock, we can either prohibit event λ_1 from occurring at state $(0, 2)$ and event λ_2 from occurring at state $(1, 0)$ such that the system would not reach state $(1, 2)$ or allow event ρ to occur (i.e., to exchange the two blocked jobs) at state $(1^*, 2^*)$. Certainly, each action has an adequate cost. Then the problem of which action is better arises. We use the model and analysis discussed in the previous sections to solve this problem.

Now, prohibiting event λ_1 at state $(0,2)$ is equivalent to the occurrence of event μ_{22} at state $(0,2)$ and prohibiting event λ_2 at state $(1,0)$ is equivalent to the occurrence of event μ_{11} at state $(1,0)$. So, we introduce a cost c_1 for occurring event μ_{22} at state $(0,2)$ and a cost c_2 for occurring event μ_{11} at state $(1,0)$. Moreover, suppose that ρ is the cost for exchanging the two blocked jobs at state $(1^*,2^*)$ (we use the same symbol ρ to represent the cost and the event of exchange). It is assumed that c_1, c_2, and ρ are nonnegative. We define the cost function as

$$c((1,0),\mu_{11}) = c_2, \;\; c((0,2),\mu_{22}) = c_1, \;\; c((1^*,2^*),\rho) = \rho$$

and all other $c(q,\sigma) = 0$.

Here, we consider the maximal discounted total cost. Then the stationary I-optimality equation (7.6) is given as follows.

$$\begin{aligned}
V(0,0) &= \max\{\beta V(1,0), \beta V(0,2)\},\\
V(1,0) &= \max\{\beta V(1,2), c_2 + \beta V(0,1)\},\\
V(0,2) &= \max\{\beta V(1,2), c_1 + \beta V(2,0)\},\\
V(0,1) &= \max\{\beta V(1,1), \beta V(0,0)\},\\
V(2,0) &= \max\{\beta V(2,2), \beta V(0,0)\},\\
V(1,1) &= \max\{\beta V(1^*,1), \beta V(1,0)\},\\
V(2,2) &= \max\{\beta V(2,2^*), \beta V(0,2)\},\\
V(1,2) &= \max\{\beta V(1^*,2), \beta V(1,2^*)\},\\
V(2,1) &= \max\{\beta V(0,1), \beta V(2,0)\},\\
V(1^*,1) &= \beta V(0,1),\\
V(2,2^*) &= \beta V(2,0),\\
V(1^*,2) &= \beta V(1^*,2^*),\\
V(1,2^*) &= \beta V(1^*,2^*),\\
V(1^*,2^*) &= \rho + \beta V(2,1).
\end{aligned}$$

We simplify the above equations. By substituting the last equation for $V(1^*,2^*)$ into the equations for $V(1^*,2)$ and $V(1,2^*)$ we have that

$$V(1^*,2) = V(1,2^*) = \beta\rho + \beta^2 V(2,1).$$

Again by substituting these two equations together with the equations for $V(1^*,1)$ and $V(2,2^*)$ into the equations for $V(1,1), V(2,2)$, and $V(1,2)$ we can obtain that

$$\begin{aligned}
V(1,1) &= \max\{\beta^2 V(0,1), \beta V(1,0)\},\\
V(2,2) &= \max\{\beta^2 V(2,0), \beta V(0,2)\},\\
V(1,2) &= \beta^2\rho + \beta^3 V(2,1).
\end{aligned}$$

Hence, to solve the stationary I-optimality equation, it suffices to solve first the following set of equations.

$$\begin{cases} V(0,0) = \max\{\beta V(1,0), \beta V(0,2)\}, \\ V(1,0) = \max\{\beta^3\rho + \beta^4 V(2,1), c_2 + \beta V(0,1)\}, \\ V(0,2) = \max\{\beta^3\rho + \beta^4 V(2,1), c_1 + \beta V(2,0)\}, \\ V(0,1) = \max\{\beta V(1,1), \beta V(0,0)\}, \\ V(2,0) = \max\{\beta V(2,2), \beta V(0,0)\}, \\ V(1,1) = \max\{\beta^2 V(0,1), \beta V(1,0)\}, \\ V(2,2) = \max\{\beta^2 V(2,0), \beta V(0,2)\}, \\ V(2,1) = \max\{\beta V(0,1), \beta V(2,0)\}. \end{cases} \tag{7.26}$$

The above set of eight equations can be computed by successive approximation. That is, for any given initial values of $V_0(i,j)$ for i,j, (e.g., $V_0(i,j) = 0$ for all i,j), we iteratively compute $V_{n+1}(i,j)$ for $n = 0, 1, \ldots$ by

$$\begin{cases} V_{n+1}(0,0) = \max\{\beta V_n(1,0), \beta V_n(0,2)\}, \\ V_{n+1}(1,0) = \max\{\beta^3\rho + \beta^4 V_n(2,1), c_2 + \beta V_n(0,1)\}, \\ V_{n+1}(0,2) = \max\{\beta^3\rho + \beta^4 V_n(2,1), c_1 + \beta V_n(2,0)\}, \\ V_{n+1}(0,1) = \max\{\beta V_n(1,1), \beta V_n(0,0)\}, \\ V_{n+1}(2,0) = \max\{\beta V_n(2,2), \beta V_n(0,0)\}, \\ V_{n+1}(1,1) = \max\{\beta^2 V_n(0,1), \beta V_n(1,0)\}, \\ V_{n+1}(2,2) = \max\{\beta^2 V_n(2,0), \beta V_n(0,2)\}, \\ V_{n+1}(2,1) = \max\{\beta V_n(0,1), \beta V_n(2,0)\}. \end{cases} \tag{7.27}$$

Similarly to that in Markov decision processes (see Section 2.4), it can be proven that

$$\lim_{n\to\infty} V_n(i,j) = V(i,j), \quad \forall i,j.$$

So, for a given small constant $\varepsilon > 0$, when

$$|V_{n+1}(i,j) - V_n(i,j)| < \varepsilon, \quad \forall i,j,$$

we stop the above iterative computing procedure and take $V_{n+1}(i,j)$ as an approximate value of $V(i,j)$ for i,j. Moreover, by substituting these values into previous equations we can compute $V(1^*,2), V(1,2^*), V(1^*,2^*)$, and $V(1,2)$.

But, fortunately, the above set of equations (7.26) can be solved directly.

Suppose first that $c_1 \geq c_2$. Then from the successive approximation of Eq. (7.27) with $V_0(i,j) = 0$ for all i,j, we have that for each $n \geq 1$,

$$V_n(0,2) \geq V_n(1,0), \quad V_n(2,0) \geq V_n(0,1), \quad V_n(2,2) \geq V_n(1,1).$$

Hence,

$$V(0,2) \geq V(1,0), \quad V(2,0) \geq V(0,1), \quad V(2,2) \geq V(1,1).$$

This together with Eq. (7.26) implies that

$$V(0,0) = \beta V(0,2), \quad V(2,1) = \beta V(2,0).$$

In the following, we solve the set of equations (7.26) in three cases.

Case 1. $\beta^3\rho + \beta^4 V(2,1) \leq c_2 + \beta V(0,1)$. In this case, the set of equations (7.26) becomes that

$$\begin{aligned}
V(0,0) &= \beta V(0,2), \\
V(1,0) &= c_2 + \beta V(0,1), \\
V(0,2) &= c_1 + \beta V(2,0), \\
V(1,1) &= \max\{\beta^2 V(0,1), \beta c_2 + \beta^2 V(0,1)\} = \beta c_2 + \beta^2 V(0,1), \\
V(2,2) &= \max\{\beta^2 V(2,0), \beta c_1 + \beta^2 V(2,0)\} = \beta c_1 + \beta^2 V(2,0), \\
V(0,1) &= \max\{\beta^2 c_2 + \beta^3 V(0,1), \beta^2 c_1 + \beta^3 V(2,0)\} = \beta^2 c_1 + \beta^3 V(2,0), \\
V(2,0) &= \max\{\beta^2 c_1 + \beta^3 V(2,0), \beta^2 c_1 + \beta^3 V(2,0)\} = \beta^2 c_1 + \beta^3 V(2,0), \\
V(2,1) &= \beta V(2,0) = \beta^3 c_1 + \beta^4 V(2,0).
\end{aligned}$$

By solving it we obtain that

$$\begin{aligned}
&V(0,0) = V(2,2) = \tfrac{\beta}{1-\beta^3} c_1, \quad && V(1,0) = c_2 + \tfrac{\beta^3}{1-\beta^3} c_1, \\
&V(0,2) = \tfrac{1}{1-\beta^3} c_1, && V(0,1) = V(2,0) = \tfrac{\beta^2}{1-\beta^3} c_1, \\
&V(1,1) = \beta c_2 + \tfrac{\beta^4}{1-\beta^3} c_1, && V(2,1) = \tfrac{\beta^3}{1-\beta^3} c_1.
\end{aligned}$$

Then

$$\begin{aligned}
&V(1,2) = \beta^2\rho + \tfrac{\beta^6}{1-\beta^3} c_1, \quad && V(1^*,2) = V(1,2^*) = \beta\rho + \tfrac{\beta^5}{1-\beta^3} c_1, \\
&V(1^*,1) = \tfrac{\beta^3}{1-\beta^3} c_1, && V(2,2^*) = \tfrac{\beta^3}{1-\beta^3} c_1, \\
&V(1^*,2^*) = \rho + \tfrac{\beta^4}{1-\beta^3} c_1.
\end{aligned}$$

Moreover, the condition $\beta^3\rho + \beta^4 V(2,1) \leq c_2 + \beta V(0,1)$ is equivalent to $\rho \leq (1/\beta^3)c_2 + ((1-\beta^4)/(1-\beta^3))c_1$.

Case 2. $c_2 + \beta V(0,1) \leq \beta^3\rho + \beta^4 V(2,1) \leq c_1 + \beta V(2,0)$. In this case, the set of equations (7.26) becomes that

$$\begin{aligned}
V(0,0) &= \beta V(0,2), \\
V(2,1) &= \beta V(2,0), \\
V(1,0) &= \beta^3\rho + \beta^4 V(2,1) = \beta^3\rho + \beta^5 V(2,0), \\
V(0,2) &= c_1 + \beta V(2,0), \\
V(0,1) &= \max\{\beta V(1,1), \beta^2 c_1 + \beta^3 V(2,0)\}, \\
V(2,0) &= \max\{\beta V(2,2), \beta^2 c_1 + \beta^3 V(2,0)\}, \\
V(1,1) &= \max\{\beta^2 V(0,1), \beta^4\rho + \beta^6 V(2,0)\}, \\
V(2,2) &= \max\{\beta^2 V(2,0), \beta c_1 + \beta^2 V(2,0)\} = \beta c_1 + \beta^2 V(2,0).
\end{aligned}$$

By solving it we obtain that

$$\begin{array}{ll} V(0,0) = V(2,2) = \frac{\beta}{1-\beta^3}c_1, & V(1,0) = \beta^3\rho + \frac{\beta^7}{1-\beta^3}c_1, \\ V(0,2) = \frac{1}{1-\beta^3}c_1, & V(1,1) = \beta^4\rho + \frac{\beta^8}{1-\beta^3}c_1, \\ V(0,1) = V(2,0) = \frac{\beta^2}{1-\beta^3}c_1, & V(2,1) = \frac{\beta^3}{1-\beta^3}c_1. \end{array}$$

Then

$$\begin{array}{ll} V(1,2) = \beta^2\rho + \frac{\beta^6}{1-\beta^3}c_1, & V(1^*,2) = V(1,2^*) = \beta\rho + \frac{\beta^5}{1-\beta^3}c_1, \\ V(1^*,1) = V(2,2^*) = \frac{\beta^3}{1-\beta^3}c_1, & V(1^*,2^*) = \rho + \frac{\beta^4}{1-\beta^3}c_1. \end{array}$$

Moreover, the condition $c_2 + \beta V(0,1) \leq \beta^3\rho + \beta^4 V(2,1) \leq c_1 + \beta V(2,0)$ is equivalent to $(1/\beta^3)c_2 + ((1-\beta^4)/(1-\beta^3))c_1 \leq \rho \leq ((1-\beta^7)/(\beta^3(1-\beta^3)))c_1$.

Case 3. $\beta^3\rho + \beta^4 V(2,1) \geq c_1 + \beta V(2,0)$. In this case, the set of equations (7.26) becomes that

$$\begin{aligned} V(0,0) &= \beta V(0,2), \\ V(2,1) &= \beta V(2,0), \\ V(1,0) &= V(0,2) = \beta^3\rho + \beta^5 V(2,0), \\ V(0,1) &= \max\{\beta V(1,1), \beta^5\rho + \beta^7 V(2,0)\}, \\ V(2,0) &= \max\{\beta V(2,2), \beta^5\rho + \beta^7 V(2,0)\}, \\ V(1,1) &= \max\{\beta^2 V(0,1), \beta^4\rho + \beta^6 V(2,0)\}, \\ V(2,2) &= \max\{\beta^2 V(2,0), \beta^4\rho + \beta^6 V(2,0)\}. \end{aligned}$$

By solving it we obtain that

$$\begin{array}{ll} V(0,0) = V(2,2) = V(1,1) = \frac{\beta^4}{1-\beta^7}\rho, & V(1,0) = V(0,2) = \frac{\beta^3}{1-\beta^7}\rho, \\ V(0,1) = V(2,0) = \frac{\beta^5}{1-\beta^7}\rho, & V(2,1) = \frac{\beta^6}{1-\beta^7}\rho. \end{array}$$

Then

$$\begin{array}{ll} V(1,2) = \frac{\beta^2}{1-\beta^7}\rho, & V(1^*,2) = V(1,2^*) = \frac{\beta}{1-\beta^7}\rho, \\ V(1^*,1) = V(2,2^*) = \frac{\beta^6}{1-\beta^7}\rho, & V(1^*,2^*) = \frac{1}{1-\beta^7}\rho. \end{array}$$

Moreover, the condition $\beta^3\rho + \beta^4 V(2,1) \geq c_1 + \beta V(2,0)$ is equivalent to $\rho \geq ((1-\beta^7)/\beta^3(1-\beta^3))c_1$.

From the above three cases, we have the following proposition, which is obvious because each cost increases the discounted total cost.

Proposition 7.1: *For each state (i,j), $V(i,j)$ is increasing, respectively, in the costs c_1, c_2, and ρ.* □

Let

$$\rho_1^* = \frac{1-\beta^7}{\beta^3(1-\beta^3)}c_1, \quad \rho_2^* = \frac{1}{\beta^3}c_2 + \frac{1-\beta^4}{1-\beta^3}c_1$$

be two constants. The following proposition solves the comparison of prohibiting event λ_i with resolving the deadlock.

Proposition 7.2: *When $c_1 \geq c_2$, it is better to prohibit event λ_1 at state $(0,2)$ than to resolve the deadlock if and only if $\rho \geq \rho_1^*$ and it is better to prohibit event λ_2 at state $(1,0)$ than to resolve the deadlock if and only if $\rho \geq \rho_2^*$.*

Proof: From the optimality equation, it is better to prohibit event λ_1 at state $(0,2)$ than to resolve the deadlock if and only if the discounted total cost of a state feedback that prohibits λ_1 at $(0,2)$ is smaller than that of a state feedback which let λ_1 occur at $(0,2)$; that is,

$$c_1 + \beta V(2,0) \leq \beta V(1,2).$$

Due to the previous three cases, we know that the above condition is true if and only if $\rho \geq \rho_1^*$; that is, case 3 happens.

Similarly, it is better to prohibit event λ_2 at state $(1,0)$ than to resolve the deadlock if and only if

$$c_2 + \beta V(0,1) \leq \beta V(1,2),$$

which is equivalent to $\rho \geq \rho_2^*$; that is, cases 2 and 3 happen. □

Similarly, we can obtain the following proposition for $c_1 \leq c_2$, where the two constants are, respectively,

$$\rho_1^0 = \frac{1}{\beta^3}c_1 + \frac{1-\beta^4}{1-\beta^3}c_2, \quad \rho_2^0 = \frac{1-\beta^7}{\beta^3(1-\beta^3)}c_2.$$

Proposition 7.3: *When $c_1 \leq c_2$, it is better to prohibit event λ_1 at state $(0,2)$ than to resolve the deadlock if and only if $\rho \geq \rho_1^0$ and is better to prohibit event λ_2 at state $(1,0)$ than to resolve the deadlock if and only if $\rho \geq \rho_2^0$.*

The above two propositions say that there are threshold values for comparison of prohibiting event λ_i with resolving the deadlock.

Remark 7.2: *If there is a cost of the occurrence of any event, then the stationary I-optimality equation will be more complex and difficult to solve. But we can still prove the existence of the threshold values.*

6. Notes and References

Considering the reward of occurring events at states, Passino and Antsaklis [100] studied the optimal control problem of minimizing the total reward among strings from the initial state to some given target state subset and presented a heuristic algorithm to search for the string with the minimal total reward by using a branching-bounding algorithm. Tsitsiklis [140] presented a dynamic programming model to solve some special synthesizing problem in the supervisory control. It is shown that the problem is NP-hard. Kumar and Garg [86] studied an optimal static control problem with two reward functions $c(q, \sigma)$ and $p(q)$, but they assumed that these two rewards occurred only once. So the problem is, in fact, a static designing problem. They used the maximal-flow-minimal-cut theorem to solve the problem. Based on the theorem, they presented algorithms to compute the maximal subcontrollable languages for the supervisory control.

Yamsaki and Ushio [153] proposed a method to construct a supervisor based on reinforcement learning for state feedback control of partially observed DESs. They introduced a probability structure for the system based on some other parameters. Moon and Wardi [95] studied optimal control of processing times in single-stage DESs with blocking based on queues. They decomposed the problem into a finite sequence of reduced-order problems based on convex programming.

Some studies on optimal control problems have been used to solve stability problems of DESs. Considering the reward for occurring events at states, Brave and Heymann [9] calculated the optimal attraction by minimizing the total reward among all possible strings from an arbitrary state to a given global attraction, found conditions for the existence of supervisors achieving optimal attraction, and provided efficient algorithms for their synthesis. Hu and Liu [70] used Markov decision processes to study the static stability problems in DESs.

But all of the above researches either related special optimal control problems to special reward functions to solve some problem in supervisory control, or they were concerned with the static control of DESs, and were not concerned with general frameworks for optimal control of DESs [86].

Other related works on the optimal control of DESs include the paper of Fu et al. [44], where they presented a state-based method for optimal control of regular languages with the performance measure being a signed real measure of the supervised sublanguage. Ray et al. [107] generalized the model studied in [44] by considering the disabling event cost. Under their performance measure, the costs from one state to the next state in one transition is summed over all possible occurring events and the cost for a string occurring is the product of all costs for occurring events in the string. Moreover, their supervisors are given without any structure.

Over the last twenty years, many aspects of supervisory control have been researched, such as: partially observable information [90], [91], and [134], decentralized supervisory control [116], hierarchical supervisory control [19], stability analysis [101], robustness [135], and fault diagnosis [4] and [155]. Due to the complexity of the models and the methods in supervisory control, reduction of the model is often discussed in the literature, for example, see Su and Wonham [133]. Our model can be considered to be generalized for these models.

The contents of this chapter are from Hu and Yue [77] and [78].

Problems

1. In the model discussed in this chapter, we study the problem among the stationary policies. Whether or not the optimal value will be improved when Markov policies are considered?

2. In the resource allocation system studied in Section 5, only three costs are considered. They are the cost c_1 for occurring event μ_{22} at state $(0, 2)$, the cost c_2 for occurring event μ_{11} at state $(1, 0)$, and the cost ρ for exchanging the two blocked jobs at state $(1^*, 2^*)$. Now, suppose that there is a cost for occurring any event at any state, write the optimality equation and show the existence of the threshold values.

3. The cost function considered in this chapter for the controlled DESs is $c(q, \sigma)$ for occurring event σ at state q. A more general case is $c(s, \sigma)$ for occurring event σ at string s. Set this up as a Markov decision process model and write the optimality equation. Do you think that the same results as in Subsection 2.1 and those in Section 3 can be proved?

4. Study the resource allocation system with the minimal discounted total reward criterion J^*.

5. For any given DES $G = (Q, \Sigma, \delta, q_0)$ with the cost function $c(q, \sigma)$. Let $\Sigma(q) = \{\sigma \in \Sigma | \delta(\sigma, q)!\}$ for each $q \in Q$. We have defined two criteria I^* and J^*, which are the best case and the worst case, respectively, we have for the discounted total rewards.

Now, we revise the DES by introducing probabilities as follows. For each $q \in Q$, there is a probability distribution in $\Sigma(q)$, denoted by $\{\pi(q, \sigma), \sigma \in \Sigma(q)\}$. $\pi(q, \sigma)$ represents the probability that the event σ occurs at state q. If the event σ occurs at state q then the next state is $\delta(\sigma, q)$. At this new state, the above process repeats. For such a probability distribution π, we can define the expected discounted total cost from the initial state q_0 as $V(\pi, q) = E_\pi\{\sum_{k=0}^{\infty} \beta^k c(q_k, \sigma_k) | q_0 = q\}$, which is similar as that in MDPs. Show that for any π, we have $J^*(q) \leq V(\pi, q) \leq I^*(q)$ for all $q \in Q$.

Chapter 8

OPTIMAL CONTROL OF DISCRETE EVENT SYSTEMS: II

In this chapter, we present another model for optimal control of DESs with an arbitrary control pattern. This model differs from that in Chapter 7 in three ways. First, the discrete event system is defined as a collection of event sets that depend on strings. When the system generates a string, the next event that occurs should be in the corresponding event set. Second, the rewards are for choosing control inputs at strings. Finally, the control pattern (which consists of sets of available control inputs) depends on strings. The performance measure is to find a supervisor under the condition where the discounted total reward among strings from the initial state is maximized. Similarly to Chapter 7, we study the problem also by applying ideas from Chapter 2 for Markov decision processes. Surely, the problem here is more complex than that in Chapter 7. Moreover, we present a new supervisory control problem of DESs with the control pattern being dependent on strings. We study the problem in both event feedback control and state feedback control by generalizing concepts of invariant and closed languages/predicates from the supervisory control literature. Finally, we apply our model and the results to a job-matching problem.

1. System Model

Let Σ be a finite set of events and Σ^* be the set of all finite length strings formed with elements of Σ, including the null string ε. Any subset of Σ^* is called a language on Σ. The discrete event system we consider here is defined by

$$G = \{\Sigma(s),\ s \in \Sigma^*\}, \tag{8.1}$$

where $\Sigma(s) \subset \Sigma$ is the set of events that can occur after string s. The system evolves as follows.

Initially, an event $\sigma \in \Sigma(\varepsilon)$ occurs.

Inductively, if string $s \in \Sigma^*$ occurs then the next event should be in the event set $\Sigma(s)$.

Hence, the language generated by the system, denoted also by $L(G)$, is defined recursively as follows.

(a) $\varepsilon \in L(G)$, and
(b) If $s \in L(G)$ then $s\sigma \in L(G)$ for each $\sigma \in \Sigma(s)$.

It is apparent that such a language $L(G)$ is well defined. Moreover, for any string $r \in L(G)$, we define by $L(G, r)$ the language similarly to $L(G)$ where (a) $\varepsilon \in L(G)$ is replaced by (a) $r \in L(G)$. Certainly,

$$L(G, r) = \{s \in \Sigma^* \mid rs \in L(G)\}.$$

The following remark compares the DES given in Eq. (8.1) with the automaton type given in Chapter 7.

Remark 8.1: *1. In order to determine the language $L(G)$, it suffices to know $\Sigma(s)$ for $s \in L(G)$. In fact, $\{\Sigma(s),\ s \in L(G)\}$ and $L(G)$ are determined by each other because if $L(G)$ is given then*

$$\Sigma(s) = \{\sigma \in \Sigma \mid s\sigma \in L(G)\}, \quad s \in L(G).$$

2. In the supervisory control literature, a system is defined by an automaton:

$$G := \{\Sigma, Q, \delta, q_0\}, \tag{8.2}$$

as described in Section 7.1. But if we define $\Sigma(s) := \{\sigma \in \Sigma \mid \delta(s\sigma, q_0)!\}$ for $s \in L(G)$ then $L(G)$ can be determined completely by $\Sigma(s)$. Hence, a DES described by an automaton can be expressed also by the DES given in Eq. (8.1). On the other hand, the DES given in Eq. (8.1) can also be expressed by an automaton with its state space Σ^*, event set Σ, state transition function $s \xrightarrow{\sigma} s\sigma$, and the initial state ε. Thus, the expression given by Eq. (8.1) for DES is equivalent to that given by Eq. (8.2). But in general, an automaton is well structured and the DES in Eq. (8.1) has a universal form. In this chapter, we use Eq. (8.1) to study some theoretical problems because the expression Eq. (8.1) is simpler. On the other hand, the automaton is better structured and so we can expect better results when G is an automaton. This is explored in Section 8.4 below. □

The concepts of prefix, closure, infinite languages, and so on, are the same as those defined in Chapter 7.

The event set Σ is also divided into two disjoint subsets: an uncontrollable event set Σ_u and a controllable event set Σ_c. A control input is defined as a

subset of Σ satisfying $\Sigma_u \subset \gamma \subset \Sigma$. Let Γ be the set of all control inputs. The joint operation in Γ is defined as usual:

$$\gamma_1 \wedge \gamma_2 = \gamma_1 \cap \gamma_2.$$

There is a control pattern attached to the DES G. Suppose that for $s \in L(G)$ there is $\Gamma(s) \subset \Gamma$. $\Gamma(s)$ represents the set of available control inputs at string s. That is, it is required that the control input at string $s \in L(G)$ should be restricted in $\Gamma(s)$. We call $\{\Gamma(s), s \in L(G)\}$ a control pattern. Then, we define a controlled discrete event system (CDES) in a twofold manner:

$$G_c = \{\Sigma(s), \Gamma(s), \ \ s \in L(G)\}. \tag{8.3}$$

A supervisor for G_c is defined as a map $\boldsymbol{\pi} : L(G) \to \Gamma$ satisfying $\boldsymbol{\pi}(s) \in \Gamma(s)$ for each $s \in L(G)$. The set of supervisors is denoted by Π. For each supervisor $\boldsymbol{\pi} \in \Pi$, the system controlled under $\boldsymbol{\pi}$, denoted by $\boldsymbol{\pi}/G$, is

$$\boldsymbol{\pi}/G = \{\Sigma(s) \wedge \boldsymbol{\pi}(s), \ \ s \in \Sigma^*\},$$

which is also a DES according to Eq. (8.1). We denote by $L(\boldsymbol{\pi}/G)$ and $L^\omega(\boldsymbol{\pi}/G)$, respectively, the language and the infinite language generated by $\boldsymbol{\pi}/G$. Surely, $L(\boldsymbol{\pi}/G) \subset L(G)$ and $L^\omega(\boldsymbol{\pi}/G) \subset L^\omega(G)$.

As in Golaszewski and Ramadge [45], it is assumed that

$$\Sigma(s) = \bigcup\{\gamma \mid \gamma \in \Gamma(s)\}. \tag{8.4}$$

This condition requires that each event that can occur at string s should be included in some control input available at string s. Under this condition, it is easy to show that

$$L(G) = \bigcup_{\boldsymbol{\pi} \in \Pi} L(\boldsymbol{\pi}/G).$$

Otherwise, the above equality may not be true and in general we have that $\bigcup_{\boldsymbol{\pi} \in \Pi} L(\boldsymbol{\pi}/G) \subset L(G)$. However, this is not reasonable. So when condition Eq. (8.4) is true we call $\{\Gamma(s), s \in L(G)\}$ the reasonable control pattern. Certainly, if Eq. (8.4) is not true then we can reduce $\Sigma(s)$ such that Eq. (8.4) is true.

It should be noted that the control pattern $\{\Gamma(s)\}$ here generalizes that in the last chapter where $\Gamma(s) = \Gamma$ for all s.

Now, we define an optimal control problem for the DES G as the following triple

$$\{\Sigma(s), \Gamma(s), (c(s, \gamma), \gamma \in \Gamma(s)), \ \ s \in L(G)\}, \tag{8.5}$$

where $c(s, \gamma) \in [-\infty, +\infty]$ is an extended real-valued reward function, for choosing control input γ at string s for $s \in L(G)$ and $\gamma \in \Gamma(s)$.

We consider the optimal control problem Eq. (8.5) in infinite languages. Hence, we assume that G is alive; that is, $\Sigma(s)$ is nonempty for each string $s \in L(G)$. This assumption can be relaxed. In fact, we can introduce a fictitious event $\sigma_J \notin \Sigma$ and let $\Sigma_u := \Sigma_u \bigcup \{\sigma_J\}$, $\Sigma(s) := \{\sigma_J\}$ if $\Sigma(s) = \emptyset$, and let $c(s, \gamma) = c(s, \gamma - \{\sigma_J\})$ for each s and γ.

For each supervisor $\boldsymbol{\pi} \in \Pi$, we let

$$c(s, \boldsymbol{\pi}) = c(s, \boldsymbol{\pi}(s))$$

be the reward at string s under $\boldsymbol{\pi}$ and let

$$v_s(\boldsymbol{\pi}, t) = \sum_{k=0}^{\infty} \beta^k c(st_k, \boldsymbol{\pi})$$

be the discounted total reward for string $t = \sigma_1\sigma_2 \cdots \in L^\omega(G, s)$ under $\boldsymbol{\pi}$ when string s has occurred, where $t_k = \sigma_1\sigma_2 \cdots \sigma_k$ is a prefix of t for $k = 1, 2, \ldots, n$ and $t_0 = \varepsilon$. $\beta > 0$ is the discount factor. We simply call $v_s(\boldsymbol{\pi}, t)$ *the reward for t at s under $\boldsymbol{\pi}$*.

In general, there are infinite possible strings that may be generated by the system (G or $\boldsymbol{\pi}/G$), but there is only one string that will finally be generated. We cannot know which string will be generated before the end of the system. Thus we consider the maximal discounted total reward of all possible strings that may occur in the system controlled under $\boldsymbol{\pi}$. Formally, we define

$$V(\boldsymbol{\pi}, s) = \sup_{t \in L^\omega(\boldsymbol{\pi}/G, s)} v_s(\boldsymbol{\pi}, t), \ s \in L(G) \tag{8.6}$$

as the maximal discounted total reward of the system controlled under $\boldsymbol{\pi}$ when string $s \in L(G)$ has occurred, where $L^\omega(\boldsymbol{\pi}/G, s)$ is the infinite language similar to $L^\omega(\boldsymbol{\pi}/G)$ but with the initial string s.

We define the optimal value function by

$$V^*(s) \ = \ \sup_{\boldsymbol{\pi} \in \Pi} V(\boldsymbol{\pi}, s), \ \ s \in L(G). \tag{8.7}$$

$V^*(s)$ is the best case we have for the discounted total reward. We call $\boldsymbol{\pi}^*$ an optimal supervisor at string s if $V(\boldsymbol{\pi}^*, s) = V^*(s)$ and call $\boldsymbol{\pi}^*$ optimal if $\boldsymbol{\pi}^*$ is optimal at each $s \in L(G)$.

We introduce the following condition on the reward function.

Condition 8.1: *The discounted total reward $v_s(\boldsymbol{\pi}, t)$ for string t at s is well defined for each finite string $s \in L(G)$ and each infinite string $t \in \Sigma^\omega$ with $st \in L^\omega(G)$.*

We should point out that $v_s(\boldsymbol{\pi}, t)$ is well defined as a series because t is infinite. Condition 8.1 will be true, for example, when the reward function

$c(\cdot,\cdot)$ is nonnegative, or is nonpositive, or is uniformly bounded and $\beta \in (0,1)$. Condition 8.1 implies that the objective function $V(\boldsymbol{\pi}, s)$ is well defined for each supervisor $\boldsymbol{\pi}$ and string $s \in L(G)$. So, we say that the optimal control problem is well defined. Surely, this condition is the basis for discussing the optimal control problem and is thus assumed throughout this chapter.

2. Optimality Equation and Optimal Supervisors

In this section, we study the optimality equation and optimal supervisors for the optimal control problem, Eq. (8.5).

First, for $s \in L(G)$, we define

$$\Gamma_1(s) := \{\gamma \in \Gamma(s) \mid c(s,\gamma) > -\infty\} \tag{8.8}$$

as the set of control inputs at string s where the reward is not negative infinite. Let

$$L^{-\infty}(G) := \{s \in L(G) \mid \Gamma_1(s) = \emptyset\}$$

be a sublanguage of $L(G)$. $\Gamma_1(s) = \emptyset$ means that $c(s,\gamma) = -\infty$ for all $\gamma \in \Gamma(s)$. Hence, for each $s \in L^{-\infty}(G)$ and $\boldsymbol{\pi} \in \Pi$, $c(s,\boldsymbol{\pi}) = -\infty$ and so $V(\boldsymbol{\pi}, s) = -\infty$. This shows the following lemma.

Lemma 8.1: *$V^*(s) = -\infty$ for all $s \in L^{-\infty}(G)$ and so each supervisor $\boldsymbol{\pi} \in \Pi$ is optimal at $s \in L^{-\infty}(G)$.* □

From the above lemma, it suffices to discuss the optimality in $L(G) - L^{-\infty}(G)$. Moreover, for each $s \in L(G) - L^{-\infty}(G)$, if $\gamma \in \Gamma(s) - \Gamma_1(s)$ then $c(s,\gamma) = -\infty$. Hence, each supervisor $\boldsymbol{\pi} \in \Pi$ with $\boldsymbol{\pi}(s) \in \Gamma(s) - \Gamma_1(s)$ must satisfy $V(\boldsymbol{\pi}, s) = -\infty$ and so we would not consider such a supervisor. Let Π_1 be the set of supervisors $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi}(s) \in \Gamma_1(s)$ for all $s \in L(G) - L^{-\infty}(G)$. Surely,

$$V^*(s) = \sup_{\boldsymbol{\pi} \in \Pi_1} V(\boldsymbol{\pi}, s), \quad s \in L(G) - L^{-\infty}(G).$$

So, we limit our attention to supervisors in Π_1.

Hence, we limit our discussion on $L(G) - L^{-\infty}(G)$ to the set of available control inputs at string $s \in L(G) - L^{-\infty}(G)$ being $\Gamma_1(s)$ in the following. For notational simplicity, we let

$$\Sigma_{\boldsymbol{\pi}}(s) = \Sigma(s) \cap \boldsymbol{\pi}(s), \ \Sigma_{\gamma}(s) = \Sigma(s) \cap \gamma$$

for $\boldsymbol{\pi} \in \Pi, \gamma \in \Gamma(s)$ and $s \in L(G)$. $\Sigma_{\boldsymbol{\pi}}(s)$ is the set of events that is available at string s under controlled by supervisor π.

The following lemma characterizes the criterion $V(\boldsymbol{\pi}, s)$.

Lemma 8.2: *For any supervisor* $\boldsymbol{\pi} \in \Pi_1$,

$$V(\boldsymbol{\pi}, s) = c(s, \boldsymbol{\pi}) + \beta \max_{\sigma \in \Sigma_{\boldsymbol{\pi}}(s)} V(\boldsymbol{\pi}, s\sigma), \quad s \in L(G) - L^{-\infty}(G).$$

Proof: If $c(s, \boldsymbol{\pi}) = +\infty$ then both sides of the above equation are infinite and so the above equation holds. Otherwise, $c(s, \boldsymbol{\pi})$ is finite due to $\boldsymbol{\pi} \in \Pi_1$. Then, we have that for $s \in L(G) - L^{-\infty}(G)$,

$$\begin{aligned} V(\boldsymbol{\pi}, s) &= \sup_{t \in L^{\omega}(\boldsymbol{\pi}/G, s)} \sum_{k=0}^{\infty} \beta^k c(st_k, \boldsymbol{\pi}) \\ &= c(s, \boldsymbol{\pi}) + \beta \sup_{t \in L^{\omega}(\boldsymbol{\pi}/G, s)} \sum_{k=1}^{\infty} \beta^{k-1} c(st_k, \boldsymbol{\pi}) \\ &= c(s, \boldsymbol{\pi}) + \beta \max_{\sigma \in \Sigma_{\boldsymbol{\pi}}(s)} \sup_{t' \in L^{\omega}(\boldsymbol{\pi}/G, s\sigma)} \sum_{k=0}^{\infty} \beta^k c(s\sigma t'_k, \boldsymbol{\pi}) \\ &= c(s, \boldsymbol{\pi}) + \beta \max_{\sigma \in \Sigma_{\boldsymbol{\pi}}(s)} V(\boldsymbol{\pi}, s\sigma). \end{aligned}$$

Hence, the lemma is true. □

The above lemma separates the discounted total reward for an infinite string into two parts: the reward $c(s, \boldsymbol{\pi})$ for the first period and the maximal discounted total reward for the remaining periods.

We divide the language $L(G)$ into three disjoint sublanguages:

$$L(G) = L^{+}(G) \bigcup L^{-}(G) \bigcup L^{0}(G),$$

where $L^{+}(G) := \{s \in L(G) \mid V^{*}(s) = +\infty\}$ is the sublanguage with positive infinite optimal value, $L^{-}(G) := \{s \in L(G) \mid V^{*}(s) = -\infty\}$ is the sublanguage with negative infinite optimal value, and $L^{0}(G) := \{s \in L(G) \mid V^{*}(s) \in (-\infty, +\infty)\}$ is the sublanguage with finite optimal values. Surely, $L^{-\infty}(G) \subset L^{-}(G)$. Furthermore, we let $L^{+\infty}(G) := \{s \in L(G) \mid$ there is a supervisor $\boldsymbol{\pi} \in \Pi_1$ such that $V(\pi, s) = +\infty\}$ be a sublanguage of $L^{+}(G)$. We have the following results, which divide the optimality into three subcases in $L^{-}(G)$, $L^{+}(G)$, and $L^{0}(G)$, respectively.

Theorem 8.1:

1. *Each supervisor* $\boldsymbol{\pi} \in \Pi$ *is optimal in* $L^{-}(G)$, *there is an optimal supervisor in* $L^{+\infty}(G)$ *but there is no optimal supervisor in* $L^{+}(G) - L^{+\infty}(G)$.
2. V^{*} *satisfies the following optimality equation in the sublanguage* $L^{0}(G)$,

$$V(s) = \max_{\gamma \in \Gamma_1(s)} \{c(s, \gamma) + \beta \max_{\sigma \in \Sigma_{\gamma}(s)} V(s\sigma)\}, \quad s \in L^{0}(G). \quad (8.9)$$

Proof: 1. It is easy to see that the result in Lemma 8.1 is also true for $L^-(G)$ because when $V^*(s) = -\infty$, $V(\boldsymbol{\pi}, s) = -\infty$ for each supervisor $\boldsymbol{\pi} \in \Pi$. The remaining results are obvious.

2. For $s \in L^0(G)$, all $V^*(s)$ and $c(s, \gamma)$ for $\gamma \in \Gamma_1(s)$ are finite. Then, from Lemma 8.2 we have that

$$\begin{aligned} V^*(s) &= \sup_{\boldsymbol{\pi} \in \Pi_1} V(\boldsymbol{\pi}, s) \\ &= \sup_{\boldsymbol{\pi} \in \Pi_1} \{c(s, \boldsymbol{\pi}) + \beta \max_{\sigma \in \Sigma_{\boldsymbol{\pi}}(s)} V(\boldsymbol{\pi}, s\sigma)\} \\ &= \sup_{\gamma \in \Gamma_1(s), \boldsymbol{\pi} \in \Pi_1} \{c(s, \gamma) + \beta \max_{\sigma \in \Sigma_\gamma(s)} V(\boldsymbol{\pi}, s\sigma)\} \\ &= \max_{\gamma \in \Gamma_1(s)} \{c(s, \gamma) + \beta \sup_{\boldsymbol{\pi} \in \Pi_1} \max_{\sigma \in \Sigma_\gamma(s)} V(\boldsymbol{\pi}, s\sigma)\} \\ &= \max_{\gamma \in \Gamma_1(s)} \{c(s, \gamma) + \beta \max_{\sigma \in \Sigma_\gamma(s)} V^*(s\sigma)\}. \end{aligned}$$

Hence, the theorem is true. □

In the following theorem, we characterize solutions of the optimality equation (8.9). Due to the above theorem, we limit our attention to $L^0(G)$. For any function V on $\Pi \times L(G)$, we denote by

$$\Pi_1^{nn}(V) := \{(\boldsymbol{\pi}, s) \mid \boldsymbol{\pi} \in \Pi_1,\ s \in L^0(G) \text{ satisfying } V(\boldsymbol{\pi}, s) \neq -\infty\}$$

a subset of $\Pi \times L(G)$ in which the value of V does not equal the negative infinity.

Lemma 8.3: *We have the following four statements.*

1. *V^* satisfies the following condition,*

$$\limsup_{n \to \infty} \beta^n \sup_{t \in L^\omega(\boldsymbol{\pi}/G, s)} V(st_n) \geq 0,\ \ \forall (\boldsymbol{\pi}, s) \in \Pi_1^{nn}(V). \tag{8.10}$$

2. *$V \geq V^*$ if V is a solution of the optimality equation (8.9) and satisfies Eq. (8.10).*

3. *$V \leq V^*$ if V is a solution of the optimality equation (8.9) and satisfies*

$$\limsup_{n \to \infty} \beta^n \sup_{t \in L^\omega(\boldsymbol{\pi}/G, s)} V(st_n) \leq 0,\ \ \forall (\boldsymbol{\pi}, s) \in \Pi_1^{nn}(V). \tag{8.11}$$

4. *$V = V^*$ if V is a solution of the optimality equation (8.9) and satisfies*

$$\limsup_{n \to \infty} \beta^n \sup_{t \in L^\omega(\boldsymbol{\pi}/G, s)} V(st_n) = 0,\ \ \forall (\boldsymbol{\pi}, s) \in \Pi_1^{nn}(V). \tag{8.12}$$

Proof: 1. Because $V(\boldsymbol{\pi}, s) \neq -\infty$ and $s \in L^0(G)$, we know that $V(\boldsymbol{\pi}, s)$ is finite. So there is an infinite string $t^* \in L^\omega(\boldsymbol{\pi}/G, s)$ with its prefixes t_k^* for

$k = 0, 1, ...$ such that

$$V(\boldsymbol{\pi}, s) = \sup_{t \in L^\omega(\boldsymbol{\pi}/G,s)} \sum_{k=0}^{\infty} \beta^k c(st_k, \boldsymbol{\pi}) \geq \sum_{k=0}^{\infty} \beta^k c(st_k^*, \boldsymbol{\pi}) > -\infty.$$

Thus, the series $\sum_{k=0}^{\infty} \beta^k c(st_k^*, \boldsymbol{\pi})$ is convergent to some finite value. Moreover, from the definition of V^* we know that

$$\beta^n V^*(st_n) \geq \beta^n V(\boldsymbol{\pi}, st_n) = \beta^n \sup_{u \in L^\omega(\boldsymbol{\pi}/G,st_n)} \sum_{k=0}^{\infty} \beta^k c(st_n u_k, \boldsymbol{\pi}).$$

Hence,

$$\begin{aligned}
\sup_{t \in L^\omega(\boldsymbol{\pi}/G,s)} \beta^n V^*(st_n) &\geq \sup_{t \in L^\omega(\boldsymbol{\pi}/G,s)} \beta^n \sup_{u \in L^\omega(\boldsymbol{\pi}/G,st_n)} \sum_{k=0}^{\infty} \beta^k c(st_n u_k, \boldsymbol{\pi}) \\
&= \sup_{t \in L^\omega(\boldsymbol{\pi}/G,s)} \sum_{k=n}^{\infty} \beta^k c(st_k, \boldsymbol{\pi}) \\
&\geq \sum_{k=n}^{\infty} \beta^k c(st_k^*, \boldsymbol{\pi}) \\
&\to 0, \quad \text{as } n \to \infty.
\end{aligned}$$

Thus the result is true.

2. For each supervisor $\boldsymbol{\pi} \in \Pi_1$, finite string $s \in L(G)$, and infinite string $t \in L^\omega(G)$, we let

$$\begin{aligned}
v_s^n(\boldsymbol{\pi}, t) &= \sum_{k=0}^{n} \beta^k c(st_k, \boldsymbol{\pi}), \\
V^n(\boldsymbol{\pi}, s) &= \sup_{t \in L^\omega(\boldsymbol{\pi}/G,s)} v_s^n(\boldsymbol{\pi}, t).
\end{aligned}$$

Suppose that V satisfies the given conditions. Then for each $s \in L^0(G)$ and $\boldsymbol{\pi} \in \Pi_1$ with $V(\boldsymbol{\pi}, s) \neq \infty$, from the optimality equation (8.9) we have that

$$\begin{aligned}
V(s) &\geq c(s, \boldsymbol{\pi}) + \beta \max_{\sigma_1 \in \Sigma_{\boldsymbol{\pi}}(s)} V(s\sigma_1) \\
&\geq c(s, \boldsymbol{\pi}) + \beta \max_{\sigma_1 \in \Sigma_{\boldsymbol{\pi}}(s)} \{c(s\sigma_1, \boldsymbol{\pi}) + \beta \max_{\sigma_2 \in \Sigma_{\boldsymbol{\pi}}(s\sigma_1)} V(s\sigma_1\sigma_2)\} \\
&= V^1(\boldsymbol{\pi}, s) + \beta^2 \sup_{t \in L^\omega(\boldsymbol{\pi}/G,s)} V(st_2).
\end{aligned}$$

Based on the above formula, we can prove by the inductive method that

$$V(s) \geq V^{n-1}(\boldsymbol{\pi}, s) + \beta^n \sup_{t \in L^\omega(\boldsymbol{\pi}/G,s)} V(st_n), \quad n \geq 1.$$

By taking $\limsup_{n\to\infty}$ in the above inequality we obtain that $V(s) \geq V(\boldsymbol{\pi}, s)$. Due to the arbitrariness of $\boldsymbol{\pi}$ we get that $V \geq V^*$.

3. Due to the finiteness of Σ, there is supervisor $\boldsymbol{\pi}^* \in \Pi_1$ such that

$$V(s) \quad = \quad c(s, \boldsymbol{\pi}^*) + \beta \max_{\sigma_1 \in \Sigma_{\boldsymbol{\pi}^*}(s)} V(s\sigma_1).$$

Based on this, the result can be proved similarly to 2.

4. The result follows 2 and 3. □

From the above lemma, especially result 1, condition Eq. (8.11) is equivalent to condition Eq. (8.12) for $V = V^*$. So, the following theorem is true.

Theorem 8.2:

1. *V^* is the smallest solution of the optimality equation (8.9) satisfying condition Eq. (8.10).*

2. *V^* is the unique solution of the optimality equation (8.9) satisfying condition Eq. (8.11) or equivalently condition Eq. (8.12) if and only if the optimality equation (8.9) has a solution satisfying condition Eq. (8.11) or equivalently condition Eq. (8.12).*

In the above theorem, the optimal value is characterized as a solution of the optimality equation.

A sufficient condition for Eq. (8.11) is the following

$$\limsup_{n\to\infty} \beta^n \sup_{t \in L^\omega(G,s)} V(st_n) \leq 0, \ \ \forall s \in L^0(G).$$

This condition is simpler and may be verified more easily than Eq. (8.11).

The following two theorems relate the optimality of supervisors to the optimality equation.

Theorem 8.3: *For each supervisor $\boldsymbol{\pi}^*$, if for all $s \in L^0(G)$ with $V(\boldsymbol{\pi}^*, s) \neq -\infty$ we have*

$$\limsup_{n\to\infty} \beta^n \sup_{t \in L^\omega(\boldsymbol{\pi}^*/G,s)} V^*(st_n) = 0, \tag{8.13}$$

then $\boldsymbol{\pi}^*$ is optimal if and only if $\boldsymbol{\pi}^*$ attains the maximum of the optimality equation (8.9).

Proof: Sufficiency. Similarly to proof 2 in Lemma 8.3, we have

$$\begin{aligned} V^*(s) \quad &= \quad c(s, \boldsymbol{\pi}^*) + \beta \max_{\sigma \in \Sigma_{\boldsymbol{\pi}^*}(s)} V^*(s\sigma) \\ &= \quad V^{n-1}(\boldsymbol{\pi}^*, s) + \beta^n \sup_{t \in L^\omega(\boldsymbol{\pi}^*/G,s)} V^*(st_n), \end{aligned} \tag{8.14}$$

for $s \in L^0(G), n \geq 1$. By letting $n \to \infty$ in the above equation, we get $V(\boldsymbol{\pi}^*, s) = V^*(s)$ for all $s \in L^0(G)$.

Necessity. If $\boldsymbol{\pi}^*$ is optimal, then $V(\boldsymbol{\pi}^*, s) = V^*(s)$ for all $s \in L^0(G)$. With this and Lemma 8.2 we have

$$\begin{aligned} & \max_{\gamma \in \Gamma_1(s)} \{c(s,\gamma) + \beta \max_{\sigma \in \Sigma_\gamma(s)} V^*(s\sigma)\} \\ = \; & V^*(s) = V(\boldsymbol{\pi}^*, s) \\ = \; & c(s, \boldsymbol{\pi}^*) + \beta \max_{\sigma \in \Sigma_{\boldsymbol{\pi}^*}(s)} V^*(s\sigma). \end{aligned}$$

This implies that $\boldsymbol{\pi}^*$ attains the maximum of the optimality equation (8.9). Hence, the theorem is true. □

The above theorem characterizes the optimal supervisors with the optimality equation (8.9), and the following theorem characterizes the structure of the set of optimal supervisors. We let the set of optimal control inputs at string $s \in L^0(G)$ be

$$\Gamma_1^*(s) = \{\gamma \in \Gamma_1(s) \mid \gamma \text{ attains the maximum in Eq. (8.9)}\}.$$

These sets play an important role in optimal supervisors.

Theorem 8.4: *A supervisor $\boldsymbol{\pi}^*$ is optimal in $L^0(G)$ if and only if (1) for $s \in L^0(G)$, $\Gamma_1^*(s)$ is nonempty and $\boldsymbol{\pi}^*(s) \in \Gamma_1^*(s)$, and (2) $(\boldsymbol{\pi}^*, V^*)$ satisfies Eq. (8.13) or for all $s \in L^0(G)$ with $V(\boldsymbol{\pi}^*, s) \neq -\infty$,*

$$\limsup_{n \to \infty} \beta^n \sup_{t \in L^\omega(\boldsymbol{\pi}^*/G, s)} V^*(st_n) \leq 0. \tag{8.15}$$

Proof: Necessity. Suppose that supervisor $\boldsymbol{\pi}^*$ is optimal in $L^0(G)$. Then from Lemma 8.2,

$$\begin{aligned} V^*(s) &= V(\boldsymbol{\pi}^*, s) \\ &= c(s, \boldsymbol{\pi}^*(s)) + \beta \max_{\sigma \in \Sigma_{\boldsymbol{\pi}^*}(s)} V(\boldsymbol{\pi}^*, s\sigma) \\ &= c(s, \boldsymbol{\pi}^*(s)) + \beta \max_{\sigma \in \Sigma_{\boldsymbol{\pi}^*}(s)} V^*(s\sigma), \;\; s \in L^0(G). \end{aligned}$$

Hence, $\boldsymbol{\pi}^*(s) \in \Gamma_1^*(s)$ and so $\Gamma_1^*(s)$ is nonempty for each $s \in L^0(G)$. Moreover, we have Eq. (8.13). With this, we know that 2 is also true.

Sufficiency. Suppose that supervisor $\boldsymbol{\pi}^*$ satisfies the given conditions 1 and 2. Then we have from the optimality equation (8.9) that

$$V^*(s) = c(s, \boldsymbol{\pi}^*(s)) + \beta \max_{\sigma \in \Sigma_{\boldsymbol{\pi}^*}(s)} V^*(s\sigma), \;\; s \in L^0(G).$$

Thus, we can also get Eq. (8.13). By letting $\limsup_{n\to\infty}$ in the above equation, we can get from the given conditions that $V^*(s) = V(\boldsymbol{\pi}^*, s)$ for $s \in L^0(G)$. □

At the end of this section, we give the following corollary for two special cases. The proof is simple and is omitted here.

Corollary 8.1: *We have the following three statements.*

1. *Suppose that for each $s \in L^0(G)$, $c(s,\gamma)$ is increasing in $\gamma \in \Gamma_1(s)$ and there is the unique maximal element in $\Gamma_1(s)$, denoted by $\Gamma_M(s)$. Let a supervisor $\boldsymbol{\pi}_M$ by $\boldsymbol{\pi}_M(s) = \Gamma_M(s)$ for $s \in L^0(G)$. Then $\boldsymbol{\pi}_M$ attains the maximum of the optimality equation (8.9).*

2. *Suppose that for each $s \in L^0(G)$, $c(s,\gamma)$ is decreasing in $\gamma \in \Gamma_1(s)$ and there is the unique minimal element in $\Gamma_1(s)$, denoted by $\Gamma_m(s)$. Let a supervisor $\boldsymbol{\pi}_m$ by $\boldsymbol{\pi}_m(s) = \Gamma_m(s)$. Then $\boldsymbol{\pi}_m$ attains the maximum of the optimality equation (8.9).*

3. *If $c(s,\gamma)$ is bounded uniformly in $\gamma \in \Gamma_1(s)$ and $s \in L^0(G)$ and $\beta \in (0,1)$, then $\boldsymbol{\pi}^*$ is optimal if and only if $\boldsymbol{\pi}^*$ attains the maximum of the optimality equation (8.9).*

When $\Sigma \in \Gamma_1(s)$, the unique maximal element in $\Gamma_1(s)$ exists and is just Σ, whereas when $\Sigma_u \in \Gamma_1(s)$, the unique minimal element in $\Gamma_1(s)$ exists and is just Σ_u.

Remark 8.2: *In our CDES G_c, $\Gamma(s) \subset \Gamma$. This can be fitted to the case of $\Gamma(s) = \Gamma$ by defining the reward function $c(s,\gamma) = -\infty$ for $\gamma \in \Gamma - \Gamma(s)$. On the other hand, for the optimal control problem Eq. (8.5), even if $\Gamma(s) = \Gamma$, we have still $\Gamma_1(s) \subset \Gamma$. This makes the sets of available control inputs dependent on strings.* □

In this section, we study the optimal control problem by the ideas and methods in Chapter 2 for MDPs. We divide the language $L(G)$ into three sublanguages: $L^-(G)$, $L^+(G)$, and $L^0(G)$, and show and characterize the optimality equation and optimal supervisors in $L^0(G)$.

3. Language Properties

In this section, we study some properties of the sublanguages. For this we use some ideas from languages, automatons, and the supervisory control of DESs.

We first introduce several concepts.

Suppose that $\Sigma'(s) \subset \Sigma(s)$ is an event subset for each $s \in L(G)$. Then

$$G' := \{\Sigma'(s), s \in L(G)\}$$

is also a discrete event system and surely $L(G') \subset L(G)$. We call G' a subsystem of G. For two languages L_1 and L_2, if there are strings $s \in L_1$ and $r \in L(G', s)$ such that $sr \in L_2$ then we say that L_2 can be reached from L_1 through G'.

Definition 8.1: *For a language $L \subset L(G)$ and a subsystem G' of G, L is said to be*

(a) a G'-invariant language if for arbitrary strings $s \in L$ and $r \in L(G', s)$ with $sr \in L(G)$, and sr must belong to L itself;

(b) a G'-closed language if for arbitrary strings $s \in L(G)$ and $r \in L(G', s)$ with $sr \in L$, and s must belong to L itself. □

It is clear that the closed language and the controllable language defined in the literature of supervisory control (Ramadge and Wonham 1989) are exactly the G-closed language and G_u-invariant language, respectively. A language L is G'-closed means that the past of L in G' is included in L itself. L is G'-invariant means that L includes its future under G'; that is, the system will remain in L through G' whenever the system begins in L.

Now, we return to properties of the sublanguages of $L(G)$. We define a subsystem

$$G_1 := \{\Sigma_1(s), s \in \Sigma^*\}$$

by

$$\Sigma_1(s) = \bigcup\{\gamma \mid \gamma \in \Gamma_1(s)\}, \quad s \in L(G) - L^{-\infty}(G)$$

and $\Sigma_1(s) = \emptyset$ for $s \in L^{-\infty}(G)$. The following theorem gives some properties.

Theorem 8.5: *$L^-(G)$ is G_1-invariant and $L^+(G)$ is G_1-closed.*

Proof: 1. We show first that $L^-(G)$ is G_1-invariant. For any string $s \in L^-(G)$ and event $\sigma \in \Sigma_1(s)$ with $s\sigma \in L(G)$, if $s\sigma \notin L^-(G)$; that is, $V^*(s\sigma) > -\infty$, then there is a supervisor $\boldsymbol{\pi}$ such that $V(\boldsymbol{\pi}, s\sigma) > -\infty$. By taking any $\gamma \in \Gamma_1(s)$ with $\sigma \in \gamma$, we define a supervisor $\boldsymbol{\pi}'$ by $\boldsymbol{\pi}'(s) = \gamma$ and $\boldsymbol{\pi}'(t) = \boldsymbol{\pi}(t)$ for all other strings $t \neq s$. This supervisor $\boldsymbol{\pi}'$ differs from $\boldsymbol{\pi}$ only at string s. Then, from Lemma 8.2,

$$\begin{aligned} V(\boldsymbol{\pi}', s) &= c(s, \boldsymbol{\pi}') + \beta \max_{\sigma' \in \Sigma_\gamma(s)} V(\boldsymbol{\pi}', s\sigma') \\ &= c(s, \gamma) + \beta \max_{\sigma' \in \Sigma_\gamma(s)} V(\boldsymbol{\pi}, s\sigma') \\ &\geq c(s, \gamma) + \beta V(\boldsymbol{\pi}, s\sigma) \\ &> -\infty \end{aligned}$$

due to the finiteness of $c(s, \gamma)$. Hence, $V^*(s) \geq V(\boldsymbol{\pi}', s) > -\infty$, which contradicts the fact that $s \in L^-(G)$. Therefore, $s\sigma \in L^-(G)$. This shows that $L^-(G)$ is G_1-invariant.

2. In order to show that $L^+(G)$ is G_1-closed, it suffices to show that if $s\sigma \in L^+(G)$ with $\sigma \in \Sigma_1(s)$ then $s \in L^+(G)$. For the string s and the event σ, there must be a control input $\gamma \in \Gamma_1(s)$ such that $\sigma \in \gamma$. Then for each supervisor $\boldsymbol{\pi} \in \Pi_1$, we define another supervisor $\boldsymbol{\pi}'$ by letting $\boldsymbol{\pi}'(s) = \gamma$ and $\boldsymbol{\pi}'(t) = \boldsymbol{\pi}(t)$ for all other strings $t \neq s$. Then from Lemma 8.2 again we know that

$$\begin{aligned} V^*(s) &\geq V(\boldsymbol{\pi}', s) \\ &= c(s,\gamma) + \beta \max_{\sigma' \in \Sigma_\gamma(s)} V(\boldsymbol{\pi}, s\sigma') \\ &\geq c(s,\gamma) + \beta V(\boldsymbol{\pi}, s\sigma). \end{aligned}$$

Due to the arbitrariness of $\boldsymbol{\pi}$ and $s \in L^+(G)$, we have that $V^*(s) \geq c(s,\gamma) + \beta V^*(s\sigma) = +\infty$. Hence, $s \in L^+(G)$. □

With the above theorem, we have the following reachable relationships among the three sublanguages: $L^+(G)$ can reach both $L^0(G)$ and $L^-(G)$ through G_1 and $L^0(G)$ can reach $L^-(G)$ through G_1. However, the reverse reachable relationships among these three sub-languages do not hold.

4. System Based on Automaton

When the discrete event system is modeled by an automaton, the problem arises of whether any better results can be attained.

For the given optimal control problem Eq. (8.5), suppose that G can be described by an automaton $G = \{\Sigma, Q, \delta, q_0\}$ and furthermore for $\gamma \in \Gamma(s)$ and $s \in L(G)$,

$$\Sigma(s) = \Sigma(\delta(s,q_0)),\ \Gamma(s) = \Gamma(\delta(s,q_0)),\ c(s,\gamma) = c(\delta(s,q_0),\gamma), \quad (8.16)$$

where for $q \in Q$, $\Sigma(q) = \{\sigma \in \Sigma \mid \delta(\sigma,q)!\}$, $\Gamma(q) \subset \Gamma$, and $c(q,\gamma)$ is an extended real-valued reward function defined on $\{(q,\gamma) \mid \gamma \in \Gamma(q), q \in Q\}$. When condition Eq. (8.16) is true, we say that the controlled discrete event system G_c and the optimal control problem Eq. (8.5) are stationary.

We define a state feedback f as a map: $Q \to \Gamma$ satisfying $f(q) \in \Gamma(q)$. Obviously, a state feedback is also a supervisor $\boldsymbol{\pi} : L(G) \to \Gamma$ with $\boldsymbol{\pi}(s) = f(\delta(s,q_0))$. Let F be the set of all state feedback.

For any given $q \in Q$ and infinite string $t = \sigma_0\sigma_1\cdots$, we define $q_{k+1} = \delta(\sigma_k, q_k)$ for $k \geq 0$ with $q_0 = q$, and

$$v_q(f,t) = \sum_{k=0}^{\infty} \beta^k c(q_k, f), \quad f \in F$$

as the discounted total reward for string t from state q under state feedback f. Moreover, we define

$$V(f,q) = \sup_{t \in L^{\omega}(f/G,q)} v_q(f,t)$$

as the maximal discounted total reward from state q under state feedback f. The optimal value function within state feedbacks is defined by

$$V^0(q) = \sup_{f \in F} V(f,q), \quad q \in Q.$$

As in Section 8.2, we divide the state set Q into the following three subsets:

$$\begin{aligned} Q^- &= \{q \in Q \mid V^0(q) = -\infty\}, \\ Q^+ &= \{q \in Q \mid V^0(q) = +\infty\}, \\ Q^0 &= \{q \in Q \mid V^0(q) \in (-\infty, +\infty)\}. \end{aligned}$$

Moreover, we define

$$Q^{+\infty} = \{q \in Q \mid \text{there is a state feedback } f \text{ such that } V(f,q) = +\infty\}.$$

Let

$$\Gamma_1(q) = \{\gamma \in \Gamma(q) \mid c(q,\gamma) > -\infty\}, \quad q \in Q.$$

Certainly, $\Gamma_1(s) = \Gamma_1(\delta(s,q))$ and so

$$\Sigma_1(s) = \Sigma_1(\delta(s,q)) := \bigcup \{\gamma \mid \gamma \in \Gamma_1(\delta(s,q))\}.$$

Therefore, $G_1 = \{\Sigma_1(s), s \in \Sigma^*\}$ can be described by an automaton

$$G_1 = \{Q, \Sigma, \delta_1, q_0\},$$

where $\delta_1(\sigma, q) = \delta(\sigma, q)$ for $\sigma \in \Sigma_1(q)$ and is undefined otherwise.

If we restrict the problem under the condition Eq. (8.16) within the state feedback set F and the state set Q, then all the results obtained in Sections 8.3 and 8.4 are still true. For example, the following lemma is similar to Theorem 8.1 and Theorem 8.5.

Lemma 8.4: *We have the following statements.*

1. *Each state feedback $f \in F$ is optimal in Q^-. There is an optimal state feedback in $Q^{+\infty}$ but there is no optimal state feedback in $Q^+ - Q^{+\infty}$.*

2. *V^0 satisfies the following equation in the subset Q^0.*

$$V(q) = \max_{\gamma \in \Gamma_1(q)} \{c(q,\gamma) + \beta \max_{\sigma \in \Sigma_\gamma(q)} V(\delta(\sigma,q))\}, \quad q \in Q^0. \tag{8.17}$$

This equation is called the stationary optimality equation.

3. *Q^- is a G_1-invariant predicate and Q^+ is a G_1-closed predicate.*

Here, G_1-invariant predicates and G_1-closed predicates can be similarly defined to G_1-invariant languages and G_1-closed languages.

The following theorem discusses some relationships between $V^*(s)$ and $V^0(\delta(s, q_0))$ for $s \in L(G)$.

Theorem 8.6: *$V^*(s) \geq V^0(\delta(s, q_0))$ for all $s \in L(G)$. Moreover, if*

$$\limsup_{n\to\infty} \beta^n \sup_{t\in L^\omega(\boldsymbol{\pi}/G,s)} V^0(\delta(st, q_0)) \geq 0, \tag{8.18}$$

for all $\pi \in \Pi_1$ and all $s \in L(G)$ with $\delta(s, q_0) \in Q^0$, then $V^*(s) = V^0(\delta(s, q_0))$ for all $s \in L(G)$ with $\delta(s, q_0) \in Q^0$.

Proof: First, it is apparent that for $s \in L(G), st \in L^\omega(G)$, and $f \in F$,

$$v_s(f, t) = v_{\delta(s,q_0)}(f, t), \;\; V(f, s) = V(f, \delta(s, q_0)).$$

Then for each $s \in L(G)$,

$$\begin{aligned} V^*(s) &= \sup_{\boldsymbol{\pi}} V(\boldsymbol{\pi}, s) \\ &\geq \sup_{f} V(f, s) = \sup_{f} V(f, \delta(s, q_0)) \\ &= V^0(\delta(s, q_0)). \end{aligned}$$

On the other hand, for each supervisor $\boldsymbol{\pi}$, we have from Eq. (8.17) that

$$V^0(\delta(s, q_0)) \geq c(s, \boldsymbol{\pi}) + \beta \max_{\sigma\in\Sigma_{\boldsymbol{\pi}}(s)} V^0(\delta(s, q_0)),$$

for all $s \in L(G)$ with $\delta(s, q_0) \in Q^0$. With this, we can prove as in Lemma 8.3 that for all $s \in L(G)$ with $\delta(s, q_0) \in Q^0$,

$$V^0(\delta(s, q_0)) \geq V^{n-1}(\boldsymbol{\pi}, s) + \beta^n \sup_{t\in L^\omega(\boldsymbol{\pi}/G,s)} V^0(\delta(st, q_0)).$$

By letting $n \to \infty$ in the above inequality and according to the conditions in Eq. (8.18), we get that $V^0(\delta(s, q_0)) \geq V(\boldsymbol{\pi}, s)$. Because $\boldsymbol{\pi}$ is arbitrary, $V^0(\delta(s, q_0)) \geq V^*(s)$. Hence, $V^0(\delta(s, q_0)) = V^*(s)$ for all $s \in L(G)$ with $\delta(s, q_0) \in Q^0$. □

From the above theorem, the next corollary immediately follows.

Corollary 8.2: *We have*

$$\{\delta(s, q_0) \mid s \in L^-(G)\} \subset Q^-, \;\; Q^+ \subset \{\delta(s, q_0) \mid s \in L^+(G)\}.$$

Moreover, when V^0 satisfies Eq. (8.18),

$$\{\delta(s, q_0) \mid s \in L^0(G)\} = Q^0.$$

From Theorem 8.6 and Corollary 8.2, we know that under the condition given by Eq. (8.18) the optimality in state feedback set F is equivalent to that in the whole supervisor set Π.

When V^0 satisfies Eq. (8.18), there may be

$$\{\delta(s, q_0) \mid s \in L^-(G)\} = Q^-, \;\; Q^+ = \{\delta(s, q_0) \mid s \in L^+(G)\}.$$

But we could not prove it.

At the end of this section, we make the following remark.

Remark 8.3: *1. In the stationary conditions given in Eq. (8.16), the first two conditions on $\Sigma(s)$ and $\Gamma(s)$ are conditions for the controlled DES G_c Eq. (8.3). Under these two conditions, we say that G_c is an automaton type. But without the third condition on the reward function in Eq. (8.16), we cannot obtain the above results. From the above results, we say that when the problem is stationary, the optimality equation and the optimal supervisors are also stationary.*

2. For the optimal control problem of an automaton-type controlled DES, there is an interesting problem: when is there an automaton-type supervisor? Here, we call $\boldsymbol{\pi}$ an automaton-type supervisor if there is an automaton $P = \{Y, \Sigma, \eta, y_0\}$ and a map $\phi : Y \to \Gamma$ such that

$$\boldsymbol{\pi}(s) = \phi(\eta(s, y_0)), \quad s \in L(G).$$

5. Supervisory Control Problems

The most basic problem in the supervisory control of DESs is the supervisory control problem. In this section, we study the problem in the framework of both event feedback control and state feedback control.

5.1 Event Feedback Control

For a given language $K \subset L(G)$, we consider the problem of whether there is a supervisor $\boldsymbol{\pi}$ such that $L(\boldsymbol{\pi}/G) = K$. If so, we say that K can be synthesized by $\boldsymbol{\pi}$, or $\boldsymbol{\pi}$ synthesizes K. In the standard model (i.e., $\Gamma(s) = \Gamma$ for all $s \in L(G)$), a necessary and sufficient condition for synthesizing a language K is that K is G-closed and G_u-invariant (Ramadge and Wonham [104]).

In this section, we discuss the problem for CDES G_c Eq. (8.5). Because the set $\Gamma(s)$ does not equal Γ we need some other conditions for synthesizing a

language. We denote by

$$\mathbf{\Gamma} = \{\Gamma(s), s \in L(G)\}$$

the collection of sets of available control inputs.

Definition 8.2: *For $K \subset \Sigma^*$, K is said to be*

(a) *A $\mathbf{\Gamma}$-invariant language if for each string $s \in K$ there is $\gamma_s \in \Gamma(s)$ satisfying*

$$s\gamma_s \cap L(G) \subset K. \tag{8.19}$$

(b) *A $\mathbf{\Gamma}$-closed language if for each $s\sigma \in K$ there is $\gamma_s \in \Gamma(s)$ satisfying $\sigma \in \gamma_s$ and Eq. (8.19).*

(c) *A $\mathbf{\Gamma}$-controllable language if it is $\mathbf{\Gamma}$-invariant and $\mathbf{\Gamma}$-closed.*

The following remark concerns the concepts defined above.

Remark 8.4: *When $\Sigma_u \in \Gamma(s)$ for all $s \in K$, Σ_u is the minimal element in all $\Gamma(s)$. In this case, Eq. (8.19) is equivalent to $s\Sigma_u \cap L(G) \subset K$ for all $s \in K$, which is exactly the definition of controllable language for the standard model (Ramadge and Wonham [104]). In fact, a $\mathbf{\Gamma}$-invariant language must be controllable but the reverse is not true in general. Hence, the concept of $\mathbf{\Gamma}$-invariant languages generalizes that of controllable languages. Similarly, the concept of $\mathbf{\Gamma}$-closed languages generalizes that of closed languages: a $\mathbf{\Gamma}$-closed language is closed but the reverse is not true in general.* □

We have the following result on synthesizing a language.

Theorem 8.7: *For any given language $K \subset L(G)$, there is a supervisor $\boldsymbol{\pi}_K$ such that $L(\boldsymbol{\pi}_K/G) = K$ if and only if K is $\mathbf{\Gamma}$-controllable and the maximal element of*

$$\Gamma_K(s) = \{\gamma \mid s\gamma \cap L(G) \subset K, \gamma \in \Gamma(s)\}$$

exists uniquely for each $s \in K$. Moreover, the supervisor $\boldsymbol{\pi}_K$ can be taken by

$$\begin{aligned} \boldsymbol{\pi}_K(s) &= \max \Gamma_K(s) \\ &= \max\{\gamma \mid s\gamma \cap L(G) \subset K, \gamma \in \Gamma(s)\}, \quad s \in K. \end{aligned} \tag{8.20}$$

Proof: Necessity. For the given language $K \subset L(G)$, if there is a supervisor $\boldsymbol{\pi}_K$ such that $L(\boldsymbol{\pi}_K/G) = K$, then it is easy to see that

$$s\boldsymbol{\pi}_K(s) \cap L(G) = \{s\sigma \mid s\sigma \in K\}, \quad s \in K. \tag{8.21}$$

With this and Definition 8.2 we know that K is $\mathbf{\Gamma}$-controllable and the maximal element of $\Gamma_K(s)$ exists uniquely for each $s \in K$.

Sufficiency. Because K is $\mathbf{\Gamma}$-invariant, the set $\Gamma_K(s)$ is nonempty. Then,

$$\boldsymbol{\pi}_K(s) = \max \Gamma_K(s)$$

satisfies Eq. (8.19); that is, $s\boldsymbol{\pi}_K(s) \cap L(G) \subset K$ for all $s \in K$. On the other hand, for each $s\sigma \in K$, because K is $\mathbf{\Gamma}$-closed and the maximum of $\boldsymbol{\pi}_K(s)$, we have $\sigma \in \boldsymbol{\pi}_K(s)$. Hence, Eq. (8.21) is true and so $L(\boldsymbol{\pi}_K/G) = K$. □

The above theorem says that the supervisor $\boldsymbol{\pi}_K$ defined by Eq. (8.19) synthesizes the given language K.

Remark 8.5: *1. Surely, K is $\mathbf{\Gamma}$-controllable and $\max \Gamma_K(s)$ exists uniquely for each $s \in K$ if and only if Eq. (8.21) is true for some π_K. Equation (8.21) is exactly the condition 1 presented in Takai [134], where it is assumed that there is $\Gamma' \subset \Gamma$ such that $\Gamma(s) = \Gamma'$ for all strings s. Here, we show that Eq. (8.21) is also a necessary and sufficient condition for our synthesizing problem for a more general control pattern than that in [134]. On the other hand, we characterize condition 1 by $\mathbf{\Gamma}$-invariant and $\mathbf{\Gamma}$-closed languages. These two concepts generalize the corresponding concepts of invariant and closed languages, respectively, for the standard control pattern [104]. Our characterization is more essential than condition 1 in [134].*

2. When $\Gamma(s)$ is closed under union $\vee$, the maximal element of $\Gamma_K(s)$ exists uniquely for each $s \in K$. Then under this condition, Theorem 8.7 says just that there is a supervisor $\boldsymbol{\pi}_K$ such that $L(\boldsymbol{\pi}_K/G) = K$ if and only if K is $\mathbf{\Gamma}$-controllable. □

The above theorem solves the synthesizing problem if the given language K is $\mathbf{\Gamma}$-controllable. Otherwise, we want to know if there is a unique maximal sublanguage of K that is $\mathbf{\Gamma}$-controllable. In the following, we assume that $\Gamma(s)$ is closed under union $\vee$.

For any given language K, let K_1 and K_2 be two $\mathbf{\Gamma}$-controllable sublanguages of K. It is easy to see from the definitions that $K_1 \cup K_2$ is also a $\mathbf{\Gamma}$-controllable sublanguage of K. Hence, the set of $\mathbf{\Gamma}$-controllable sublanguages of K is closed under union and so has the unique maximal element. We denote this maximal element by

$$K^{\uparrow} = \max\{K' \mid K' \subset K \text{ is } \mathbf{\Gamma}\text{-controllable}\},$$

and call it the maximal $\mathbf{\Gamma}$-controllable sublanguages of K. This shows the following lemma.

Lemma 8.5: *For any given language $K \subset L(G)$, its maximal $\mathbf{\Gamma}$-controllable sublanguage $K^{\uparrow}$ exists uniquely and the supervisor synthesizing $K^{\uparrow}$ is $\boldsymbol{\pi}_{K^{\uparrow}}$.*

Similarly, for any given language $K \subset L(G)$, its maximal $\mathbf{\Gamma}$-invariant sublanguage and maximal $\mathbf{\Gamma}$-closed sublanguage exist uniquely.

In the following, we introduce an optimal control problem to compute $K^{\uparrow}$. In fact, the supervisor $\boldsymbol{\pi}_{K^{\uparrow}}$ is constructed. But before doing this, we introduce some concepts.

We define $\gamma_1 \leq \gamma_2$ by the set inclusion $\gamma_1 \subset \gamma_2$ for two control inputs γ_1 and γ_2. Then $\leq$ is a partial order in Γ and also in $\Gamma(s)$ for each s. Moreover, we define $\boldsymbol{\pi}_1 \leq \boldsymbol{\pi}_2$ for two supervisors $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ if $\boldsymbol{\pi}_1(s) \leq \boldsymbol{\pi}_2(s)$ for all $s \in L(G)$. For a supervisor set Π', we call $\boldsymbol{\pi}^* \in \Pi'$ the maximum supervisor of Π' if $\boldsymbol{\pi} \leq \boldsymbol{\pi}^*$ for all $\boldsymbol{\pi} \in \Pi'$. Especially, for an optimal control problem, we refer to the maximum supervisor of the set of all its optimal supervisors as the maximum optimal supervisor, if it exists.

For the given language K, we define an optimal control problem based on the CDES G_c with the reward function

$$c(s,\gamma) = \begin{cases} 0, & \text{if } s \in K, s\gamma \cap L(G) \subset K \\ -1, & \text{else.} \end{cases}$$

This reward function is uniformly bounded. We take the discount factor by any $\beta \in (0,1)$. Then, $V(\boldsymbol{\pi}, s)$ is well defined and uniformly bounded. Therefore, $L^{+}(G) = L^{-}(G) = \emptyset$, $L^{0}(G) = L(G)$, and $\Gamma_1(s) = \Gamma(s)$. Let

$$K^* = \{s \in K \mid V^*(s) = 0\}.$$

We have the following result about K^*.

Theorem 8.8: *K^* is the maximal $\mathbf{\Gamma}$-invariant sublanguage of K and $\boldsymbol{\pi}_{K^*}$ is the maximum optimal supervisor of the optimal control problem. Moreover,*

$$K^{\uparrow} = L(\boldsymbol{\pi}_{K^*}/G) = \{s \in K^* \mid t \in K^*, \forall t \leq s\}$$

is the maximal $\mathbf{\Gamma}$-closed sublanguage of K^*.

Proof: 1. For each $s \in K^*$, $V^*(s) = 0$ and so there is $\gamma \in \Gamma(s)$ such that $c(s,\gamma) = 0$. This results in γ satisfies Eq. (8.19). Thus, K^* is a $\mathbf{\Gamma}$-invariant sublanguage of K.

Now, if $K' \subset K$ is $\mathbf{\Gamma}$-invariant, then from the optimality equation (8.9) and the definition of $\mathbf{\Gamma}$-invariant language, we know that $V^*(s) = 0$ for $s \in K'$. Thus, $K' \subset K^*$.

Hence, K^* is the maximal $\mathbf{\Gamma}$-invariant sublanguage of K.

2. It is apparent that $\Gamma_{K^*}(s)$ is exactly the set of control inputs that attains the maximum of the optimality equation (8.9). It is nonempty and closed under

union. Hence, $\boldsymbol{\pi}_{K^*}(s) = \max \Gamma_{K^*}(s)$ is well defined. Then, due to the proof of Theorem 8.7, $\boldsymbol{\pi}_{K^*}$ is the maximal optimal supervisor of the optimal control problem.

3. We then prove that $L(\boldsymbol{\pi}_{K^*}/G) = K^{\uparrow}$. First, it is obvious that $L(\boldsymbol{\pi}_{K^*}/G) \subset K^* \subset K$ and $L(\boldsymbol{\pi}_{K^*}/G)$ is $\boldsymbol{\Gamma}$-controllable. Second, if $K' \subset K$ is $\boldsymbol{\Gamma}$-controllable, then it is obvious that $V(\boldsymbol{\pi}_{K'}, s) = 0$ for $s \in K'$. Thus, $V^*(s) = 0$ for $s \in K'$. This shows that $K' \subset K^*$. Therefore, $\boldsymbol{\pi}_{K'}(s) \leq \boldsymbol{\pi}_{K^*}(s)$ for all $s \in K'$. This implies that $K' = L(\boldsymbol{\pi}_{K'}/G) \subset L(\boldsymbol{\pi}_{K^*}/G)$.

Overall, $K^{\uparrow} = L(\boldsymbol{\pi}_{K^*}/G)$. Obviously,

$$L(\boldsymbol{\pi}_{K^*}/G) = \{s \in K^* \mid t \in K^*, \forall t \leq s\}$$

is the maximal $\boldsymbol{\Gamma}$-closed sublanguage of K^*. □

With the above theorem, we know that in order to compute the maximal $\boldsymbol{\Gamma}$-controllable language $K^{\uparrow}$ of K, we can first compute the maximal $\boldsymbol{\Gamma}$-invariant language K^* of K and then compute the maximal $\boldsymbol{\Gamma}$-closed language of K^* which is exactly $K^{\uparrow}$.

5.2 State Feedback Control

First, for a state feedback f, we let $f/G = (Q, \Sigma, \delta_f, q_0)$ be a subsystem of G with $\delta_f(\sigma, q) = \delta(\sigma, q)$ if $\sigma \in f(q)$ and otherwise undefined, and

$$R(f/G) = \{\delta(s, q_0) \mid s \in L(f/G)\}$$

be the reachable state set of the system f/G (Ramadge and Wonham [105]). The supervisory control problem in state feedback control is whether there is a state feedback f such that $R(f/G) = P$ for a given predicate $P \subset Q$. If so, we say that P can be synthesized by f, or f synthesizes P. In the standard model with $\Gamma(q) = \Gamma$ for all $q \in Q$, a necessary and sufficient condition for synthesizing a predicate P is that P is controllable [105].

Similarly to that in the previous subsection, we denote by

$$\boldsymbol{\Gamma} = \{\Gamma(q), q \in Q\}$$

the collection of sets of available control inputs. Moreover, for $P \subset Q$, P is said to be

(a) a $\boldsymbol{\Gamma}$-invariant predicate if for each state $q \in P$ there is $\gamma_q \in \Gamma(q)$ such that $\delta(\sigma, q) \in P$ for all $\sigma \in \Sigma(q) \wedge \gamma_q$, (in this case we let $\Gamma_P(q)$ be the set of all such γ_q).

(b) a $\boldsymbol{\Gamma}$-closed predicate if for each state $q \in P$ there is integer $n \geq 0$ and states $q_k \in P$ and control inputs $\gamma_k \in \Gamma_P(q_k)$ for $k = 0, 1, \ldots, n-1$ such that for $k = 0, 1, \ldots, n-1$ there is $\sigma_k \in \gamma_k$ with $q_{k+1} = \delta(\sigma_k, q_k)$ with $q_n = p$.

(c) a $\mathbf{\Gamma}$-controllable predicate if P is $\mathbf{\Gamma}$-invariant and $\mathbf{\Gamma}$-closed.

The above concepts correspond to those defined in Definition 8.2 for languages. Also, when $\Gamma(q) = \Gamma$ for all $q \in Q$, $\mathbf{\Gamma}$-controllable predicates are exactly the controllable predicates for the standard model [105].

The following theorem on synthesizing a given predicate P can be proved similarly to Theorem 8.7.

Theorem 8.9: *For any given predicate $P \subset Q$, there is a state feedback f_P such that $R(f_P/G) = P$ if and only if P is $\mathbf{\Gamma}$-controllable and $\max \Gamma_P(q)$ exists uniquely for each $q \in P$. Moreover, the state feedback f_P can be taken by*

$$\begin{aligned} f_P(q) &= \max \Gamma_P(q) \\ &= \max\{\gamma \mid \gamma \in \Gamma(q), \delta(\sigma, q) \in P, \forall \sigma \in \Sigma(q) \wedge \gamma\} \end{aligned} \tag{8.22}$$

for $q \in P$.

When the given predicate P is not $\mathbf{\Gamma}$-controllable, we assume that $\Gamma(q)$ is closed under union $\vee$ for each $q \in P$. Then the maximal $\mathbf{\Gamma}$-controllable subpredicate of P, denoted by $P^{\uparrow}$, is unique.

Similarly to the previous section, we introduce a stationary reward function by

$$c(q, \gamma) = \begin{cases} 0, & \text{if } q \in P, \delta(\sigma, q) \in P, \forall \sigma \in \Sigma(q) \wedge \gamma \\ -1, & \text{else.} \end{cases}$$

We still let $V^0(q)$ be the optimal value function and let

$$P^* = \{q \in P \mid V^0(q) = 0\}.$$

We have the following theorem about P^* and $P^{\uparrow}$, which can be proved similarly to Theorem 8.8.

Theorem 8.10: *P^* is the maximal $\mathbf{\Gamma}$-invariant subpredicate of P and f_{P^*} is the maximum optimal supervisor of the corresponding stationary optimal control problem. Moreover, $P^{\uparrow} = R(f_{P^*}/G)$ is the maximal $\mathbf{\Gamma}$-closed sub-predicate of P^*.*

The above theorem says that the supervisory control problem can be solved by the optimal control problem, although the former problem is in the logical level and the latter problem is in the performance level. Moreover, the method for solving the supervisory control problem by the optimal control problem is easy.

6. Job-Matching Problem

We consider a job shop with two machines, M_1 and M_2, and two job types, J_1 and J_2. A job of J_1 must be processed first in M_1 and then in M_2, and a job of

J_2 must be processed first in M_2 and then in M_1. A job is said to be completed if it is completed in both machines. Suppose that completed jobs are output to another system to be equipped and that the number of jobs of J_1 should equal the number of jobs of J_2 in each output (which is called job-matching).

The shop can be modeled by an automaton G as given in Figure 8.1.

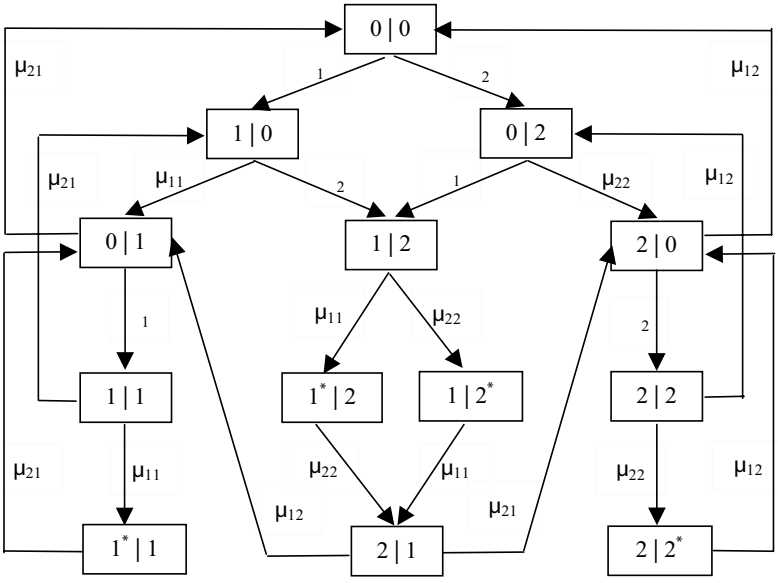

Figure 8.1. A job-matching problem: the automaton G.

The state variable is defined as (i, j) with $i \in \{0, 1, 2, 1^*\}$ and $j \in \{0, 1, 2, 2^*\}$. Here, $i = 0$ (or $j = 0$) represents that Machine M_1 (or M_2) is empty, $i \in \{1, 2\}$ (or $j \in \{1, 2\}$) represents that a job of J_i (or J_j) is being processed in M_1 (or M_2), whereas $i = 1^*$ (or $j = 2^*$) represents that a job of J_1 (or J_2) is completed and waiting in M_1 (or M_2). A job of J_i is waiting in M_i means that there is another job (of J_1 or J_2) being processed or waiting in M_j ($j \neq i$). Let Q be the set of all states.

The event set is $\Sigma = \{\lambda_i, \mu_{ij} \mid i, j = 1, 2\}$, where event λ_i represents the arrival of a job of J_i, event μ_{ij} represents completion of a job of J_i in machine j. Here it is assumed that only the arrival events can be controlled. Hence, the uncontrollable event set is $\Sigma_u = \{\mu_{ij} \mid i, j = 1, 2\}$, and the controllable event set is $\Sigma_c = \{\lambda_1, \lambda_2\}$.

Remark 8.6: *The above automaton G is based on the example discussed in Section 7.5. But there is no deadlocked state here. The main problem*

discussed in Section 7.5 is about deadlock, whereas it is the job-matching problem here. □

Suppose that there are N trays in the shop to take jobs. Each tray is common; that is, each tray can take a job either of J_1 or of J_2. When there is an empty tray, then a job can join the shop. The tray will take the job to the two machines which will then output the job. However, it is not allowed for all N trays to take jobs of one type.

For any $s \in L(G)$, if we let $|s|_{12}$ be the difference in the number of events λ_1 in s minus the number of events λ_2 in s, then $|s|_{12}$ represents the number of jobs of J_1 that joined the shop minus the number of J_2 that joined the shop. Due to the requirement of job-matching and that there are N trays,

$$|s|_{12} \in \{-N, -N+1, \ldots, 0, \ldots, N-1, N\}.$$

But if $|s|_{12} = -N$ or N then all the trays are taking jobs of the same type and so the system is deadlocked. Hence, we need some control mechanism. Here, we assume that the set of control inputs available at string $s \in L(G)$ is given by

$$\Gamma(s) = \begin{cases} \Gamma^-, & \text{if } |s|_{12} = -N+1 \\ \Gamma^+, & \text{if } |s|_{12} = N-1 \\ \Gamma, & \text{otherwise,} \end{cases}$$

where

$$\Gamma^- = \{\gamma \in \Gamma \mid \lambda_2 \notin \gamma\}, \;\; \Gamma^+ = \{\gamma \in \Gamma \mid \lambda_1 \notin \gamma\}.$$

Under any control input $\gamma \in \Gamma(s)$ at string s with $||s|_{12}| < N$, the system will never be deadlocked.

Moreover, the cost function is assumed as follows.

$$\begin{aligned} c(s,\gamma) &= c_1\chi(\lambda_1 \notin \gamma, \lambda_1 \in \Sigma(s)) + c_2\chi(\lambda_2 \notin \gamma, \lambda_2 \in \Sigma(s)) \\ &\quad + \infty\chi(\gamma \wedge \Sigma(s) = \emptyset), \;\; s \in L(G), \;\; \gamma \in \Gamma. \end{aligned}$$

Here χ is the indicator function, $c_i \geq 0$ is the cost for prohibiting event λ_i for $i = 1, 2$, and the last term $\infty\chi(\gamma \wedge \Sigma(s) = \emptyset)$ means that the deadlock is not allowed. Furthermore, for string s, if $\lambda_i \notin \Sigma(s)$ then event λ_i could not occur at string s and so there is no cost to prohibit λ_i at s. We take the discount factor by any $\beta \in (0, 1)$. Because $c(s, \gamma)$ is bounded, $\Gamma_1(s) = \Gamma(s)$ for all string s. It should be noted that if $\Gamma(s) = \Gamma$ then we will have the same result as in the following except that $\Gamma_1(s)$ is just the $\Gamma(s)$ defined previously due to the third term in $c(s, \gamma)$.

By noting that the cost function $c(s, \gamma)$ depends on string s only through state $q = \delta(s, q_0)$, we let

$$\begin{aligned} c(q,\gamma) &= c_1\chi(\lambda_1 \notin \gamma, \lambda_1 \in \Sigma(q)) + c_2\chi(\lambda_2 \notin \gamma, \lambda_2 \in \Sigma(q)) \\ &\quad + \infty\chi(\gamma \wedge \Sigma(q) = \emptyset), \;\; q \in Q, \;\; \gamma \in \Gamma. \end{aligned}$$

Then

$$c(s,\gamma) = c(\delta(s,q_0),\gamma), \quad s \in L(G), \quad \gamma \in \Gamma.$$

With this, and the fact that $\Gamma(s)$ depends on s only through $|s|_{12}$, it can be proved from Theorems 1 and 2 that $V^*(s)$, the minimal discounted total cost, depends on s only through $\delta(s,q_0)$ and $|s|_{12}$. It is easy to see that under each supervisor,

$$|s|_{12} \in \mathcal{N} := \{-N+1,\ldots,0,\ldots,N-1\}.$$

Then we have the function $V^*(q,n)$, defined on $Q \times \mathcal{N}$, such that

$$V^*(s) = V^*(\delta(s,q_0),|s|_{12}).$$

We introduce a variable $n = |s|_{12}$, which has the following transition law based on the changes of the system for any $n \in \mathcal{N}$.

$$\eta(\sigma,n) \quad = \quad \begin{cases} n+1, & \text{if } \sigma = \lambda_1 \\ n-1, & \text{if } \sigma = \lambda_2 \\ n, & \text{otherwise.} \end{cases}$$

Because we want to minimize the discounted total cost, "max" in the optimality equation (8.9) should be replaced by "min". Then, the optimal value function $V^*(i,j;n)$ (here $q = (i,j)$) is the unique bounded solution of the following optimality equation,

$$V(i,j;n) \quad = \quad \min_{\gamma\in\Gamma_n}\{c(i,j,\gamma) + \beta \max_{\sigma\in\Sigma_\gamma(i,j)} V(\delta(\sigma,i,j),\eta(\sigma,n))\},$$
$$(i,j) \in Q, \quad n \in \mathcal{N}, \qquad (8.23)$$

where $\Gamma_{-N+1} = \Gamma^-, \Gamma_n = \Gamma$ for $n = -N+2,\ldots,N-2$ and $\Gamma_{N-1} = \Gamma^+$.

Equation (8.23) can be simplified. We consider equations for

$$(i,j) \in Q^* := \{(1,2),(1,1),(2,2),(2,1),(1^*,2),(1,2^*),(1^*,1),(2,2^*)\}$$

as follows.

$$\begin{aligned} V(1,2;n) &= \beta\max\{V(1^*,2;n),V(1,2^*;n)\}, \\ V(1,1;n) &= \beta\max\{V(1,0;n),V(1^*,1;n)\}, \\ V(2,2;n) &= \beta\max\{V(0,2;n),V(2,2^*;n)\}, \\ V(2,1;n) &= \beta\max\{V(0,1;n),V(2,0;n)\}, \\ V(1^*,2;n) &= V(1,2^*;n) = \beta V(2,1;n), \\ V(1^*,1;n) &= \beta V(0,1;n), \quad V(2,2^*;n) = \beta V(2,0;n). \end{aligned}$$

The above set of equations is obviously equivalent to the following set.

$$V(1,2;n) \quad = \quad \beta^2 V(2,1;n),$$

$$\begin{aligned}
V(1,1;n) &= \beta\max\{V(1,0;n),\beta V(0,1;n)\},\\
V(2,2;n) &= \beta\max\{V(0,2;n),\beta V(2,0;n)\},\\
V(2,1;n) &= \beta\max\{V(0,1;n),V(2,0;n)\},\\
V(1^*,2;n) &= V(1,2^*;n)=\beta V(2,1;n),\\
V(1^*,1;n) &= \beta V(0,1;n),\ \ V(2,2^*;n)=\beta V(2,0;n). \qquad (8.24)
\end{aligned}$$

Hence, it suffices to solve the following equations.

$$\begin{aligned}
V(i,j;n) &= \min_{\gamma\in\Gamma_n}\{c(i,j,\gamma)+\beta\max_{\sigma\in\Sigma_\gamma(i,j)}V(\delta(\sigma,i,j),\eta(\sigma,n))\},\\
&\qquad\qquad (i,j)\notin Q^*,\ \ n\in\mathcal{N},\\
V(1,2;n) &= \beta^2 V(2,1;n),\ \ n\in\mathcal{N},\\
V(1,1;n) &= \beta\max\{V(1,0;n),\beta V(0,1;n)\},\ \ n\in\mathcal{N}, \qquad (8.25)\\
V(2,2;n) &= \beta\max\{V(0,2;n),\beta V(2,0;n)\},\ \ n\in\mathcal{N},\\
V(2,1;n) &= \beta\max\{V(0,1;n),V(2,0;n)\},\ \ n\in\mathcal{N}.
\end{aligned}$$

The above set of equations can be computed by successive approximation. That is, for any given initial value of $V_0(i,j;n)$ (e.g., $V_0(i,j;n)=0$ for all i,j,n), we iteratively compute $V_{k+1}(i,j)$ for $k=0,1,\ldots$ by

$$\begin{aligned}
V_{k+1}(i,j;n) &= \min_{\gamma\in\Gamma_n}\{c(i,j,\gamma)+\beta\max_{\sigma\in\Sigma_\gamma(i,j)}V_k(\delta(\sigma,i,j),\eta(\sigma,n))\},\\
&\qquad\qquad (i,j)\notin Q^*,\ \ n\in\mathcal{N},\\
V_{k+1}(1,2;n) &= \beta^2 V_k(2,1;n),\ \ n\in\mathcal{N},\\
V_{k+1}(1,1;n) &= \beta\max\{V_k(1,0;n),\beta V_k(0,1;n)\},\ \ n\in\mathcal{N}, \qquad (8.26)\\
V_{k+1}(2,2;n) &= \beta\max\{V_k(0,2;n),\beta V_k(2,0;n)\},\ \ n\in\mathcal{N},\\
V_{k+1}(2,1;n) &= \beta\max\{V_k(0,1;n),V_k(2,0;n)\},\ \ n\in\mathcal{N}.
\end{aligned}$$

Similarly to that in Markov decision processes (see Section 2.4), it can be proven that

$$\lim_{k\to\infty}V_k(i,j;n)=V(i,j;n),\ \ \forall i,j,n.$$

So, for a given small constant $\varepsilon>0$, when $|V_{k+1}(i,j;n)-V_k(i,j;n)|<\varepsilon$ for all i,j,n, we stop the above iterative procedure and take $V_{k+1}(i,j;n)$ as an approximating value of $V(i,j;n)$ for i,j,n.

In the following, we solve the optimality equation for $N=2$; that is, there are only two trays in the shop. First, we write Eq. (8.25) in the following three cases.

Case 1. $n=0$. In this case,

$$V(0,0;0) = \min\{c_2+\beta V(1,0;1),c_1+\beta V(0,2;-1),$$

$$
\begin{aligned}
& & \beta \max\{V(1,0;1), V(0,2;-1)\}\},\\
V(1,0;0) &= \min\{c_2 + \beta V(0,1;0), \beta \max\{V(0,1;0), V(1,2;-1)\}\},\\
V(0,2;0) &= \min\{c_1 + \beta V(2,0;0), \beta \max\{V(1,2;1), V(2,0;0)\}\},\\
V(0,1;0) &= \min\{c_1 + \beta V(0,0;0), \beta \max\{V(0,0;0), V(1,1;1)\}\},\\
V(2,0;0) &= \min\{c_2 + \beta V(0,0;0), \beta \max\{V(0,0;0), V(2,2;-1)\}\},\\
V(1,2;0) &= \beta^2 V(2,1;0),\\
V(1,1;0) &= \beta \max\{V(1,0;0), \beta V(0,1;0)\},\\
V(2,2;0) &= \beta \max\{V(0,2;0), \beta V(2,0;0)\},\\
V(2,1;0) &= \beta \max\{V(0,1;0), V(2,0;0)\}.
\end{aligned}
$$

Case 2. $n = -1$. In this case,

$$
\begin{aligned}
V(0,0;-1) &= c_2 + \beta V(1,0;0),\\
V(1,0;-1) &= c_2 + \beta V(0,1;-1),\\
V(0,2;-1) &= \min\{c_1 + \beta V(2,0;-1), \beta \max\{V(1,2;0), V(2,0;-1)\}\},\\
V(0,1;-1) &= \min\{c_1 + \beta V(0,0;-1), \beta \max\{V(0,0;-1), V(1,1;0)\}\},\\
V(2,0;-1) &= c_2 + \beta V(0,0;-1),\\
V(1,2;-1) &= \beta^2 V(2,1;-1),\\
V(1,1;-1) &= \beta \max\{V(1,0;-1), \beta V(0,1;-1)\},\\
V(2,2;-1) &= \beta \max\{V(0,2;-1), \beta V(2,0;-1)\},\\
V(2,1;-1) &= \beta \max\{V(0,1;-1), V(2,0;-1)\}.
\end{aligned}
$$

Case 3. $n = 1$. In this case,

$$
\begin{aligned}
V(0,0;1) &= c_1 + \beta V(0,2;0),\\
V(1,0;1) &= \min\{c_2 + \beta V(0,1;1), \beta \max\{V(1,2;0), V(0,1;1)\}\},\\
V(0,2;1) &= c_1 + \beta V(2,0;1),\\
V(0,1;1) &= c_1 + \beta V(0,0;1),\\
V(2,0;1) &= \min\{c_2 + \beta V(0,0;1), \beta \max\{V(0,0;1), V(2,2;0)\}\},\\
V(1,2;1) &= \beta^2 V(2,1;1),\\
V(1,1;1) &= \beta \max\{V(1,0;1), \beta V(0,1;1)\},\\
V(2,2;1) &= \beta \max\{V(0,2;1), \beta V(2,0;1)\},\\
V(2,1;1) &= \beta \max\{V(0,1;1), V(2,0;1)\}.
\end{aligned}
$$

We use successive approximation to solve the above equations with $c_1 = 1, c_2 = 5, \beta = 0.99$, and $\varepsilon = 0.01$. Successive approximation stops when the iteration steps are 517 and the result is given in Table 8.1.

Table 8.1. Optimal values for $c_1 = 1, c_2 = 5$, and $\beta = 0.99$.

$V(i,j;n)$	$(0,0)$	$(1,0)$	$(0,2)$	$(0,1)$	$(2,0)$
$n=0$	163.51	163.00	163.00	161.87	164.65
$n=-1$	166.37	168.05	168.00	164.71	169.71
$n=1$	162.37	160.12	160.13	161.75	160.75
$V(i,j;n)$	$(1,2)$	$(1,1)$	$(2,2)$	$(2,1)$	–
$n=0$	159.76	161.37	161.37	163.00	–
$n=-1$	164.65	166.36	166.32	168.00	–
$n=1$	156.92	158.52	158.52	160.12	–

From this, with the optimality equation, we obtain the optimal supervisor as given in Table 8.2, where

$$\Sigma_1 = \Sigma - \{\lambda_1\}, \;\; \Sigma_2 = \Sigma - \{\lambda_2\}.$$

In fact, the optimal supervisor takes the maximum among all available control inputs except at $(0,0;0)$ where it takes Σ_2, a real subset of the maximum Σ among all available control inputs.

Table 8.2. Optimal supervisor for $c_1 = 1, c_2 = 5$, and $\beta = 0.99$.

$f^*(i,j;n)$	$(0,0)$	$(1,0)$	$(0,2)$	$(0,1)$	$(2,0)$	$(1,2)$	$(1,1)$	$(2,2)$	$(2,1)$
$n=0$	Σ_2	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ
$n=-1$	Σ_2	Σ_2	Σ	Σ	Σ_2	Σ	Σ	Σ	Σ
$n=1$	Σ_1	Σ	Σ_1	Σ_1	Σ	Σ	Σ	Σ	Σ

When the discount factor β is smaller, the number of steps that is needed will be smaller. For example, with the same parameters as above but with the discount factor $\beta = 0.95$, the number of iteration steps for stopping the successive approximation is 103.

Now we discard the restriction on $\Gamma(s)$ and the finiteness of trays. Suppose that $\Gamma(s) = \Gamma$ for all $s \in L(G)$ and there are infinite trays in the shop. The objective is to control the system such that the number of completed jobs of any type that are waiting for output is at most one. That is, the language to be synthesized is $K = \{s \in L(G) : ||s|_{12}| \leq 1\}$.

Then from the results in Section 5, it is easy to see that $K^* = K$,

$$\pi_K(s) = \begin{cases} \Sigma - \{\lambda_1\}, & \text{if } |s|_{12} = 1 \\ \Sigma - \{\lambda_2\}, & \text{if } |s|_{12} = -1 \\ \Sigma, & \text{otherwise} \end{cases}$$

and

$$K^{\uparrow} = L(\pi_K/G) = \{s \in K : ||t|_{12}| \leq 1, \forall t \leq s\}$$

is the maximal closed sublanguage of K. The optimal supervisor π_K can be realized by

(a) A three-states automaton $\mathcal{S} = \{S, \Sigma_c, \eta\}$ with the state space $S = \{-1, 0, 1\}$, the event set $\Sigma_c = \{\lambda_1, \lambda_2\}$, and the state transition function $\eta(\sigma, n)$ defined previously, and

(b) A map $\phi : S \to \Gamma$ such that $\phi(1) = \Sigma - \{\lambda_1\}$, $\phi(-1) = \Sigma - \{\lambda_2\}$, and $\phi(0) = \Sigma$.

This means that the control input is $\phi(n)$ whenever the automaton $\mathcal{S}$ is at state n. This automaton $\mathcal{S}$ is simpler than the original system G described in Figure 8.1.

7. Notes and References

The control pattern in the standard models (e.g., [104]) of supervisory control is Γ, the set of all control inputs. Golaszewski and Ramadge [45] and Li et al. [89] studied supervisory control problems with a control pattern that is a subset of Γ. This includes the supervisory control problem with forced events [45] and control of timed DESs [8] as special cases. Golaszewski and Ramadge [45] studied the full observation case and Li et al. [89] studied the partial observation case under the restrictability condition. Takai [134] also studied the partial observation case but discarded the restrictability condition. All these papers discussed event feedback control.

The control pattern in this chapter is more general than that in Golaszewski and Ramadge [45], Li et al. [89], and Takai [134].

This chapter is from Hu and Yue [79].

Further research may include optimal control and supervisory control with arbitrary control pattern of DESs with incomplete information (such as partially observable DESs and decentralized supervisory control), and so on.

Problems

1. In the model discussed in this chapter, we study the problem among the stationary policies. Whether or not the optimal value will be improved when Markov policies are considered?

2. The DES $G = (Q, \Sigma, \delta, q_0)$ discussed in this chapter is called deterministic. On the other hand, a non-deterministic DES is defined as $G = (Q, \Sigma, \delta, q_0)$, where if an event σ occurs at state q then the next state is in the state subset $\delta(\sigma, q) \subset Q$. The difference between the deterministic DES and the non-deterministic DES is that whether $\delta(\sigma, q)$ is a state or a state subset. Please Generalize the model discussed in Section 1 to the non-deterministic DES.

3. Study the job-matching problem if the deadlock is allowed but a cost ρ is incurred when to resolve the deadlock.

Chapter 9

OPTIMAL REPLACEMENT UNDER STOCHASTIC ENVIRONMENTS

Optimal replacement has been an interesting research area for a long time. It considers a system (or a machine) that will deteriorate as it operates and thus should be replaced by a new one when it is too bad. There are two types of deterioration considered in reliability literature. The first one is due to the operation of the system itself, and the second one is caused by the influence of the environment, for example, shocks to the system. We call these two types, respectively, system deterioration and environment deterioration.

This chapter applies Markov decision processes in semi-Markov environments discussed in Chapter 6 to investigate two optimal replacement problems of systems in changing environments. In each problem, the system deteriorates according to a Markov process and is further influenced by its environment. In the first problem, the system deteriorates according to a discrete time Markov chain and is further subject to random shocks from its environment. Optimal extended control limit policies are shown for the discounted expected total costs with finite and infinite horizons and for the average criterion. A transformation from infinite states to finite states is discussed. In the second problem, both the system and its environment are described by semi-Markov processes, and each change of the environment's state will change the parameters of the semi-Markov process modeling the system and also cause damage to the system. Optimal extended control limit policies are also shown for the discounted expected total costs with both finite and infinite horizons. It is shown that the result is robust with respect to the d.f.s of sojourning times. Moreover, the Markov environment is studied with a simplification on the results and a transformation from infinite states to finite states. Finally, numerical example is given.

1. Optimal Replacement: Discrete Time

1.1 Problem and Model

The replacement model investigated in this section is described as follows.

The system is observed periodically at discrete time periods $t = 0, t_0, 2t_0, \ldots$ for some $t_0 > 0$. For convenience, we set $t_0 = 1$. Its state at the observation periods forms a Markov chain with the state space $S = \{0, 1, 2, \cdots\}$, where state 0 represents that the system is new, and states $1, 2, \ldots$ represent the various degrees of deterioration of the system. The larger the value, the more serious the deterioration will be. When the system is observed in state $i \in S$, one of the following two actions can be chosen.

1. Operate the system continually (denoted by O). Then the cost in one period is $b(i)$ and the probability that the state will be j at the next period is p_{ij} (which is called the natural state transition probability).

2. Replace the system with a new one (denoted by R) (the time of the replacement is assumed to be one period). Then the cost of the replacement is $d(i)$ and the state at the next period will be 0 with probability one.

In addition, the system is subject to random shocks from its environment. It is assumed that the shocks occur right before some discrete time periods $t = 0, 1, 2, \ldots$. Following each shock, an instantaneous state transition occurs according to a probability law $\{q_{ij}\}$ with a cost $R(i, j)$. Let

$$R(i) = \sum_j q_{ij} R(i, j)$$

be the expected cost at state i caused by the environment's shocks. We call $b(i)$, $d(i)$, and $R(i)$ the operation cost, the replacement cost, and the shocking cost, respectively. Here, it is assumed that the transition caused by shocks is preemptive and that the shock occurred when the system to be replaced has no functions. The times between two adjacent shocks are independent and identical distributed (i.i.d.) and ξ is used to represent such a time. So ξ is a random variable of discrete type with the probability law

$$p(k) := Pr\{\xi = k\}, \quad k = \infty, 1, 2, \ldots.$$

We introduce the following condition as it is usual in the literature.

Condition 9.1: *1. For each $j \geq 0$, both $\sum_{m=j}^{\infty} p_{im}$ and $\sum_{m=j}^{\infty} q_{im}$ are nondecreasing in i.*

2. $b(i), d(i), b(i)-d(i), R(i)$ are all nondecreasing in i and $b(i), d(i), R(i)$ are all nonnegative.

Obviously, 1 of Condition 9.1 above represents that the larger the deteriorative degree of the system, the larger the speed of deterioration resulting from

the natural state transition or from shocks will be. $b(i) - d(i)$ is nondecreasing indicates that the operation cost increases faster than the replacement cost as the deterioration degree of the system increases.

The following lemma is well known and is preparation for later discussions.

Lemma 9.1: *Let (r_{ij}) be a transition probability matrix. Then the following two statements are equivalent.*

1. *For each $m \geq 0, \sum_{j=m}^{\infty} r_{ij}$ is nondecreasing in i.*
2. *For each nonnegative and nondecreasing function $h(j)$, $v(i) := \sum_j r_{ij} h(j)$ is nondecreasing.*

Proof: Suppose that 1 holds and $h(j)$ is a nonnegative and nondecreasing function. Let the following nonnegative constants

$$C_0 = h(0), \;\; C_m = h(m) - h(m-1), \;\; m = 1, 2, \ldots$$

and the following nondecreasing functions

$$u_m(j) = \begin{cases} 1, & \text{if } m \leq j, m = 0, 1, 2, \ldots \\ 0, & \text{if } m > j. \end{cases}$$

Then

$$h(i) = \sum_{m=0}^{\infty} C_m u_m(i), \;\; i = 0, 1, 2, \ldots$$

and

$$\begin{aligned} v(i) &= \sum_j r_{ij} h(j) = \sum_j r_{ij} \sum_{m=0}^{\infty} C_m u_m(j) \\ &= \sum_{m=0}^{\infty} C_m \sum_j r_{ij} u_m(j) \\ &= \sum_{m=0}^{\infty} C_m \sum_{j=m}^{\infty} r_{ij}. \end{aligned}$$

Thus, $v(i)$ is nondecreasing in i.

Now, suppose that 2 holds. Then, for the nondecreasing function $u_m(j)$ defined above, the function

$$v(i) := \sum_j r_{ij} u_m(j) = \sum_{j=m}^{\infty} r_{ij}$$

is obviously nondecreasing for each $m \geq 0$. □

The replacement problem described previously is surely a problem of Markov decision processes in a stochastic environment. Because both the system and its environment are described by discrete time models, we can simply use Markov decision processes to model it. Hence, we construct a MDP model for the optimal replacement problem.

For $k \geq 0, i \in S$, let (k, i) denote that the system is in state i and the last shock occurred before k periods. We call (k, i) the (mathematical) state of the system. The feasible action set at (k, i) is $A = \{O, R\}$. If action O is chosen at state (k, i), then the probability that the next state will be (k', j) is

$$P((k', j) \mid (k, i), O) = \begin{cases} [1 - \lambda(k)]p_{ij}, & \text{if } k' = k + 1, j \in S \\ \lambda(k)q_{ij}, & \text{if } k' = 0, j \in S \\ 0, & \text{otherwise,} \end{cases}$$

where

$$\lambda(k) = p(k + 1)/P(k)$$

and

$$P(k) = \sum_{m=k+1}^{\infty} p(m) = P\{\xi > k\}.$$

Here, $\lambda(k)$ represents the probability that the shock will occur at period $k + 1$ conditioned on that the shock has not occurred during the first k periods. When $P(k) = 0$ we set $\lambda(k) = 0$. If action R is chosen at state (k, i), then the probability that the next state will be (k', j) is

$$P((k', j) \mid (k, i), R) = \begin{cases} 1 - \lambda(k), & \text{if } k' = k + 1, j = 0 \\ \lambda(k), & \text{if } k' = 0, j = 0 \\ 0, & \text{otherwise.} \end{cases}$$

When action O or R is chosen, the expected total cost in one period will be, respectively,

$$\begin{aligned} r(k, i, O) &= b(i) + \beta\lambda(k)R(i), \\ r(k, i, R) &= d(i), \end{aligned} \tag{9.1}$$

where $\beta \in [0, 1]$ is the discount factor.

We denote by $V_{\beta,n}(k, i)$ the minimal expected discounted total cost when the state is (k, i) and there remain n periods to go. From [52] or [118], $V_{\beta,n}(k, i)$ satisfies the following optimality equation for the finite horizon problem.

$$\begin{aligned} V_{\beta,n}(k, i) &= \min\{V_{\beta,n}^{(O)}(k, i), V_{\beta,n}^{(R)}(k, i)\}, \quad n \geq 0, \\ V_{\beta,0}(k, i) &= 0, \end{aligned} \tag{9.2}$$

where

$$\begin{aligned} V_{\beta,n}^{(O)}(k,i) &= b(i) + \beta\lambda(k)R(i) + \beta[1-\lambda(k)]\sum_j p_{ij}V_{\beta,n-1}(k+1,j) \\ &\quad + \beta\lambda(k)\sum_j q_{ij}V_{\beta,n-1}(0,j), \\ V_{\beta,n}^{(R)}(k,i) &= d(i) + \beta[1-\lambda(k)]V_{\beta,n-1}(k+1,0) + \beta\lambda(k)V_{\beta,n-1}(0,0) \end{aligned}$$

are the minimal expected discounted total costs when the state is (k,i) and action O or R is chosen at the current period and an optimal policy is used in the remaining $n-1$ periods.

Now, we denote by $V_\beta(k,i)$ the minimal expected discounted total cost for infinite horizons when the initial state is (k,i). From Chapter 2, we know that $V_\beta(k,i)$ satisfies the following optimality equation for the infinite horizons (here $\beta < 1$ is assumed)

$$V_\beta(k,i) = \min\{V_\beta^{(O)}(k,i), V_\beta^{(R)}(k,i)\}, \tag{9.3}$$

where the two components $V_\beta^{(O)}(k,i)$ and $V_\beta^{(R)}(k,i)$ are similarly defined as, respectively, $V_{\beta,n}^{(O)}(k,i)$ and $V_{\beta,n}^{(R)}(k,i)$ by only deleting the subscripts n and $n-1$. Moreover, we have

$$\lim_{n\to\infty} V_{\beta,n}(k,i) = V_\beta(k,i), \quad k \geq 0, i \geq 0 \tag{9.4}$$

and so

$$\lim_{n\to\infty} V_{\beta,n}^{(a)}(k,i) = V_\beta^{(a)}(k,i), \quad a = O, R.$$

If we define

$$\begin{aligned} v_n(k,i) &= V_{\beta,n}^{(O)}(k,i) - V_{\beta,n}^{(R)}(k,i), \\ f_n^*(k,i) &= \begin{cases} O, & \text{if } v_n(k,i) \leq 0 \\ R, & \text{otherwise,} \end{cases} \\ v(k,i) &= V_\beta^{(O)}(k,i) - V_\beta^{(R)}(k,i), \\ f^*(k,i) &= \begin{cases} O, & \text{if } v(k,i) \leq 0 \\ R, & \text{otherwise,} \end{cases} \end{aligned}$$

then from [52], [125], and Chapter 2 we have that

1. for $N > 0, (f_0^*, f_1^*, \ldots, f_N^*)$ is an optimal policy for the discounted expected total cost criterion with N horizons.
2. $(f^*)^\infty$ is an optimal policy for discounted expected total cost criterion with infinite horizons.

It should be noted that if there is k_0 such that $P\{\xi \leq k_0\} = 1$, then the mathematical state set $\{(k,i) : k \geq 0, i \in S\}$ can obviously be restricted to $\{(k,i) : k \leq k_0, i \in S\}$. This set will be finite whenever S is finite.

1.2 Total Cost Criterion

First, we prove the following lemma on the monotone of the optimal values.

Lemma 9.2: *For each $n \geq 0$ and $k \geq 0$, both $V_{\beta,n}(k,i)$ and $v_n(k,i)$ are nondecreasing in i. Thus, both $V_\beta(k,i)$ and $v(k,i)$ are also nondecreasing in i.*

Proof: We use the induction method to prove that $V_{\beta,n}(k,i)$ is nondecreasing in i.

When $n = 0$, the result is trivial. Suppose that the result is true for some $n \geq 0$. Then, from Condition 9.1, Lemma 9.1, and $\lambda(k) \in [0,1]$, we have that both $V^{(O)}_{\beta,n+1}(k,i)$ and $V^{(R)}_{\beta,n+1}(k,i)$ are nondecreasing in i. This implies together with Eq. (9.2) that $V_{\beta,n+1}(k,i)$ is also nondecreasing in i.

From the definition of $v_n(k,i)$ we get that

$$\begin{aligned} v_n(k,i) &= b(i) - d(i) + \beta\lambda(k)R(i) \\ &\quad + \beta[1-\lambda(k)][\sum_j p_{ij} V_{\beta,n-1}(k+1,j) - V_{\beta,n-1}(k+1,0)] \\ &\quad + \beta\lambda(k)[\sum_j q_{ij} V_{\beta,n-1}(0,j) - V_{\beta,n-1}(0,0)]. \end{aligned} \tag{9.5}$$

With this, Lemma 9.1, Condition 9.1, and the results proved above, we have that $v_n(k,i)$ is also nondecreasing in i.

From the above results and Eq. (9.4), we conclude that both $V_\beta(k,i)$ and $v(k,i)$ are also nondecreasing in i. □

For $n \geq 0$ and $k \geq 0$, we define

$$i^*_n(k) = \min\{i \mid v_n(k,i) \geq 0\}.$$

As $v_n(k,i)$ is nondecreasing in i, $v_n(k,i) < 0$ if and only if $i < i^*_n(k)$. Then

$$f^*_n(k,i) = \begin{cases} O, & \text{if } i < i^*_n(k) \\ R, & \text{if } i \geq i^*_n(k). \end{cases} \tag{9.6}$$

That is, when n periods remain and the last shock occurred before k periods, the optimal action is to operate the system if and only if the system's deterioration degree is less than $i^*_n(k)$. We call $i^*_n(k)$ a state limit at (n,k) and call such a policy *an extended control limit policy*.

Similarly, for the infinite horizons, we define

$$i^*(k) = \min\{i \mid v(k,i) \geq 0\}.$$

Also, $f^*(k,i) = O$ if and only if $i < i^*(k)$. Hence, f^* is also an extended control limit policy for the infinite horizons. We have the following theorem on the optimal policies.

Theorem 9.1: *We have the following two statements.*

1. *There are extended control limit policies* $(f_1^*, f_2^*, \ldots, f_N^*)$ *and* f^* *for the discounted expected total cost criteria in finite and infinite horizons, respectively.*
2. *Let*

$$i_0(k) = \min\{i \mid b(i) - d(i) + \beta\lambda(k)R(i) \geq 0\}$$

($i_0(k) = +\infty$ *when the above set* $\{\cdot\}$ *is empty). Then* $i_n^*(k) \leq i_0(k)$ *for all* $n \geq 1$ *and* $i^*(k) \leq i_0(k)$.

Proof: It suffices to prove 2. With Condition 9.1, Lemmas 9.1 and 9.2, and Eq. (9.5), we have that

$$v_n(k,i) \geq b(i) - d(i) + \beta\lambda(k)R(i), \quad n \geq 1, k \geq 0, \quad i \in S.$$

This implies with the definition of $i_0(k)$ that $v_n(k, i_0(k)) \geq 0$ for $n \geq 1$ and $k \geq 0$. Hence, $i_0(k) \geq i_n^*(k)$ and $i_0(k) \geq i^*(k)$. □

Result 2 in Theorem 9.1 gives a upper bound for the state limit $i_n^*(k)$ and $i^*(k)$. So, the system must be replaced when the state is in (k, i) with $i \geq i_0(k)$, irrespectively of how many periods remain. Moreover, if we denote

$$i_0 = \min\{i : b(i) - d(i) \geq 0\},$$

then $i_n^*(k) \leq i_0$ and $i^*(k) \leq i_0$ for all n and k. So, the system must be replaced when the deterioration degree exceeds i_0, which is irrespective of n and k. We thus conjecture that if i_0 is finite then we can contract the state subset $\{i_0, i_0 + 1, i_0 + 2, \ldots\}$ to one state. This is proved in the following.

Denote

$$V_{\beta,n}(k) = \beta[1 - \lambda(k)]V_{\beta,n-1}(k+1, 0) + \beta\lambda(k)V_{\beta,n-1}(0,0).$$

Then,

$$V_{\beta,n}^{(R)}(k,i) = d(i) + V_{\beta,n}(k), \quad i \geq 0, \quad k \geq 0, \quad n \geq 0.$$

Due to Theorem 9.1,

$$V_{\beta,n}(k,i) = d(i) + V_{\beta,n}(k), \quad i \geq i_0, \quad n \geq 0, \quad k \geq 0. \tag{9.7}$$

For $i \leq i_0$, one can get from Eq. (9.7) that

$$\begin{aligned} V_{\beta,n+1}^{(O)}(k,i) &= b(i) + \beta\lambda(k)R(i) \\ &\quad + \beta[1 - \lambda(k)]\{\sum_{j<i_0} p_{ij}V_{\beta,n}(k+1,j) + \sum_{j\geq i_0} p_{ij}[d(j) + V_{\beta,n}(k+1)]\} \\ &\quad + \beta\lambda(k)\{\sum_{j<i_0} q_{ij}V_{\beta,n}(0,j) + \sum_{j\geq i_0} q_{ij}[d(j) + V_{\beta,n}(0)]\}. \end{aligned}$$

For $i \le i_0$, let

$$\begin{aligned}
\bar{b}(i) &= b(i) + \beta \sum_{j \ge i_0} p_{ij}[d(j) - d(i_0)], \\
\bar{R}(i) &= R(i) + \sum_{j \ge i_0} (q_{ij} - p_{ij})[d(j) - d(i_0)], \\
\bar{p}_{ij} &= \begin{cases} p_{ij}, & \text{if } j < i_0 \\ \sum_{j \ge i_0} p_{ij}, & \text{if } j = i_0, \end{cases} \\
\bar{q}_{ij} &= \begin{cases} q_{ij}, & \text{if } j < i_0 \\ \sum_{j \ge i_0} q_{ij}, & \text{if } j = i_0, \end{cases}
\end{aligned}$$

then

$$\begin{aligned}
V^{(O)}_{\beta,n+1}(k,i) &= \bar{b}(i) + \beta\lambda(k)\bar{R}(i) + \beta[1-\lambda(k)] \sum_{j \le i_0} \bar{p}_{ij} V_{\beta,n}(k+1,j) \\
&\quad + \beta\lambda(k) \sum_{j \le i_0} \bar{q}_{ij} V_{\beta,n}(0,j), \;\; i \le i_0, \; k \ge 0, \; n \ge 0. \quad (9.8)
\end{aligned}$$

Now, we define a new replacement problem (NRP, for short) as follows:

1. The state space of NRP is $\{0, 1, \ldots, i_0\}$.

2. The natural and shock state transition probability matrices are $(\bar{p}_{ij})$ and $(\bar{q}_{ij})$, respectively.

3. The operation cost per period in state i is $\bar{b}_i$, the replacement cost in state i is $d(i)$, and the shocking cost at state i is $\bar{R}(i)$.

Let $V'_{\beta,n}(k,i)$ and $V'_{\beta}(k,i)$ be the minimal expected discounted total cost starting from state (k,i) for the finite and infinite horizons, respectively.

If we assume that

$$\begin{aligned}
&\textstyle\sum_{j \ge i_0}(q_{ij} - p_{ij})[d(j) - d(i_0)] \quad \text{is nonnegative and} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{nondecreasing in } i \le i_0, \qquad (9.9)
\end{aligned}$$

then NRP also satisfies Condition 9.1 and all the above results are still true for NRP. Thus, the following theorem is obtained.

Theorem 9.2: *NRP is equivalent to the original replacement problem (ORP for short) under the condition given in Eq. (9.9) in the following manner.*

1. $V'_{\beta,n}(k,i) = V_{\beta,n}(k,i)$ *and* $V'_{\beta}(k,i) = V_{\beta}(k,i)$ *for all* $i \le i_0$ *and* $n, k \ge 0$.

2. *Both optimality equations are the same for* $i \le i_0$*: Eq. (9.2) is for the finite horizons and Eq. (9.3) is for the infinite horizons.*

3. *Because the optimal policies $f_n^*(i)$ and $f^*(i)$ for ORP equal R when $i \geq i_0$, they can be viewed as policies for NRP (and the reverse is also true) and thus the optimal policies of NRP are the same as that of ORP.*

Compared with ORP, the main advantage of NRP is that it has finite states. This allows NRP to be solved simply. The above theorem means that under certain conditions, finite states are enough to model general optimal replacement problems.

Remark 9.1: *1. Similarly to the above, we can also contract the mathematical state subset $\{(k,i) : i \geq i_0(k)\}$ to one state. Although, in this case, there are no corresponding new replacement problems, the number of mathematical states is smaller than that in NRP defined above because $i_0 \geq i_0(k)$.*

2. If $\bar{b}(i) = +\infty$ for some i, that is, the operation cost per period at state i in NRP is infinite, then the optimal policy in state i is replacement (R) for both the finite and infinite horizons.

If we define

$$i^* = \max\{i : i \leq i_0, \bar{b}(i) < +\infty\},$$

then $\bar{b}(i) = +\infty$ for $i = i^* + 1, i^* + 2, \ldots, i_0$ and the optimal actions in states $i^* + 1, \ldots, i_0$ are R. Thus, from the above discussion, we can again contract $\{i^* + 1, i^* + 2, \ldots, i_0\}$ to one state. Especially, if $\bar{b}(0) = +\infty$, then the state set of the system can be contracted to one state and the optimal action is always R.

1.3 Average Criterion

In this subsection, we discuss the optimal replacement problem for the average criterion. We assume that i_0 is finite throughout this subsection.

For ORP, the set of decision functions is

$$F = \{f : f(i) = O \text{ or } R, i \in S\}.$$

We denote by $(Y_t, X_t), \Delta_t$ the mathematical state and action chosen, respectively, at period t for $t \geq 0$. The average objective function is defined by

$$\bar{V}(f,k,i) \quad = \quad \liminf_{n\to\infty} \frac{1}{n} V_{1,n}(f,k,i,), \quad f \in F, k \geq 0, i \in S,$$

where for $\beta \in [0,1]$ and $n \geq 0$,

$$V_{\beta,n}(f,k,i,) = \sum_{t=0}^{n-1} \beta^t E_f\{r(Y_t, X_t, f(Y_t, X_t)) \mid Y_0 = k, X_0 = i\}$$

is the discounted expected total cost starting from state (k, i) when n periods remain under policy f.

For NRP, its decision function set is

$$F' = \{f : f(i) = O \text{ or } R, i \le i_0\}$$

and the average objective functions $\bar{V}'(f, k, i)$ and $V'_{\beta,n}(f, k, i)$ are defined similarly.

For a decision function $f \in F$, we define another decision function $g_f \in F$ by

$$g_f(k, i) = \begin{cases} f(k, i), & \text{if } i < i_0, k \ge 0 \\ R, & \text{if } i \ge i_0, k \ge 0. \end{cases} \tag{9.10}$$

The difference between f and g_f is that under g_f the action chosen is always R when $i \ge i_0$. For these two policies, we have the following result.

Lemma 9.3: *$\bar{V}(f, k, i) \ge \bar{V}(g_f, k, i)$ for all $k \ge 0$ and $i \in S$.*

Proof: First, we use the induction method to prove that for $n \ge 0$,

$$V_{\beta,n}(f, k, i) \ge V_{\beta,n}(g_f, k, i), \ \ \beta \in [0, 1], \ \ k \ge 0, \ \ i \in S. \tag{9.11}$$

When $n = 0$, both sides equal zero. Suppose that Eq. (9.11) holds for some $n \ge 0$. Due to the definition of i_0 and g_f, $r(k, i, f(k, i)) \ge r(k, i, g_f(k, i))$ for $k \ge 0$ and $i \in S$. Hence,

$$\begin{aligned} V_{\beta,n+1}(f, k, i) &= r(k, i, f(k, i)) + \beta[1 - \lambda(k)] \sum_j p_{ij} V_{\beta,n}(f, k + 1, j) \\ &\quad + \beta\lambda(k) \sum_j q_{ij} V_{\beta,n}(f, 0, j) \\ &\ge r(k, i, g_f(k, i)) + \beta[1 - \lambda(k)] \sum_j p_{ij} V_{\beta,n}(g_f, k + 1, j) \\ &\quad + \beta\lambda(k) \sum_j q_{ij} V_{\beta,n}(g_f, 0, j) \\ &= V_{\beta,n+1}(g_f, k, i), \ \ k \ge 0, \ \ i \in S. \end{aligned}$$

Thus Eq. (9.11) is true, which implies that $\bar{V}(f, k, i) \ge \bar{V}(g_f, k, i)$ for all k and i. □

From Lemma 9.3, we know that ORP can be considered only in a subset, F_0, of F. Here we define

$$F_0 := \{f \in F : f(k, i) = R \text{ for } i \ge i_0\},$$

which is isomorphically equivalent to F'.

The following lemma immediately follows Theorem 9.2.

Lemma 9.4: $\bar{V}(f,k,i) = \bar{V}'(f,k,i)$ *for* $f \in F', k \geq 0, i \leq i_0$.

The above lemma implies that for the average criterion, any optimal policy for NRP is also optimal for ORP. Conversely, if f is an optimal policy for ORP, then the policy g_f defined by Eq. (9.10) is also optimal for NRP. So, NRP is equivalent to ORP for the average criterion.

Now, for the average criterion, we have proved that ORP can be transformed to NRP when $i_0 < +\infty$. However, does an optimal policy exist? Moreover, does an optimal extended control limit policy exist? For the first question, there are many conditions presented in the MDPs literature to ensure the existence of an optimal stationary policy (for example, see [36] and Chapter 3 of this book).

In the following, we only discuss the latter problem. If there is $k_0 > 0$ such that

$$P\{\xi \leq k_0\} = 1 \tag{9.12}$$

(i.e., ξ is bounded), then the mathematical state set $\{(k,i) : k \leq k_0, i \leq i_0\}$ is finite. In this case we have the following theorem.

Theorem 9.3: *Suppose that ξ is bounded. Then there is an optimal extended control limit policy for both the ORP and NRP.*

Proof: It suffices to prove the result for NRP. Suppose that $\{\beta_n, n \geq 1\}$ is an arbitrary nonnegative sequence such that β_n is increasing and tends to 1 and f_n^* is the optimal policy for the discounted criterion with the discount factor being β_n. Here, f_n^* can be chosen to be the extended control limit. Because the mathematical state set is finite, there is a subsequence $\{\beta_{n_k}, k \geq 1\}$ of $\{\beta_n, n \geq 1\}$ such that all f_{n_k} are same (denoted by f^*). Then from the Abel theorem (see Lemma 3.4), we have that for any policy π,

$$\begin{aligned} \bar{V}(f^*,k,i) &= \lim_{k\to\infty}(1-\beta_{n_k})V_{\beta_{n_k}}(f_{n_k},k,i) \\ &\leq \liminf_{k\to\infty}(1-\beta_{n_k})V_{\beta_{n_k}}(\pi,k,i) \\ &= \bar{V}(\pi,k,i). \end{aligned}$$

So, f^* is optimal for the average criterion and is certainly an extended control limit policy. □

At the end of this section, we give the following remark.

Remark 9.2:

1. *There are two special cases of the shock from the environment. The first case is that $P\{\xi = +\infty\} = 1$; that is, there is no shock. In this case, the model and the corresponding results are classical in the literature on optimal replacement problems. The second case is that*

$q_{ij} \equiv 0$; *that is, the system shall be terminated when the first shock occurs. Then, the problem can be called the replacement problem with stochastic termination. So, all the results in this chapter are also true for this special case.*

2. *The countable (physical) state set of the system is transformed into a finite (physical) state set under the condition that i_0 is finite. This condition is always true in practical problems and implies that there exists a state i such that the operation cost is larger than the replacement cost in i.*

3. *For the case with finite physical state set $S = \{0, 1, 2, \ldots, L\}$, an assumption made in the literature is that the system must be replaced in state L, which is not necessary from a mathematical viewpoint. For example, when*

$$b(L) - d(L) + \beta\lambda(k)R(L) \leq 0, \tag{9.13}$$

$v_1(k, L) \leq 0$ and action O is optimal in (k, L). In fact, L means that the system has a failure, or cannot be used, and then must be replaced. This indicates that the cost functions determined are not adequate and must be redetermined when the condition given in Eq. (9.13) is true. For NRP, $L = i_0$ and $b(L) - d(L) \leq 0$, so action R is optimal in state L.

2. Optimal Replacement: Semi-Markov Processes

2.1 Problem

The optimal replacement problem considered in this section is as follows.

1. The system is in a semi-Markov environment $\{(J_n, T_n), n \geq 0\}$ on a set K of countable environment states, where J_n is the state of the environment immediately after its nth transition epoch T_n and $0 = T_0 < T_1 < T_2 < \cdots$. We denote the state's kernel of the semi-Markov environment by

$$G_{kk'}(t) = P\{T_{n+1} - T_n \leq t, J_{n+1} = k' | J_n = k\}, \;\; k, k' \in K, \;\; t \geq 0.$$

For $k, k' \in K$, let

$$\psi_{kk'} = G_{kk'}(\infty), \;\; G_k(t) = \sum_{k' \in K} G_{kk'}(t)$$

be the transition probability of the environment from state k to k' and the distribution function of the sojourning time of the environment in state k, respectively. The environment considered here is the same as that in Chapter 6.

2. While the environment is in state k with $k \in K$, the system itself operates according to a semi-Markov process with a kernel $\{P_{ij}^k(t), i, j \in S\}$ and a set

$S = \{0, 1, 2, \ldots\}$ of countable states. Here, state 0 represents a new system, and states $1, 2, \ldots$ represent the different degrees of deterioration of the system, and the larger the value, the more serious the deterioration will be. For $i, j \in S, k \in K$, and $t \geq 0$, let

$$p_{ij}^k = P_{ij}^k(\infty), \;\; T_{ij}^k(t) = P_{ij}^k(t)/p_{ij}^k, \;\; T_i^k(t) = \sum_{j \in S} P_{ij}^k(t).$$

Here, p_{ij}^k is the transition probability of the system from state i to state j, $T_{ij}^k(t)$ and $T_i^k(t)$ are the distribution functions of the sojourning times at the state i provided whether or not the next state will be j, under the condition that the environment is in state k.

3. Suppose that the environment is in state k. Then one of the following two actions can be chosen if the system state transfers to state i.

(a) Operate the system continually (denoted by O) with a cost rate $b^k(i)$.

(b) Replace the system with a new one (denoted by R) with a cost rate $d^k(i)$, and the time of the replacement is assumed to be a random variable with distribution function $F^k(t)$, and the state after replacement is 0.

4. When the environment state changes from k to k' and the system is in state i immediately before the transition of the environment, if action O is chosen then the system state will change immediately according to a probability distribution $\{q_{ij}^k, j \in S\}$ and an instantaneous cost $R^k(i, O)$ occurs, whereas if action R is chosen then the replacement is immediately completed with the system being in state 0 and an instantaneous cost $R^k(i, R)$ occurs.

5. The objective of the system is to minimize the expected discounted total cost in $[0, \infty)$ with discount rate $\alpha > 0$.

Such a system can be modeled by a semi-Markov decision process in a semi-Markov environment, presented and studied in Section 6.2, as follows.

During the environment state k (i.e., $J_n = k$ for some $n \geq 0$) the system can be modeled by the following SMDPs,

$$\text{SMDPs}^k := \{S, A, p^k(j|i, a), T^k(\cdot|i, a, j), r^k(i, a, j, u)\}, \tag{9.14}$$

where the state space S and the action set $A = \{O, R\}$ are given in the above. The transition probability p^k, the distribution function T^k of the sojourning time, and the one period cost function r^k are given, respectively, by

$$\begin{aligned}
&p^k(j|i, O) = p_{ij}^k, \\
&p^k(j|i, R) = \delta_{j0}, \\
&T^k(t|i, O, j) = T_{ij}^k(t), \\
&T^k(t|i, R, O) = F^k(t), \\
&r^k(i, O, j, u) = b^k(i)\alpha^{-1}(1 - e^{-\alpha u}), \\
&r^k(i, R, j, u) = \delta_{j0} d^k(i)\alpha^{-1}(1 - e^{-\alpha u})
\end{aligned} \tag{9.15}$$

for $i, j \in S, k \in K, t, u \geq 0$, where $\delta_{j0} = 1$ if $j = 0$ and $\delta_{j0} = 0$ otherwise. The decision epoch is the time when the system or the environment changes its states. We call the duration between two adjacent two decision epochs an horizon. The detailed meanings of the above elements can be found in Section 6.2 for SMDPs-SE and Chapter 5 for SMDPs.

For a SMDP model in a semi-Markov environment, when the environment state changes from k to k', that is, at T_{n+1} for some $n \geq 0$ with $J_n = k$ and $J_{n+1} = k'$, the system's state changes immediately to j with a probability $q(j|i, a, k, k')$ if the system's state is i at $T_{n+1} - 0$ and the last action taken before T_{n+1} is a, and at the same time, an instantaneous cost $R^k(i, a)$ occurs, where for $i, j \in S, k, k' \in K$,

$$q(j|i, O, k, k') = q_{ij}^k, \qquad q(j|i, R, k, k') = \delta_{j0}$$

and $R^k(i, O)$ and $R^k(i, R)$ are given previously.

To simplify the notations, for $k \in K$ and $s, t \geq 0$, we denote

$$\begin{aligned} b^k(s,t) &= \alpha^{-1}(1 - e^{-\alpha t})\,[1 - G_k(t+s)] \\ &\quad + \alpha^{-1}\int_{s+}^{s+t}(1 - e^{-\alpha(u-s)})dG_k(u), \\ g^{k,k'}(s,t) &= \int_{s+}^{s+t} e^{-\alpha(u-s)}dG_{kk'}(u), \\ g^k(s,t) &= \sum_{k' \in K} g^{k,k'}(s,t) = \int_{s+}^{s+t} e^{-\alpha(u-s)}dG_k(u). \end{aligned} \tag{9.16}$$

Let $x = (k, s, i) \in \Omega$, where

$$\Omega := \{(k, s, i) : k \geq 0, s \geq 0, i \in S\}$$

is a mathematical state which means that the environment has been in state k just since time s ago and the system's state has just transferred to i. For simplicity, we call x a state when no confusion results. Then, we define

$$\begin{aligned} r(x, O) &= b^k(i)\int_0^\infty b^k(s,t)dT_i^k(t) + R^k(i, O)\int_0^\infty g^k(s,t)dT_i^k(t), \\ r(x, R) &= d^k(i)\int_0^\infty b^k(s,t)dF^k(t) + R^k(i, R)\int_0^\infty g^k(s,t)dF^k(t), \\ \beta(x, O, k') &= \int_0^\infty g^{kk'}(s,t)dT_i^k(t), \\ \beta(x, R, k') &= \int_0^\infty g^{kk'}(s,t)dF^k(t). \end{aligned} \tag{9.17}$$

Here, $r(x, a)$ is the expected discounted cost occurring when state x is reached and action a is taken, and $\beta(x, a, k')$ corresponds to a discount factor depending on state x, action a, and the next environment state k'.

Now, due to [67] (see also Theorem 6.8), the minimal expected discounted total cost in $[0,\infty)$ starting from the initial state x, $V^*(x)$, is the minimal nonnegative solution of the following optimality equation

$$V^*(x) = \min\{V^*(x,O), V^*(x,R)\}, \tag{9.18}$$

where for $x=(k,s,i)\in\Omega$,

$$\begin{aligned} V^*(x,O) &= r(x,O) + \sum_{k'\in K}\beta(x,O,k')\sum_{j\in S} q_{ij}^k V^*(k',0,j) \\ &\quad + \sum_{j\in S} p_{ij}^k \int_0^\infty e^{-\alpha t} V^*(k,s+t,j)\,dT_{ij}^k(t), \\ V^*(x,R) &= r(x,R) + \sum_{k'\in K}\beta(x,R,k')V^*(k',0,0) \\ &\quad + \int_0^\infty e^{-\alpha t} V^*(k,s+t,0)\,dF^k(t) \end{aligned} \tag{9.19}$$

are, respectively, the discounted total cost if action O or R is used in the first horizon with the mathematical state x and then the optimal policy is used in the remaining horizons.

2.2 Optimal Control Limit Policies

From the standard results in discrete time Markov decision processes, Eq. (9.18) can be considered as an optimality equation for an adequate DTMDP model with state space Ω. Thus we can consider its n-horizon problem with the optimality equation

$$V_n^*(x) = \min\{V_n^*(x,O), V_n^*(x,R)\}, \qquad x=(k,s,i)\in\Omega, \tag{9.20}$$

where $V_n^*(x)$ is the optimal value from state x for an n horizon problem, whereas

$$\begin{aligned} V_n^*(x,O) &= r(x,O) + \sum_{k'\in K}\beta(x,O,k')\sum_{j\in S} q_{ij}^k V_{n-1}^*(k',0,j) \\ &\quad + \sum_{j\in S} p_{ij}^k \int_0^\infty e^{-\alpha t} V_{n-1}^*(k,s+t,j)\,dT_{ij}^k(t), \\ V_n^*(x,R) &= r(x,R) + \sum_{k'\in K}\beta(x,R,k')V_{n-1}^*(k',0,0) \\ &\quad + \int_0^\infty e^{-\alpha t} V_{n-1}^*(k,s+t,0)\,dF^k(t) \end{aligned} \tag{9.21}$$

are the values from state x in n horizons if action O or R is used, respectively, in the first horizon and then an optimal policy in the remaining horizons. The initial conditions are

$$V_0^*(x,O) = V_0^*(x,R) = 0.$$

Let

$$v_n(x) = V_n^*(x, O) - V_n^*(x, R), \;\; v(x) = V^*(x, O) - V^*(x, R),$$

for $x = (k, s, i) \in \Omega$. Then following the standard theory in DTMDPs (see Section 2.4 of this book) we have that

$$\begin{aligned} \lim_{n\to\infty} V_n^*(x, a) &= V^*(x, a), \;\; a = O, R, \\ \lim_{n\to\infty} v_n(x) &= v(x), \end{aligned} \tag{9.22}$$

and the optimal policies can be depicted as

$$f_n^*(x) = O \Longleftrightarrow v_n(x) < 0, \qquad f^*(x) = O \Longleftrightarrow v(x) < 0.$$

So, $(f_N^*, f_{N-1}^*, \ldots, f_0^*)$ is optimal for the discounted expected total cost in N horizons and f^* is optimal for the discounted expected total cost in infinite horizons.

A concept of stochastic order between two distribution functions is needed in the following. For two distribution functions F and G, F is said to be smaller stochastically than G, denoted by $F \preceq G$, if $F(t) \geq G(t)$ for each t (see the book by Muller and Stoyan [97]). The following lemma is Theorem 1.2.8 in [97] and is used in the proof of Theorem 9.4 below.

Lemma 9.5: *For two distribution functions F and G, $F \preceq G$ if and only if*

$$\int_{-\infty}^{\infty} f(t) dF(t) \leq \int_{-\infty}^{\infty} f(t) dG(t)$$

for each nondecreasing function f.

To obtain some properties of the optimal policies, we introduce the following condition.

Condition 9.2: *For each $k \in K$,*

9.2.1. $\sum_{j=m}^{\infty} q_{ij}^k$ is nondecreasing in i for each $m \geq 0$.

9.2.2. $b^k(i), d^k(i), R^k(i, O)$, and $R^k(i, R)$ are all nonnegative and nondecreasing in i.

9.2.3. Both $b^k(i) - d^k(i)$ and $R^k(i, O) - R^k(i, R)$ are nondecreasing in i.

9.2.4. $F^k \preceq T_0^k \preceq T_1^k \preceq T_2^k \preceq T_3^k(t) \preceq \cdots$; that is, T_i^k is stochastically nondecreasing in i and $F^k(\cdot)$ is the smallest one.

9.2.5. $\int_0^{\infty} e^{-\alpha t} \sum_{j\in S} V(t, j) p_{ij}^k dT_{ij}^k(t)$ is nondecreasing in i if $V(t, j)$ is nonnegative and nondecreasing in j for each $t \geq 0$.

As usual, Condition 9.2.1 means that the larger the deterioration degree of the system, the faster the deterioration resulting from the environment state change will be. In Condition 9.2.3, that $b^k(i)-d^k(i)$ is nondecreasing in i indicates that the operation cost increases faster than the replacement cost as the deterioration degree of the system increases, and similarly for $R^k(i,O)-R^k(i,R)$. In fact, Conditions 9.2.1 to 9.2.3 are Condition 9.1 which is often found in the literature for the discrete time model, whereas Condition 9.2.4 is given for the continuous time case here. Condition 9.2.4 means that the sojourning time of the system in a state is nondecreasing as the deterioration increases and the replacement time is smaller than the sojourning time in any state. Condition 9.2.5 is true if $T_{ij}^k(t)$ is absolutely continuous with probability density function $t_{ij}^k(t)$, and $\sum_{j=m}^{\infty} p_{ij}^k t_{ij}^k(t)$ is nondecreasing in i for each $t \geq 0$ and $m \geq 0$, which is similar to Condition 9.2.1. It is easy to see that the latter two conditions are involved in defining the state.

The first main theorem of this section is given as follows.

Theorem 9.4: *Under Condition 9.2, both $V_n^*(k,s,i)$ and $v_n(k,s,i)$ are nondecreasing in i for each $n \geq 0, k \in K, s \geq 0$ and so*

$$V_n^*(k,s,i) = \begin{cases} V_n^*(k,s,i,O), & 0 \leq i < i_n^*(k,s) \\ V_n^*(k,s,i,R), & i \geq i_n^*(k,s), \end{cases} \tag{9.23}$$

where

$$i_n^*(k,s) := \min\{i | v_n(k,s,i) \geq 0\}.$$

Similarly, both $V^*(k,s,i)$ and $v(k,s,i)$ are also nondecreasing in i and

$$V^*(k,s,i) = \begin{cases} V^*(k,s,i,O), & 0 \leq i < i^*(k,s) \\ V^*(k,s,i,R), & i \geq i^*(k,s), \end{cases} \tag{9.24}$$

where

$$i^*(k,s) := \min\{i | v(k,s,i) \geq 0\}.$$

Proof: It is easy to see that $g^{kk'}(s,t)$ is nondecreasing in t, which implies that $\beta((k,s,i),O,k')$ is nondecreasing in i due to Lemma 9.5 and Condition 9.2. Then by using the induction method, it can be shown from Condition 9.2 and Lemma 9.1 that all of $V_n^*(x,O), V_n^*(x,R)$, and $V_n^*(x)$ are nondecreasing in i.

Now, for each $k \in K$ and $s \geq 0$,

$$\begin{aligned} & r(x,O) - r(x,R) \\ = \; & b^k(i) \int_0^{\infty} b^k(s,t) \left[dT_i^k(t) - dF^k(t) \right] \\ & + \left[b^k(i) - d^k(i) \right] \int_0^{\infty} b^k(s,t) dF^k(t) \end{aligned}$$

$$+ R^k(i,O)\int_0^\infty g^k(s,t)\left[dT_i^k(t) - dF^k(t)\right]$$
$$+ \left[R^k(i,O) - R^k(i,R)\right]\int_0^\infty g^k(s,t)dF^k(t) \qquad (9.25)$$

is also nondecreasing in i due to Condition 9.2, Lemma 9.1, and the fact that both $b^k(s,t)$ and $g^k(s,t)$ are nondecreasing in t for each $k \in K, s \geq 0$.

It should be noted that the latter two terms in $V_n^*(x,R)$ of Eq. (9.21) are independent of i. So $v_n(x)$ is nondecreasing in i and thus the former result follows.

The latter result for $V^*(x)$ follows the former result together with Eq. (9.22). Hence, the theorem is true. □

Theorem 9.4 says that there exists a state limit $i^*(k,s)$ just since time s ago for each $k \in K$ and $s \geq 0$ such that if the system enters a state i while the environment is in state k, then the optimal action is to replace the system with a new one if and only if the deterioration degree of the system is over the limit $i^*(k,s)$; that is, $i \geq i^*(k,s)$. Such a policy is an extended control limit policy. So Theorem 9.4 shows that there exist optimal control limit policies for both finite and infinite horizons.

In the next subsection, we discuss a special case of Markov environments where the state limits are independent of the time variable s.

2.3 Markov Environment

In this subsection, we consider that the environment is Markov as follows.

$$G_{kk'}(t) = \psi_{kk'} \cdot G_k(t), \;\; G_k(t) = 1 - e^{-\lambda_k t}, \;\; t \geq 0, \;\; k,k' \in K. \quad (9.26)$$

In this case, it is shown that the variable s in state $x = (k,s,i)$ can be eliminated.

First, we let

$$t_F^k = \int_0^\infty \left[1 - e^{-(\lambda_k+\alpha)t}\right] dF^k(t),$$
$$t_{ij}^k = \int_0^\infty \left[1 - e^{-(\lambda_k+\alpha)t}\right] dT_{ij}^k(t),$$
$$t_i^k = \sum_{j\in S} p_{ij}^k t_{ij}^k, \;\; \alpha_F^k = 1 - t_F^k, \;\; \alpha_{ij}^k = 1 - t_{ij}^k, \;\; \alpha_i^k = 1 - t_i^k, \quad (9.27)$$

where $F^k(t)$ and $T_{ij}^k(t)$ are defined in Subsection 2.1. Furthermore, we let (for $k \in K$ and $i \in S$)

$$r'(k,i,O) = \frac{t_i^k}{\lambda_k + \alpha}\left[b^k(i) + \lambda_k R^k(i,O)\right] e^{-\lambda_k s},$$
$$r'(k,i,R) = \frac{t_F^k}{\lambda_k + \alpha}\left[d^k(i) + \lambda_k R^k(i,R)\right] e^{-\lambda_k s}.$$

Then it can be calculated, due to Eq. (9.17), that the expected cost $r(x,a)$ and the discount factor $\beta(x,a,k')$ for one horizon can be simplified, respectively, as follows.

$$\begin{aligned} r(x,O) &= r'(k,i,O)e^{-\lambda_k s}, \\ r(x,R) &= r'(k,i,R)e^{-\lambda_k s}, \\ \beta(x,O,k') &= \frac{\lambda_k t_i^k}{\lambda_k+\alpha}\psi_{kk'}e^{-\lambda_k s}, \\ \beta(x,R,k') &= \frac{\lambda_k t_F^k}{\lambda_k+\alpha}\psi_{kk'}e^{-\lambda_k s}, \end{aligned} \tag{9.28}$$

where the variable s and other variables i,k,k' are separated. Based on the above equations, it can be shown that $e^{\lambda_k s}V^*(k,s,i)$ is independent of s, and so both $e^{\lambda_k s}V^*(k,s,i,O)$ and $e^{\lambda_k s}V^*(k,s,i,R)$ are also independent of s. Therefore,

$$\begin{aligned} &e^{\lambda_k s}V^*(k,s,i) = V^*(k,0,i), \\ &e^{\lambda_k s}V^*(k,s,i,a) = V^*(k,0,i,a), \quad a=O,R. \end{aligned}$$

This is to say that all of $V^*(k,s,i), V^*(k,s,i,O)$, and $V^*(k,s,i,R)$ are independent of the variable s. Hence, we denote

$$V^*(k,i) := V^*(k,0,i), \quad V^*(k,i,a) := V^*(k,0,i,O), \quad a=O,R$$

and

$$v(k,i) = V^*(k,i,O) - V^*(k,i,R).$$

Then $V^*(k,i)$ is the minimal nonnegative solution of the following optimality equation,

$$V^*(k,i) = \min\{V^*(k,i,O), V^*(k,i,R)\} \tag{9.29}$$

with corresponding

$$\begin{aligned} V^*(k,i,O) &= r'(k,i,O) + \frac{\lambda_k t_i^k}{\lambda_k+\alpha}\sum_{k'\in K}\psi_{kk'}\sum_{j\in S}q_{ij}^k V^*(k',j) \\ &\quad + \sum_{j\in S}p_{ij}^k\alpha_{ij}^k V^*(k,j), \\ V^*(k,i,R) &= r'(k,i,R) + \frac{\lambda_k t_F^k}{\lambda_k+\alpha}\sum_{k'\in K}\psi_{kk'}V^*(k',0) \\ &\quad + \alpha_F^k V^*(k,0). \end{aligned} \tag{9.30}$$

Now, the problem is simplified by eliminating the time variable s. Thus, we can solve for $V^*(k,i)$ only. From the standard results in DTMDPs, Eq. (9.29)

can also be considered as the optimality equation of an adequately defined DTMDP model with state space $S' = \{(k,i) : k \in K, i \in S\}$ and action set $A = \{O, R\}$.

In the case of a Markov environment, Conditions 9.2.4 and 9.2.5 can be replaced, respectively, by the following weaker ones.

9.2.4′. $t_F^k \le t_0^k \le t_1^k \le t_2^k \le \cdots$ for each $k \in K$.
9.2.5′. $\sum_{j=m}^{\infty} p_{ij}^k \alpha_{ij}^k$ is nondecreasing in i for each $k \in K$ and $m \ge 0$.

The following corollary can be proved exactly as that of Theorem 9.4 from the above discussions.

Corollary 9.1: *For the Markov environment case, suppose that Conditions 9.2.1, 9.2.2, 9.2.3, 9.2.4′, and 9.2.5′ hold. Then* $v(k,i) := V^*(k,i,O) - V^*(k,i,R)$ *is nondecreasing in* i *and*

$$V^*(k,i) = \begin{cases} V^*(k,i,O), & i < i^*(k) \\ V^*(k,i,R), & i \ge i^*(k), \end{cases} \tag{9.31}$$

where

$$i^*(k) := \min\{i | v(k,i) \ge 0\}.$$

The above corollary says that the state limit is also independent of the time variable s; that is, $i^*(k,s) = i^*(k)$.

Remark 9.3:

1. *If* $t_F^k \le t_0^k$ *is not true, then it can be shown similarly that Corollary 9.1 holds in* $i \ge I_k$ *with*

$$I_k := \min\{i | t_i^k \ge t_F^k\}$$

and $i^*(k)$ *should be redefined by*

$$i^*(k) := \min\{i \ge I_k | v(k,i) \ge 0\}.$$

In this case, the optimal policy is to operate the system if $I_k \le i < i^*(k)$ *and to replace the system if* $i \ge i^*(k)$*, although it is not known what the optimal action is for* $0 \le i < I_k$.

2. *Due to the expressions of* $r(x,a)$ *given in Eq. (9.28), we can know that both the optimal value and the optimal policies depend on* $T_{ij}^k(t)$ *only through* t_{ij}^k*. This is to say that the model with a Markov environment is a robust model with respect to the distribution function* $T_{ij}^k(t)$ *of the time of state transition for the system. Moreover, if*

$$P_{ij}^k(t) = p_{ij}^k T_i^k(t), \qquad \forall i, j, k,$$

then we can assume that the system itself is also Markov; that is,

$$T_i^k(t) = 1 - e^{-\mu_i^k t}, \quad i \in S, \quad k \in K,$$

where μ_i^k *and* t_i^k *are determined by each other with*

$$t_i^k = \frac{\lambda_k + \alpha}{\lambda_k + \alpha + \mu_i^k}, \quad \mu_i^k = (\lambda_k + \alpha)\frac{1 - t_i^k}{t_i^k}.$$

In Condition 9.2.4′, t_i^k is nondecreasing in i, so

$$\sum_{j=m}^{\infty} p_{ij}^k \alpha_{ij}^k = \sum_{j=m}^{\infty} p_{ij}^k (1 - t_{ij}^k)$$

may not be nondecreasing. The following lemma gives a sufficient condition for it.

Lemma 9.6: *Suppose that* $T_{ij}^k(t) = T_i^k(t)$ *for all* $i, j \in S$, $k \in K$, *and* $m \geq 0$. *Then* $\sum_{j=m}^{\infty} p_{ij}^k \alpha_{ij}^k$ *is nondecreasing in* i *if and only if*

$$\frac{t_{i+1}^k - t_i^k}{1 - t_i^k} \leq \frac{\sum\limits_{j=m}^{\infty} p_{i+1,j}^k - \sum\limits_{j=m}^{\infty} p_{ij}^k}{\sum\limits_{j=m}^{\infty} p_{ij}^k}. \tag{9.32}$$

Proof: From the given condition,

$$t_{ij}^k = t_i^k, \quad \alpha_{ij}^k = \alpha_i^k = 1 - t_i^k, \quad \sum_{j=m}^{\infty} p_{ij}^k \alpha_{ij}^k = (1 - t_i^k) \sum_{j=m}^{\infty} p_{ij}^k.$$

For two nonnegative functions $h(i)$ and $g(i)$, if $h(i)$ is nonincreasing and $g(i)$ is nondecreasing then it is obvious that $h(i)g(i)$ is nondecreasing if and only if

$$\frac{h(i)}{h(i+1)} \leq \frac{g(i+1)}{g(i)} \quad \text{or} \quad \frac{h(i) - h(i+1)}{h(i+1)} \leq \frac{g(i+1) - g(i)}{g(i)},$$

which immediately implies the lemma. □

Eq. (9.32) means that the increasing speed of $\sum_{j=m}^{\infty} p_{ij}^k$ in i for each $m \geq 0$ is larger than or equal to the decreasing speed of $(1 - t_i^k)$.

The optimal policies f_n^* and f^* are characterized by the state limits $i_n^*(k)$ and $i^*(k)$, respectively. We have the following result about the upper bound of these state limits, which is useful for the state reduction problem discussed below.

Lemma 9.7: *Under the conditions given in Corollary 9.1, if* $t_0^k = t_F^k$, *then*

$$i_n^*(k) \le i_0^*(k), \quad i^*(k) \le i_0^*(k),$$

where

$$i_0^*(k) := \min\{i : \Delta r(k,i) \ge 0\},$$
$$\Delta r(k,i) = r'(k,i,O) - r'(k,i,R).$$

Proof: If $t_0^k = t_F^k$ then $\alpha_0^k = \alpha_F^k$. So, from Lemma 9.1 and Theorem 9.4, we have that

$$\begin{aligned} & t_i^k \sum_{j\in S} q_{ij}^k V_n(k',j) - t_F^k V_n(k',0) \\ & \quad \ge \; t_F^k \sum_{j\in S} q_{ij}^k V_n(k',0) - t_F^k V_n(k',0) \\ & \quad = \; 0, \\ & \qquad \sum_{j\in S} p_{ij}^k \alpha_{ij}^k V_n(k,j) - \alpha_F^k V_n(k,0) \\ & \quad \ge \; \sum_{j\in S} p_{0j}^k \alpha_{0j}^k V_n(k,j) - \alpha_F^k V_n(k,0) \\ & \quad = \; \alpha_0^k V_n(k,0) - \alpha_F^k V_n(k,0) \\ & \quad = \; 0. \end{aligned}$$

Hence, we can get that $v_n(k,i) \ge \Delta r(k,i)$, which immediately implies the lemma. □

Condition 9.2.3 is about the cost rate. We now replace it by a new one about the expected total cost in one period.

9.2.3′. *For each* $k \in K$, *both* $b^k(i)t_i^k - d^k(i)t_F^k$ *and* $R^k(i,O)t_i^k - R^k(i,R)t_F^k$ *are nondecreasing in* i.

Here, $b^k(i)t_i^k$ and $d^k(i)t_F^k$ are, respectively, the expected operating and replacement costs in state i when the environment state is k. So the nondecreasingness of $b^k(i)t_i^k - d^k(i)t_F^k$ means that the expected operating cost increases faster than the expected replacement cost as the system's state increases. The nondecreasingness of $R^k(i,O)t_i^k - R^k(i,R)t_F^k$ has a similar meaning.

Theorem 9.5: *Under Conditions 9.2.1, 9.2.2, 9.2.3′, 9.2.4′, and 9.2.5′, for each* $k \in K$ *and* $n \ge 1, v_n(k,i) := V_n^*(k,i,O) - V_n^*(k,i,R)$ *is nondecreasing in* i, *and so*

$$v_n(k,i) < 0 \quad \textit{iff} \quad i < i_n^*(k) := \min\{i : v_n(k,i) \ge 0\}.$$

Moreover, $v(k,i) := V^*(k,i,O) - V^*(k,i,R)$ *is also nondecreasing in* i *and*

$$v(k,i) < 0 \quad \textit{iff} \quad i < i^*(k) := \min\{i : v(k,i) \geq 0\}.$$

Thus, there exist optimal control limit policies for both finite horizons and infinite horizons.

Proof: It should be noted first that under the given conditions,

$$\begin{aligned} & (\lambda_k + \alpha)\Delta r(k,i) \\ = \; & \left[b^k(i)t_i^k - d^k(i)t_F^k\right] + \lambda_k \left[R^k(i,O)t_i^k - R^k(i,R)t_F^k\right] \end{aligned}$$

is nondecreasing in i. Then the theorem can be proved exactly as that of Theorem 9.4. □

The above theorem shows the existence of optimal control limit policies whose state limit $i^*(k)$ depends only on the environment state k. Thus, the Markov environment case is simpler than the semi-Markov environment case.

In the following, we reduce the number of states of the system under the Markov environment (see Eq. (9.26)). First, we suppose that there are $j(k)$ for $k \in K$ such that

$$i_n^*(k) \leq j(k), \quad n \geq 0, k \in K, \tag{9.33}$$

where $i_n^*(k)$ is defined in Theorem 9.5. Due to Theorem 9.5, we have

$$V^*(k,i) = V^*(k,i,R) = r'(k,i,R) + V_0(k), \;\; i \geq j(k), \;\; k \in K, \tag{9.34}$$

where

$$V_0(k) = \frac{\lambda_k t_F^k}{\lambda_k + \alpha} \sum_{k' \in K} \psi_{kk'} V^*(k',0) + \alpha_F^k V^*(k,0).$$

Thus we can get from Eq. (9.30) that for $i \geq 0$,

$$\begin{aligned} V^*(k,i,O) \;\; = \;\; & r'(k,i,O) \\ & + \frac{\lambda_k t_i^k}{\lambda_k + \alpha} \sum_{k' \in K} \psi_{kk'} \\ & \cdot \left\{ \sum_{j=0}^{j(k')-1} q_{ij}^k V^*(k',j) + \sum_{j=j(k')}^{\infty} q_{ij}^k [r'(k',j,R) + V_0(k')] \right\} \\ & + \sum_{j=0}^{j(k)-1} p_{ij}^k \alpha_{ij}^k V^*(k,j) + \sum_{j=j(k)}^{\infty} p_{ij}^k \alpha_{ij}^k \left[r'(k,j,R) + V_0(k)\right] \\ = \;\; & r'(k,i,O) + \frac{\lambda_k t_i^k}{\lambda_k + \alpha} \sum_{k' \in K} \psi_{kk'} \end{aligned}$$

$$
\begin{aligned}
&\cdot \sum_{j=j(k')}^{\infty} q_{ij}^k \left[r'(k',j,R) - r'(k',j(k'),R)\right] \\
&+ \sum_{j=j(k)}^{\infty} p_{ij}^k \alpha_{ij}^k \left[r'(k,j,R) - r'(k,j(k),R))\right] \\
&+ \frac{\lambda_k t_i^k}{\lambda_k + \alpha} \sum_{k' \in K} \psi_{kk'} \\
&\cdot \left[\sum_{j=0}^{j(k')-1} q_{ij}^k V^*(k',j) + \sum_{j=j(k')}^{\infty} q_{ij}^k V^*(k',j(k')) \right] \\
&+ \sum_{j=0}^{j(k)-1} p_{ij}^k \alpha_{ij}^k V^*(k,j) + \sum_{j=j(k)}^{\infty} p_{ij}^k \alpha_{ij}^k V^*(k,j(k)). \qquad (9.35)
\end{aligned}
$$

We define that

$$
\begin{aligned}
\bar{q}_{ij}^{kk'} &= \begin{cases} q_{ij}^k, & j < j(k') \\ \sum\limits_{j=j(k')}^{\infty} q_{ij}^k, & j = j(k'), \end{cases} \\
\bar{p}_{ij}^k &= \begin{cases} p_{ij}^k, & j < j(k) \\ \sum\limits_{j=j(k)}^{\infty} p_{ij}^k, & j = j(k), \end{cases} \\
\bar{T}_{ij}^k(t) &= \begin{cases} T_{ij}^k(t), & j < j(k) \\ \sum\limits_{j=j(k)}^{\infty} p_{ij}^k T_{ij}^k(t) / \bar{p}_{i,j(k)}^k, & j = j(k). \end{cases} \qquad (9.36)
\end{aligned}
$$

Hence,

$$
\sum_{j=j(k)}^{\infty} p_{ij}^k \alpha_{ij}^k = \bar{p}_{i,j(k)}^k \bar{\alpha}_{i,j(k)}^k,
$$

where $\bar{\alpha}_{i,j(k)}^k$ is defined as $\alpha_{i,j(k)}^k$ with $T_{ij}^k(t)$ being replaced by $\bar{T}_{ij}^k(t)$. Let

$$
\begin{aligned}
\bar{b}_k(i) &= b^k(i) + \lambda_k \sum_{k' \in K} \psi_{kk'} \sum_{j=j(k')}^{\infty} q_{ij}^k \frac{t_F^{k'}}{\lambda_{k'} + \alpha} \left[d^{k'}(j) - d^{k'}(j(k'))\right] \\
&\quad + (t_i^k)^{-1} t_F^k \sum_{j=j(k)}^{\infty} p_{ij}^k \alpha_{ij}^k \left[d^k(j) - d^k(j(k))\right], \\
\bar{R}_k(i,O) &= R^k(i,O) + \sum_{k' \in K} \psi_{kk'} \sum_{j=j(k')}^{\infty} q_{ij}^k \frac{t_F^{k'}}{\lambda_{k'} + \alpha} \lambda_{k'}
\end{aligned}
$$

$$\cdot \left[R^{k'}(j,R) - R^{k'}(j(k'),R)\right]$$
$$+ (t_i^k)^{-1} t_F^k \sum_{j=j(k)}^{\infty} p_{ij}^k \alpha_{ij}^k \left[R^k(j,R) - R^k(j(k),R)\right],$$

$$\bar{r}(k,i,O) = \frac{t_i^k}{\lambda_k + \alpha}\left[\bar{b}_k(i) + \lambda_k \bar{R}_k(i,O)\right].$$

It is easy to see that $\bar{r}(k,i,O)$ is still nondecreasing in i for each k under Condition 9.2. Then for $i \geq 0$,

$$V^*(k,i,O) = \bar{r}(k,i,O) + \frac{\lambda_k t_i^k}{\lambda_k + \alpha} \sum_{k' \in K} \psi_{kk'} \sum_{j=0}^{j(k')} \bar{q}_{ij}^{kk'} V^*(k',j)$$
$$+ a \sum_{j=0}^{j(k)} \bar{p}_{ij}^k \bar{\alpha}_{ij}^k V^*(k,j). \tag{9.37}$$

Now, as in the previous section for the discrete time problem, we construct a new replacement model (NRM), similar to the original replacement model (ORM) except that

1. The state set of the system in environment k is $S_k = \{0,1,\ldots,j(k)\}$ for $k \in K$.

2. The parameters p_{ij}^k, $T_{ij}^k(t)$, q_{ij}^k, $b^k(i)$, and $R^k(i,O)$ are replaced by $\bar{p}_{ij}^k$, $\bar{T}_{ij}^k(t)$, $\bar{q}_{ij}^k$, $\bar{b}^k(i)$, and $\bar{R}^k(i,O)$, respectively, which are defined in the above.

3 The system must be replaced in state $j(k)$ during the environment state k (due to Eq. (9.33)).

From the above discussions, we know that the NRM defined above and the ORM are equivalent under the meanings that their optimal values are identical and their optimality equations are equivalent for the discounted criteria in both the finite and infinite horizons. So their optimal policies are identical. The difference between them is that the number of system states is finite for NRM. Certainly, the problem with finite states is simpler than that with infinite states. For example, the computation for the problem with finite states is feasible whereas that for the problem with infinite states should be approximated.

When $j(k) \leq j^*$ for some j^*, we can take the state set of the NRM as $S_k = \{0,1,\ldots,j^*\}$, which is irrespective of k.

At the end of this subsection, we consider two further special cases.

The first is that the system itself is Markov; that is,

$$T_{ij}^k(t) = 1 - e^{-\mu_i^k t}, \;\; F^k(t) = 1 - e^{-\mu_F^k t}, \;\; i,j \in S, \;\; k \in K. \tag{9.38}$$

Then

$$\alpha_{ij}^k = \frac{\mu_i^k}{\lambda_k + \mu_{k,i} + \alpha}, \quad t_{ij}^k = \frac{\lambda_k + \alpha}{\lambda_k + \mu_{k,i} + \alpha}, \quad i, j \in S, \quad k \in K.$$

The second special case is that the environment is a Poisson process with rate λ; that is, the Markov environment (see Eq. (9.26)) with

$$\psi_{k,k+1} = 1, \quad G_k(t) = 1 - e^{-\lambda t}, \quad t \geq 0, \quad k \in K. \tag{9.39}$$

In this case, the influence of the environment on the system is called the Poisson shock in the literature. Moreover, it is assumed that each shock increases the degree of the deterioration of the system with a probability distribution $\{q_j, j \geq 0\}$ as follows.

$$q_{ij}^k = 0 \text{ for } j < i \text{ and } q_{ij}^k = q_{j-i} \text{ for } j \geq i. \tag{9.40}$$

This means that the deterioration of the system can be cumulated from both the system itself and the environment. Furthermore, all $p_{ij}^k, T_{ij}^k(t), b^k(i), d^k(i), R^k(i, O)$, and $R^k(i, R)$ are independent of k and are denoted $p_{ij}, T_{ij}(t)$, and so on, by only deleting k in the original notations. Then t_{ij}^k, t_i^k, α_{ij}^k, α_i^k, t_F^k, α_F^k are also independent of k and are denoted by t_{ij}, t_i, and so on.

Then, it can be shown that $V^*(k, i)$ and therefore both $V^*(k, i, O)$ and $V^*(k, i, R)$ are independent of k. So the state limits $i^*(k) = i^*$ are also independent of k.

2.4 Numerical Example

In this subsection, we give a numerical example where the environment is a Markov process having two states with parameters as follows,

$$(\psi_{kk'}) = \begin{pmatrix} 0.6 & 0.4 \\ 0.5 & 0.5 \end{pmatrix}, \quad \lambda_1 = 0.08, \quad \lambda_2 = 0.1,$$

and the state transition probabilities for the system are

$$\left(p_{ij}^1\right) = \begin{pmatrix} 0.8 & 0.2 & 0 & 0 & 0 \\ 0 & 0.7 & 0.2 & 0.1 & 0 \\ 0 & 0 & 0.7 & 0.2 & 0.1 \\ 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\left(p_{ij}^2\right) = \begin{pmatrix} 0.9 & 0.1 & 0 & 0 & 0 \\ 0 & 0.8 & 0.1 & 0.1 & 0 \\ 0 & 0 & 0.7 & 0.2 & 0.1 \\ 0 & 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and two probability systems caused by the environment changes are

$$\left(q_{ij}^1\right) = \begin{pmatrix} 0.6 & 0.3 & 0.1 & 0 & 0 \\ 0 & 0.5 & 0.3 & 0.2 & 0 \\ 0 & 0 & 0.4 & 0.3 & 0.3 \\ 0 & 0 & 0 & 0.4 & 0.6 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\left(q_{ij}^2\right) = \begin{pmatrix} 0.7 & 0.2 & 0.1 & 0 & 0 \\ 0 & 0.6 & 0.3 & 0.1 & 0 \\ 0 & 0 & 0.5 & 0.3 & 0.2 \\ 0 & 0 & 0 & 0.4 & 0.6 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The cost rate functions are as follows.

$$b^1(i) = 21 + 3i,\ \ d^1(i) = 55 + i;\ \ R^1(i, O) = 50 + i,\ \ R^1(i, R) = 0,$$

$$b^2(i) = 20 + 3i,\ \ d^2(i) = 50 + i;\ \ R^2(i, O) = 45 + i,\ \ R^2(i, R) = 0.$$

Suppose that the continuous discount rate is $\alpha = 0.05$ and

$$(t_F^1, t_0^1, t_1^1, \cdots, t_4^1) = (0.49, 0.82, 0.83, 0.85, 0.86, 0.90),$$

$$(t_F^2, t_0^2, t_1^2, \cdots, t_4^2) = (0.53, 0.80, 0.81, 0.82, 0.84, 0.88).$$

Following 2 of Remark 9.3, this corresponds to

$$(\mu_F^1, \mu_0^1, \mu_1^1, \ldots, \mu_4^1) = (1.00, 0.2107, 0.01966, 0.1694, 0.1563, 0.1067),$$

$$(\mu_F^2, \mu_0^2, \mu_1^2, \ldots, \mu_4^2) = (0.97, 0.2425, 0.2274, 0.2129, 0.1848, 0.1323)$$

for the case of the exponential distribution function of Eq. (9.38).

Thus, for $i = 0, 1, 2, 3, 4$,

$$\begin{aligned} r'(1, i, O) &= \frac{t_i^1}{\lambda_1 + \alpha}(25 + 3.08i), \\ r'(2, i, O) &= \frac{t_i^2}{\lambda_2 + \alpha}(24.5 + 3.1i), \\ r'(1, i, R) &= \frac{t_F^1}{\lambda_1 + \alpha}(55 + i), \\ r'(2, i, R) &= \frac{t_F^2}{\lambda_2 + \alpha}(50 + i). \end{aligned}$$

Now we iteratively compute the optimal values $V_n^*(k, i)$ for the finite horizons by

$$V_{n+1}^*(k, i, O) = r'(k, i, O) + \frac{\lambda_k t_i^k}{\lambda_k + \alpha} \sum_{k' \in K} \psi_{kk'} \sum_{j \in S} q_{ij}^k V_n^*(k', j)$$

$$
\begin{aligned}
& + \sum_{j \in S} p_{ij}^k (1 - t_i^k) V_n^*(k, j), \\
V_{n+1}^*(k, i, R) &= r'(k, i, R) + \frac{\lambda_k t_F^k}{\lambda_k + \alpha} \sum_{k' \in K} \psi_{kk'} V_n^*(k', 0) \\
& + (1 - t_F^k) V_n^*(k, 0), \\
V_{n+1}^*(k, i) &= \min\{V_{n+1}^*(k, i, O), V_{n+1}^*(k, i, R)\}
\end{aligned}
$$

for $n \geq 0$ with $V_0^*(k, i, O) = V_0^*(k, i, R) = 0$ for all k, i.

The numerical results are shown in Table 9.1. When $n = 30$, $|V_{n+1}^*(k, i, a) - V_n^*(k, i, a)| \leq 0.01$ for all k, i, a. So, we take the optimal value $V^*(k, i) = V_{30}^*(k, i)$. Now $v(k, i)$ is shown in the last line of Table 9.1 and thus the optimal state limits for environments 1 and 2 are 3 and 2, respectively. Namely,

$$i^*(1) = 3, \;\; i^*(2) = 2.$$

For this example, the optimal policy is to replace the system if and only if the state of the system reaches or exceeds 3 or 2 when the environment state is 1 or 2, respectively.

3. Notes and References

In the literature, the optimal replacement problems are studied separately on the two types of deterioration: system deterioration and environment deterioration; see, for example, survey papers by Cho and Parlar [21] and Wang [144].

System deterioration is often described by the system's age or a Markov process with multi-state. Yeh [154] considered only the system's deterioration, modeled by a semi-Markov process, but this process is approximated by a cyclic phase-type distribution. Thus, the problem was transformed into a Markov model and an approximating optimal policy was obtained. Zhang and Love [156] considered the case with two repairing ways: one is a non-perfect repair that is performed after the failure of the system, and the other is a perfect repair. They compared whether the perfect repair is performed at fixed intervals or at variable intervals by using Markov chain theory.

Feldman [40] and Thangaraj and Stanly [138] studied optimum replacement for systems subject to shocks. Sheu [127] considered only the second type of deterioration, which is described by age and caused by a Poisson process. He obtained an optimal age replacement policy. This model was generalized in a paper by Sheu and Chang [128] where the Poisson process's intensity function depends on the number of past replacements and the time that has elapsed since the last replacement. Chiang and Yuan [20] considered a multi-state continuous time Markov chain, where the system's state transmits from state n to $n + 1$ or the failed state N. The system is inspected every T period, after which the system takes one of the following three actions: do nothing, repair,

Table 9.1. Computation results for $V_n^*(k,i)$ and $v(k,i)$.

n	$V_n^*(1,0)$	$V_n^*(1,1)$	$V_n^*(1,2)$	$V_n^*(1,3)$	$V_n^*(1,4)$
1	157.69	179.28	203.74	226.51	258.37
2	266.19	304.72	346.55	381.53	418.02
3	345.19	396.87	448.40	488.00	499.67
4	403.81	464.11	514.85	543.99	547.76
5	446.93	509.19	554.99	579.40	583.17
6	477.51	538.75	581.06	604.58	608.34
7	498.69	558.37	598.78	622.04	625.81
8	513.19	571.53	610.90	634.02	637.79
9	523.08	580.42	619.22	642.19	645.96
10	529.81	586.47	624.92	647.76	651.53
15	542.08	597.48	635.33	657.91	661.68
20	543.88	599.09	636.85	659.39	663.16
25	544.14	599.33	637.08	659.61	663.38
30	544.18	599.36	637.11	659.64	663.41
v	–104.15	–52.74	–18.76	3.87	22.58
n	$V_n^*(2,0)$	$V_n^*(2,1)$	$V_n^*(2,2)$	$V_n^*(2,3)$	$V_n^*(2,4)$
1	130.67	149.04	167.83	189.28	216.48
2	238.41	272.99	310.52	347.65	381.75
3	321.57	369.54	419.65	456.20	459.73
4	384.43	441.82	493.91	507.20	510.74
5	431.05	492.07	539.44	544.97	548.51
6	464.57	525.65	568.52	572.05	575.58
7	488.08	547.93	587.46	591.00	594.53
8	504.31	562.77	600.52	604.05	607.59
9	515.43	572.78	609.46	612.99	616.52
10	523.02	579.56	615.56	619.09	622.62
15	536.87	591.89	626.68	630.21	633.75
20	538.91	593.70	628.31	631.84	635.38
25	539.20	593.97	628.55	632.08	635.62
30	539.25	594.00	628.58	632.12	635.65
v	–82.27	–31.05	2.56	21.45	38.16

or replacement. The type of policies they considered is to take the actions "do nothing", "repair", or "replacement" if the state n belongs, respectively, to

$0 \leq n \leq i-1, i \leq n \leq j-1, j \leq n \leq N$ for some integers i and j with $i < j$. They got the optimal values of i^*, j^*, and T^*.

In the literature, there are only a few papers that consider both deterioration types. Satow et al. [130] considered an optimal replacement problem for a cumulative damage model with both deteriorations. It is assumed that the damage is observed only after shocks and the system fails only when the total amount of damage exceeds a failure level K. They check the policy by replacing the system if and only if the damage level exceeds some k, and obtain an optimal level k^*.

On the other hand, the reliability analysis for a Markov system in a Markov environment has been studied in the literature, for example, by Cao [15]. However, there is no study on the corresponding optimal replacement problems.

Section 1 of this chapter is from Hu [64] and Section 2 is from Hu and Yue [75].

Problems

1. Optimal Replacement with Quality Control. Consider a machine producing a product in each day. The product is either good or bad according to whether the machine is good or bad. Suppose that when the machine is good at the beginning of a day then it will be bad with probability q at the beginning of the next day, while the machine remains bad once it is bad, until it is replaced by a good machine. Suppose there needs no time to replace the machine.

Moreover, the manager does not know whether the machine is good or bad. But he can inspect the product to know whether the product is good or bad.

Suppose that the production cost for each product is C, the cost for each inspection is I and the cost for replacing the bad machine is R. Since we do not know the exact situation of the machine, we define the state (denoted by p) of the problem at the beginning of a day as the posterior probability that the machine is good. The objective is to minimize the expected discounted total cost (denoted by $V(p)$) by determining to make the inspection or replacement (assume that the machine is immediately replaced if it is found in bad situation by the inspection).

Set this up as a Markov decision process model, write the optimality equation. Show that $V(p)$ is increasing and concave in p, and study what special structures of the optimal policy can be obtained.

Consider the average criterion for the above problem.

2. Consider a generalized version of the model in the above problem. The problem is exact as that given in problem 1 except that 1) the product produced by the good machine is not necessarily good: the product is good with probability γ and bad with probability $1-\gamma$, and 2) the manager can observe the quality of the product. Set this up as a Markov decision process model, write

the optimality equation, and study what special structures of the optimal policy can be obtained.

3. Consider the second generalized version of the model in Problem 1. The problem is similar as that given in Problem 1 except that 1) the machine has states $1, 2, \cdots, S$, and the machine with state i produces a good product with probability γ_i and a bad product with probability $1 - \gamma_i$, and 2) the manager can observe the quality of the product. Set this up as a Markov decision process model, write the optimality equation, and study what special structures of the optimal policy can be obtained for both the discounted criterion and the average criterion.

4. Consider the third generalized version of the model in Problem 1. The problem is similar as that given in Problem 1 except that the period is not a constant (is one day in Problem 1) but a random variable η with distribution function F and the time to replace the bad machine needs a duration of ξ with distribution function G. Set this up as a Markov decision process model, write the optimality equation, and study what special structures of the optimal policy can be obtained for both the discounted criterion and the average criterion.

5. For the optimal replacement problem with stochastic termination (see Remark 9.2), simplify the structure and the formula of optimal policies.

6. For the optimal replacement problem with semi-Markov process model discussed in Section 9.2, if the environment is Poisson (described at the end of Subsection 9.2.3) then show that $V^*(k, i)$ and therefore $V^*(k, i, O)$ and $V^*(k, i, R)$ are independent of k, and that the state limit $i^*(k) = i^*$ is also independent of k.

Chapter 10

OPTIMAL ALLOCATION IN SEQUENTIAL ONLINE AUCTIONS

In this chapter, we consider a sequential Internet auction system for Web services, where a seller wants to sell a given amount of items by several sequential auctions on the Web and has a reserve price set on the items. We present two such Internet auction cases: one is where the reserve price is private (known only by the seller). The other one is where the reserve is public (known to all). The buyers arrive according to a Poisson process. The usual assumption for auctions is that the buyers value the items independently with uniform distribution functions, and they honestly bid those values for the items. The number of items allocated to each auction is determined at the beginning of the auction, and the number of items auctioned off at each auction is determined by arriving buyers' bids and the reserve price.

Due to the essential randomness in Internet auctions, we present a new realistic model using Markov decision processes (MDPs) for the sequential Internet auction system with reserve price and show that the MDP models are identical for both the private and public reserve price cases. We present an analysis for the optimal allocation problem of the sequential Internet auction system. Based on the optimality equation, we prove the monotone properties of the optimal policy. Finally, numerical results are illustrated and several corresponding problems are discussed.

1. Problem and Model

The problem we face in this chapter is described as follows.

A seller receives shipments of the items every T days, and each shipment contains K identical items. This is a problem of yield management. The seller intends to sell these items by W auctions within T days (it is shown in traditional theory that dividing one auction into several auctions can increase the aggression and the expected revenue of the seller). For simplicity, it is

assumed that the duration t_0 of each auction is the same, so $t_0 = T/W$. The seller has a reserve price v on each item, which means that he will sell an item only when the price for it is not less than the reserve price.

Bidders arrive according to a Poisson process with rate λ, and each bidder is risk-neutral. Moreover, each bidder wishes to purchase at most one item and she has a valuation on each item. This valuation is private and symmetric, that is, each bidder knows her own valuation deterministically, yet only knows other bidders' valuations as random variables which are drawn independently from the same distribution function $F(x)$. We call this type of valuations independent and private valuation, IPV for short. Furthermore, it is assumed that $F(x)$ is the uniform distribution on an interval $[\underline{v}, \overline{v}]$ and that the bidders will bid honestly, for example, when the mechanism of the auction is the first-price sealed-bid for multiple items.

At the beginning of each auction, the seller should determine how many items, say s, will be offered for the auction from those items remaining. Then when the auction is closed, each of the s highest bidders will win an item if his bid is not less than the reserve price, and pays the value of his bid for the item. The total profit of the seller from the sequential auctions is the sum of the profit gained from each auction. The seller's objective is to maximize the total profit.

Suppose that the holding cost per item per time period is h, which is a constant. When s, the amount of items allocated for auction is excessively large, more items may be auctioned off and the future holding costs would be reduced. But the price of auctioned items would be lower and the total revenue decrease. When s is too small, each auctioned item may have a high price, but the future holding costs would increase and the number of items may remain too high, which may also decrease the total revenue. So there is a problem for the seller to choose the optimal amount of items allocated to each auction to maximize total expected revenue.

Because the quantity auctioned off at each auction is random, a number of items offered for each auction may remain. Thus the amount of items remaining at the beginning of each auction may be random, and so we use a finite horizon Markov decision process to model it.

Suppose that $\beta \in (0, 1)$ is a discount factor. The index period n is referred to as the number of remaining auctions, $n = 0, 1, \ldots, W$. The state i at each period denotes the amount of items remaining at the beginning of the period, $i = 0, 1, \ldots, K$, and the action $s = 0, 1, \ldots, i$ expresses the amount of items allocated to an auction from the amount i. When the amount of the items at the beginning of the auction is i and the number of items allocated to the auction is s, then the probability that the number of items at the end of the auction is j is denoted by $p_{ij}(s)$. The reward function $r(i, s)$ is the profit from the items auctioned off minus the total holding costs at this period.

The reserve price set by the seller may be either private, or public (i.e., announced on the Web). For these two cases, we compute the expressions of the state transition probability and the reward function, respectively, in the following two sections.

2. Analysis for Private Reserve Price

In this section, it is assumed that the reserve price is private. The probability for any fixed auction, where we denote by N the number of the arriving bidders, is

$$P\{N = t\} = \frac{(\lambda t_0)^t e^{-\lambda t_0}}{t!}, \quad t \geq 0. \tag{10.1}$$

If $N = t$, let $r_1, \ldots, r_t$ be the bids of arriving bidders. Then with the assumption of an IPV we can know that $r_1, \ldots, r_t$ are independently and identically distributed uniformly on the interval $[\underline{v}, \overline{v}]$, with distribution function $F(x)$. Let $r_0 = 0$ indicate no bidder arrival.

Now, we consider a probability, $p_k(s)$, that only k items are auctioned off when the seller offers s items for the auction.

First we consider the case for $k < s$. It happens if and only if there are only k bids that are not less than v and all other $t - k$ bids are less than v, when t bidders arrive. Because the amount of items auctioned off should be less than or equal to the number of the arriving bidders, we know from the Total Probability Formula that

$$\begin{aligned} p_k(s) &= \sum_{t=k}^{\infty} P\{N = t\} \\ &\quad \cdot P\{\text{exactly } k \text{ events in } \{r_1 \geq v\}, \cdots, \{r_t \geq v\} \text{ occur} \mid N = t\}. \end{aligned}$$

According to the assumption of an IVP, the events $\{r_i \geq v\}, i = 1, 2, \ldots, t$ are independent of each other, with probability

$$\bar{F}(v) := 1 - F(v).$$

We know that

$$\begin{aligned} p_k(s) &= \sum_{t=k}^{\infty} \frac{(\lambda t_0)^t e^{-\lambda t_0}}{t!} \cdot \frac{t! \bar{F}(v)^k F(v)^{t-k}}{k!(t-k)!} \\ &= \frac{e^{-\lambda t_0}[\lambda t_0 \bar{F}(v)]^k}{k!} \sum_{t=k}^{\infty} \frac{[\lambda t_0 F(v)]^{t-k}}{(t-k)!} \\ &= e^{-\lambda t_0 \bar{F}(v)} \frac{[\lambda t_0 \bar{F}(v)]^k}{k!} \\ &= e^{-\lambda t_0 (\overline{v}-v)/(\overline{v}-\underline{v})} \frac{[\lambda t_0 \frac{\overline{v}-v}{\overline{v}-\underline{v}}]^k}{k!}, \quad 0 \leq k < s. \end{aligned} \tag{10.2}$$

Then we consider the final case of $k = s$; that is, all s items are sold at this auction. Similarly, it happens if and only if the number of the arriving bidders is more than the amount of items auctioned off (i.e., $t \geq s$), and among which there are at least s bids that are larger than or equal to v. So the probability that s items are sold out is

$$\begin{aligned} p_s(s) &= \sum_{t=s}^{\infty} P\{N=t\} P\{\text{at least } s \text{ events in} \\ &\qquad \{r_1 \geq v\}, \ldots, \{r_t \geq v\} \text{ occur} \mid N=t\} \\ &= \sum_{t=s}^{\infty} \frac{(\lambda t_0)^t e^{-\lambda t_0}}{t!} \sum_{m=s}^{t} \frac{t! \bar{F}(v)^m F(v)^{t-m}}{m!(t-m)!} \\ &= e^{-\lambda t_0} \sum_{m=s}^{+\infty} \sum_{t=m}^{+\infty} \frac{[\lambda t_0 \bar{F}(v)]^m}{m!} \frac{(\lambda t_0 F(v))^{t-m}}{(t-m)!} \\ &= e^{-\lambda t_0 \bar{F}(v)} \sum_{m=s}^{+\infty} \frac{[\lambda t_0 \bar{F}(v)]^m}{m!} \\ &= 1 - \sum_{k=0}^{s-1} p_k(s). \end{aligned} \tag{10.3}$$

For notational simplicity, we let $\delta = \lambda t_0/(\overline{v} - \underline{v})$, and

$$q_m = e^{-\delta(\overline{v}-v)} \frac{[\delta(\overline{v}-v)]^m}{m!}, \quad m \geq 0, \tag{10.4}$$

$$p_s = \sum_{m=s}^{\infty} q_m = 1 - \sum_{m=0}^{s-1} q_m, \tag{10.5}$$

where q_m is the probability of exactly m bidders arriving whose bids are larger than or equal to v, and p_s is the probability that the number of such bidders is larger than or equal to s, the number of items offered. Then the transition probability is $p_{ij}(s) = p_{i-j}(s)$. So

$$p_{ij}(s) = \begin{cases} q_{i-j}, & i - s < j \leq i \\ p_s, & j = i - s \\ 0, & j < i - s, \text{ or } j > i. \end{cases} \tag{10.6}$$

Now, we consider the reward function $r(i, s)$. We denote by b_k the kth highest bid among all bids, $k \geq 1$. Because it is impossible that the amount of items auctioned off is larger than the number of arriving bidders, we define $b_k = 0$ when $t < k$, whose probability is

$$P\{b_k = 0\} = \sum_{t=0}^{k-1} \frac{1}{t!} (\lambda t_0)^t e^{-\lambda t_0}.$$

If $t \geq k$, it is obvious that $b_k \in [\underline{v}, \overline{v}]$. Using the Total Probability Formula, we have that

$$P(b_k \geq x) = \sum_{t=k}^{\infty} P(b_k \geq x | N = t) P(N = t)$$

for $\underline{v} \leq x \leq \overline{v}$. Similarly to Eq. (10.3), we get that

$$\begin{aligned} P(b_k \geq x) &= \sum_{t=k}^{\infty} \frac{(\lambda t_0)^t e^{-\lambda t_0}}{t!} \sum_{m=k}^{t} \frac{t! \bar{F}(x)^m F(x)^{t-m}}{m!(t-m)!} \\ &= e^{-\lambda t_0 \bar{F}(x)} \sum_{m=k}^{+\infty} \frac{(\lambda t_0 \bar{F}(x))^m}{m!} \\ &= 1 - e^{-\delta(\overline{v}-x)} \sum_{m=0}^{k-1} \frac{[\delta(\overline{v}-x)]^m}{m!}, \quad \underline{v} \leq x \leq \overline{v}, \quad k \geq 1. \end{aligned}$$

So if $k \geq 1$, the kth highest bid b_k is a mixed random variable. b_k has a mass $\sum_{t=0}^{k-1} e^{-\lambda t_0} (\lambda t_0)^t / t!$ at zero and is continuous in the interval $[\underline{v}, \overline{v}]$ with its distribution function as follows.

$$P(b_k \leq x) = \begin{cases} 0, & x < 0 \\ e^{-\lambda t_0} \sum_{m=0}^{k-1} \frac{(\lambda t_0)^m}{m!}, & 0 \leq x < \underline{v} \\ e^{-\delta(\overline{v}-x)} \sum_{m=0}^{k-1} \frac{[\delta(\overline{v}-x)]^m}{m!}, & \underline{v} \leq x < \overline{v} \\ 1, & \overline{v} \leq x. \end{cases} \tag{10.7}$$

Because the seller sets the reserve price v, we only consider the price, called "trade price", at which the item is traded. If $k \geq 1$, then the kth highest trade price, denoted by $\hat{b}_k$, is the kth highest bid b_k when $b_k \geq v$, and does not exist when $b_k < v$. So the expected kth highest trade price is

$$\begin{aligned} E\hat{b}_k &= P\{b_k \geq v\} \int_v^{\overline{v}} x dP(b_k \leq x | b_k \geq v) \\ &= \int_v^{\overline{v}} x dP(b_k \leq x) \\ &= \overline{v} - v P(b_k \leq v) - \int_v^{\overline{v}} P(b_k \leq x) dx. \end{aligned} \tag{10.8}$$

We have from Eq. (10.7) that

$$\int_v^{\overline{v}} P(b_k \leq x) dx = \sum_{m=0}^{k-1} \frac{1}{m!} \int_v^{\overline{v}} e^{-\delta(\overline{v}-x)} [\delta(\overline{v}-x)]^m dx.$$

Let $y = \delta(\overline{v} - x)$. Then $x = \overline{v} - (y/\delta)$, $dx = -(1/\delta)dy$, so the above equation becomes

$$\int_v^{\overline{v}} P(b_k \le x)dx = \frac{1}{\delta}\sum_{m=0}^{k-1}\frac{1}{m!}\int_0^{\delta(\overline{v}-v)} e^{-y}y^m dy.$$

Moreover, for any constant $c > 0$, we can prove by the induction method on the integer $m \ge 1$ that

$$\int_0^c e^{-y}y^m dy = m! - e^{-c}\sum_{l=0}^{m}\frac{m!}{l!}c^l, \;\; m \ge 1.$$

So

$$\begin{aligned}
\int_v^{\overline{v}} P(b_k \le x)dx &= \frac{1}{\delta}\sum_{m=0}^{k-1}[1 - e^{-\delta(\overline{v}-v)}\sum_{l=0}^{m}\frac{[\delta(\overline{v}-v)]^l}{l!}] \\
&= \frac{k}{\delta} - \frac{e^{-\delta(\overline{v}-v)}}{\delta}\sum_{m=0}^{k-1}(k-m)\frac{[\delta(\overline{v}-v)]^m}{m!} \\
&= \frac{k}{\delta} - \frac{1}{\delta}\sum_{m=0}^{k-1}(k-m)q_m, \qquad (10.9)
\end{aligned}$$

which together with Eqs. (10.7) and (10.8) implies that

$$E\hat{b}_k = \overline{v} - \frac{k}{\delta} + \sum_{m=0}^{k-1}(\frac{k-m}{\delta} - v)q_m. \qquad (10.10)$$

So, when s items are offered for some auction, the revenue gained by the seller from this auction (not including the holding cost) is

$$\begin{aligned}
r(s) &= \sum_{k=1}^{s} E\hat{b}_k \\
&= \sum_{k=1}^{s}(\overline{v} - \frac{k}{\delta} + \sum_{m=0}^{k-1}(\frac{k-m}{\delta} - v)q_m) \\
&= s\overline{v} - \frac{s(s+1)}{2\delta} + \sum_{m=0}^{s-1}(\frac{s-m+1}{2\delta} - v)(s-m)q_m. \qquad (10.11)
\end{aligned}$$

Thus, the total expected profit for one auction is $r(i,s) = r(s) - ih$.

We define a random variable ξ_s as the number of auctioned items in one auction when s items are offered. Then for period n with $s_n = s$, the probability distribution of ξ_s is $P\{\xi_s = k\} = q_k$ for $k < s$ and $P\{\xi_s = s\} = p_s$. Certainly, ξ_s depends on the variable s.

Let $V_n(i)$ denote the maximum expected profit when there are n auctions remaining and i items in inventory. Thus, the optimality equation is

$$\begin{aligned} V_n(i) &= \max_{s=0,1,\ldots,i} \{r(s) + \beta \sum_{k=0}^{s-1} q_k V_{n-1}(i-k) + \beta p_s V_{n-1}(i-s)\} - ih \\ &= \max_{s=0,1,\ldots,i} \{r(s) + \beta E V_{n-1}(i-\xi_s)\} - ih, \ \ i \geq 0, \end{aligned} \tag{10.12}$$

with a boundary condition $V_0(i) = 0$ for $i \geq 0$.

The boundary condition $V_0(i) = 0$ implies no value of items at the end of the problem, which happens often in yield management [136]. This assumption is not essential, and can be relaxed as any nonnegative concave function in i ($i \geq 0$). On the other hand, it can be seen easily that $V_n(0) = 0$ for $n = 1, 2, \ldots, W$, which implies that if no item remains, the sequential auction will end, regardless of how many auctions remain, where W is the number of auctions.

3. Analysis for Announced Reserve Price

When the seller announces her reserve price v, the bidders whose intended bid is less than the reserve price v will obviously not bid and leave. The bidders who believe that the value of the item is more than v will be willing to place a bid. We call bidders who arrive and bid "bidding bidders". The following proposition says that the bidding bidders arrive also according to a Poisson process.

Proposition 10.1: *When the reserve price is announced, the bidding bidders arrive according to a Poisson process with rate $\lambda \bar{F}(v)$, and their valuation of the items is drawn independently and identically from a uniform distribution function on the interval $[v, \overline{v}]$.*

Proof: It is well known that the bidding bidders arrive according to a Poisson process with rate $\lambda \bar{F}(v)$. For the other result, let $\hat{r}_i$ be the valuation of ith bidding bidder. Then $\hat{r}_i \in [v, \overline{v}]$, and its distribution function is

$$F_v(x) = P(\hat{r}_i \leq x) = P(r_i \leq x | r_i \geq v) = \frac{x - v}{\overline{v} - v}$$

for $v \leq x \leq \overline{v}$. □

From Proposition 10.1, we can regard the arrival process as a case where the seller's reserve price is zero with the arrival rate $\lambda \bar{F}(v)$. For any fixed auction, if we let N denote the number of the arriving bidding bidders, then the probability distribution of N is obviously given by

$$P\{N = t\} = \frac{1}{t!} (\lambda t_0 \bar{F}(v))^t e^{-\lambda t_0 \bar{F}(v)}, \ \ t \geq 0.$$

Suppose that s items are allocated for the auction for some period. Now, we consider the probability, $p_k(s)$, of k items auctioned off in the auction.

First we consider the case of $k < s$. This means that k items among s items are sold, which happens if and only if there are exactly k bidding bidders and each of them wins an item. So we have that

$$p_k(s) = \frac{1}{k!}(\lambda t_0 \bar{F}(v))^k e^{-\lambda t_0 \bar{F}(v)} = q_k.$$

Then we consider the final case of $k = s$. It means that the number of the bidding bidders is not less than s, so

$$p_s(s) = e^{-\lambda t_0 \bar{F}(v)} \sum_{t=s}^{\infty} \frac{1}{t!}(\lambda t_0 \bar{F}(v))^t = p_s.$$

Therefore, the state transition probability $p_{ij}(s)$ is exactly the state transition probability Eq. (10.6) for the private reserve price case. This indicates that whether to announce the reserve price does not affect the amount of the items auctioned off.

Now, we consider the expected trade price. When the number of the bidding bidders is $N = t(\geq 0)$, their bids $r_1, \ldots, r_t$ are random variables independently and identically distributed uniformly on the interval $[v, \overline{v}]$ with the distribution function $F_v(x)$. Let $r_0 = 0$ denote no bid when no bidding bidders arrive.

Let b_k^o be the kth highest trade price, that is, the kth highest bid. If the number of bidding bidders t is less than k, then $b_k^o = 0$, whose probability is

$$P\{b_k^o = 0\} = \sum_{t=0}^{k-1} e^{-\lambda t_0 \bar{F}(v)} \frac{[\lambda t_0 \bar{F}(v)]^t}{t!} = \sum_{t=0}^{k-1} q_t.$$

If $t \geq k$, then $b_k^o \in [v, \overline{v}]$. Similarly to Eq. (10.3) and Eq. (10.7), we deduce that

$$P(b_k^o \geq x) = \sum_{t=0}^{\infty} P(b_k^o \geq x | N = t) P(N = t) = 1 - \sum_{m=0}^{k-1} q_m,$$

for $v \leq x \leq \overline{v}$ and $k \geq 1$.

So b_k^o is a mixed random variable, and its distribution function is exactly that given in Eq. (10.7), the distribution function of the kth trade price for the private reserve price case. Therefore, the expected revenue function $r(s)$ is also the same as that for the private reserve price case Eq. (10.11).

Therefore, we get the following theorem.

Theorem 10.1: *In the sequential Internet auction, the seller will get the same expected profit whether the reserve price is private or public (announced). Moreover, the maximum expected profit $V_n(i)$ for the seller*

satisfies the optimality equation (10.12) with the corresponding boundary condition and any policy attaining the maximum in Eq. (10.12) is optimal.

4. Monotone Properties

In this section, we study some monotone properties of the optimal policies together with the optimal value. First we have the following properties of the revenue function.

Proposition 10.2: *The revenue function $r(s)$ is strictly concave and increasing with s.*

Proof: Let

$$\Delta r(s) := r(s) - r(s-1), \;\; \Delta^2 r(s) := \Delta r(s) - \Delta r(s-1), \;\; s \geq 1$$

be, respectively, the first-order and the second-order differences of the revenue function. Then, due to Eq. (10.11),

$$\Delta r(s) = \overline{v} - \frac{s}{\delta} \sum_{m=s}^{\infty} q_m - v \sum_{m=0}^{s-1} q_m - \frac{1}{\delta} \sum_{m=0}^{s-1} m q_m$$

and

$$\begin{aligned}
\Delta^2 r(s) &= \overline{v} - \frac{s}{\delta} + \sum_{m=0}^{s-1} \left(\frac{s-m}{\delta} - v \right) q_m \\
&\quad -\overline{v} + \frac{s-1}{\delta} - \sum_{m=0}^{s-2} \left(\frac{s-1-m}{\delta} - v \right) q_m \\
&= -\frac{1}{\delta} - v q_{s-1} + \frac{1}{\delta} \sum_{m=0}^{s-1} q_m \\
&= -\frac{1}{\delta} \left\{ 1 + \delta v q_{s-1} - \sum_{m=0}^{s-1} q_m \right\} \\
&= -\frac{1}{\delta} p_s - v q_{s-1} < 0.
\end{aligned}$$

So $r(s)$ is strictly concave in s. Moreover,

$$\Delta r(\infty) := \lim_{s \to \infty} \Delta r(s) = \overline{v} - v - \frac{1}{\delta} \sum_{m=0}^{\infty} m q_m = 0,$$

which together with the concavity of $r(s)$ implies that $\Delta r(s) > 0$ and so $r(s)$ is strictly increasing with s. □

Let $V_n(i,s)$ be the term in the bracket in optimality equation (10.12); that is,

$$\begin{aligned} V_n(i,s) &= r(s) - ih + \beta \sum_{k=0}^{s-1} q_k V_{n-1}(i-k) + \beta p_s V_{n-1}(i-s) \\ &= r(s) + \beta E V_{n-1}(i-\xi_s) - ih. \end{aligned} \tag{10.13}$$

It represents the expected discounted total profit when i items are in inventory and there remain n auctions with s items offered for the current auction. Moreover, we let

$$\begin{aligned} \Delta_i V_n(i,s) &= V_n(i,s) - V_n(i-1,s), \\ \Delta_s V_n(i,s) &= V_n(i,s) - V_n(i,s-1) \end{aligned}$$

represent the first-order differences of $V_n(i,s)$ in i and s, respectively. Then we define

$$s_n^*(i) = \max\{s | \Delta_s V_n(i,s) \geq 0, s = 1,2,\ldots,i\}. \tag{10.14}$$

Later, we show that the optimal policy is to allocate a quantity of $s_n^*(i)$ for the nth auction when there are i items in inventory. That is, the optimal policy can be characterized by $s_n^*(i)$.

Now, we show that $s_n^*(i)$ is increasing with i and decreasing with n. Our methodology for the proof consists of the following three steps.

Step 1. If $V_{n-1}(i)$ is concave in i, then $V_n(i,s)$ is concave in s and supermodular in (i,s), so $s_n^*(i)$ is increasing with i, where $V_n(i,s)$ is supermodular in (i,s) means that $\Delta_i \Delta_s V_n(i,s) \geq 0$.

Step 2. $V_n(i) = \max_{s=0,1,\ldots,i} V_n(i,s)$ is concave in i.

Step 3. $s_n^*(i)$ is decreasing with n.

We first prove a lemma. For some given constant $\alpha > 0$, let

$$f(i,\lambda) = \sum_{k=i+1}^{\infty} \frac{\lambda^k}{k!} - \alpha \frac{\lambda^i}{i!}.$$

Lemma 10.1: *$f(i,\lambda)$ is nonnegative for all $i = 0,1,2,\ldots$ and $\lambda \geq 0$ if $e^\lambda - 1 - \alpha \geq 0$.*

Proof: It is easy to see that $f'_\lambda(i+1,\lambda) = f(i,\lambda)$ for $i \geq 0$. Now, the given condition that $f(0,\lambda) = e^\lambda - 1 - \alpha \geq 0$ together with the above formula implies that $f'_\lambda(1,\lambda) = f(0,\lambda) \geq 0$. So, $f(1,\lambda)$ is increasing with λ. $f(1,0) = 0$. Hence, $f(1,\lambda) \geq 0$.

Repeating this procedure, we can get the lemma. □

The following proposition is about the upper bound on the maximal profit when one item is added.

Proposition 10.3: $V_n(i) - V_n(i-1) \leq \overline{v}$ *for all* n, i.

Proof: Suppose that π^* is an optimal policy for $V_n(i)$; that is, the expected total profit under the policy π^*, denoted by $V_n(\pi^*, i)$, equals $V_n(i)$. Then

$$\begin{aligned} V_n(i) - V_n(i-1) &= V_n(\pi^*, i) - V_n(i-1) \\ &\leq V_n(\pi^*, i) - V_n(\pi^*, i-1). \end{aligned}$$

We consider a scenario where besides the $i-1$ items, another new item is added. Then the best possible case is that without influencing the original policy, this new item is auctioned off in the first auction at the highest price $\overline{v}$ without incurring any holding cost or discounting of its value. So $V_n(\pi^*, i) - V_n(\pi^*, i-1) \leq \overline{v}$. □

The proposition above is intuitive because each item can be auctioned off with the price at the uppermost valuation $\overline{v}$. The following proposition is in Step 1 of our methodology.

Proposition 10.4: $V_n(i, s)$ *is concave in* s *if* $V_{n-1}(i)$ *is concave in* i.

Proof: It follows from Eq. (10.14) that

$$\begin{aligned} \Delta_s V_n(i, s) &= \Delta r(s) + \beta E[V_{n-1}(i - \xi_s) - V_{n-1}(i - \xi_{s-1})] \\ &= \Delta r(s) - \beta p_s \Delta_i V_{n-1}(i - s + 1). \end{aligned} \tag{10.15}$$

Then

$$\begin{aligned} \Delta_s^2 V_n(i, s) &:= \Delta_s V_n(i, s) - \Delta_s V_n(i, s-1) \\ &= \Delta^2 r(s) - \beta p_s \Delta_i V_{n-1}(i - s + 1) \\ &\quad + \beta p_{s-1} \Delta_i V_{n-1}(i - s + 2) \\ &= \Delta^2 r(s) + \beta p_{s-1} \Delta_i^2 V_{n-1}(i - s + 2) \\ &\quad + \beta q_{s-1} \Delta_i V_{n-1}(i - s + 1). \end{aligned}$$

Due to the given condition in the proposition, it suffices to show that the sum of the first and third terms in the above right-hand side is negative. Because $e^x \geq x + 1$ for each $x \geq 0$, we have that

$$e^{\delta(\overline{v} - v)} \geq \delta(\overline{v} - v) + 1 \geq \delta(\beta\overline{v} - v) + 1.$$

Then by letting $\alpha = \delta(\beta\overline{v} - v)$, and due to Eq. (10.13) and Lemma 10.1 we have

$$\Delta^2 r(s) + \beta q_{s-1} \Delta_i V_{n-1}(i - s + 1)$$

$$\begin{aligned} &\leq -\frac{1}{\delta}p_s - vq_{s-1} + \beta q_{s-1}b \\ &\leq (\beta b - \frac{\alpha}{\delta} - v)q_{s-1} = 0. \end{aligned} \tag{10.16}$$

Hence, the proposition is true. □

With the above proposition, we know that $V_n(i) = V_n(i, s_n^*(i))$ and so $s_n^*(i)$ is the optimal quantity at the nth auction when i items remain if $V_{n-1}(i)$ is concave in i. The next proposition is the other result in Step 1 of our methodology.

Proposition 10.5: *If $V_{n-1}(i)$ is concave in i, then $V_n(i, s)$ is supermodular in (i, s) and so $s_n^*(i)$ is increasing with i.*

Proof: Due to Eq. (10.15), we have that

$$\begin{aligned} \Delta_i\Delta_s V_n(i, s) &= \Delta_s V_n(i, s) - \Delta_s V_n(i-1, s) \\ &= -\beta p_s \Delta_i V_{n-1}(i - s + 1) + \beta p_s \Delta_i V_{n-1}(i - s) \\ &= -\beta p_s \Delta_i^2 V_{n-1}(i - s + 1) \ \geq \ 0 \end{aligned}$$

because $V_{n-1}(i)$ is concave in i. Hence, $V_n(i, s)$ is supermodular. With Theorems 3.1 and 3.2 in Hu and Liu [69], we know that $s_n^*(i)$ is increasing with i. □

For Step 2 of our methodology, we have the following proposition.

Proposition 10.6: *If $V_{n-1}(i)$ is concave in i, then $s_n^*(i) \leq s_n^*(i+1) \leq s_n^*(i) + 1$ for all i.*

Proof: The first inequality follows Proposition 10.5. To show the second inequality it suffices to show that $\Delta_s V_n(i+1, s^*+2) < 0$ where we denote $s^* = s_n^*(i)$ for notational convenience. From Eq. (10.15), we have

$$\begin{aligned} \Delta_s V_n(i+1, s^*+2) &= \Delta r(s^*+2) - \beta p_{s^*+2}\Delta_i V_{n-1}(i - s^*) \\ &= \Delta^2 r(s^*+2) + \Delta r(s^*+1) - \beta(p_{s^*+1} - q_{s^*+1})\Delta_i V_{n-1}(i - s^*) \\ &= \Delta_s V_n(i, s^*+1) + \Delta^2 r(s^*+2) + \beta q_{s^*+1}\Delta_i V_{n-1}(i - s^*). \end{aligned}$$

Then $\Delta_s V_n(i, s^*+1) < 0$ is from the definition of $s_n^*(i) = s^*$, and $\Delta^2 r(s^*+2) + \beta q_{s^*+1}\Delta_i V_{n-1}(i - s^*) \leq 0$ is from Eq. (10.16). □

After preparing the above propositions, we show the following theorem, which is one of the two main results on the monotone properties of $s_n^*(i)$.

Theorem 10.2: *$V_n(i)$ is concave in i for each $n \geq 1$, so $s_n^*(i)$ is increasing in i.*

Proof: It follows from Proposition 10.2 that

$$V_1(i) = \max_{s=0,1,2,\ldots,i} r(s) = r(i), \ \ \forall i \geq 0.$$

So $V_1(i)$ is concave in i.

We suppose that $V_{n-1}(i)$ is concave in i for some $n > 1$. Then from Proposition 10.5 we know that $V_n(i,s)$ is supermodular in (i,s). We show in the following that $V_n(i)$ is concave in i.

First, it should be noted that if $s_n^*(i) = s_n^*(i-1) = s^*$, then

$$\begin{aligned}
\Delta_i V_n(i) &= V_n(i,s^*) - V_n(i-1,s^*) \\
&= -h + \beta E[V_{n-1}(i-\xi_{s^*}) - V_{n-1}(i-1-\xi_{s^*})] \\
&= -h + \beta E \Delta_i V_{n-1}(i-\xi_{s^*}) \\
&= -h + \beta \sum_{k=0}^{s^*-1} q_k \Delta_i V_{n-1}(i-k) \\
&\quad + \beta \sum_{k=s^*}^{\infty} q_k \Delta_i V_{n-1}(i-s^*), \qquad (10.17)
\end{aligned}$$

whereas if $s_n^*(i) = s^*+1$ and $s_n^*(i-1) = s^*$, then

$$\begin{aligned}
\Delta_i V_n(i) &= V_n(i,s^*+1) - V_n(i-1,s^*) \\
&= \Delta r(s^*+1) - h \\
&\quad + \beta E[V_{n-1}(i-\xi_{s^*+1}) - V_{n-1}(i-1-\xi_{s^*})] \\
&= \Delta r(s^*+1) - h + \beta \sum_{k=0}^{s^*} q_k \Delta_i V_{n-1}(i-k). \qquad (10.18)
\end{aligned}$$

Based on Proposition 10.6, we show $\Delta_i^2 V_n(i) \le 0$ by the following four cases, where we denote $s_n^*(i-2) = s^*$ for convenience.

Case 1: $s_n^*(i) = s_n^*(i-1) = s_n^*(i-2) = s^*$. Then,

$$\begin{aligned}
\Delta_i^2 V_n(i) &= \Delta_i V_n(i) - \Delta_i V_n(i-1) \\
&= \beta E[\Delta_i V_{n-1}(i-\xi_{s^*}) - \Delta_i V_{n-1}(i-1-\xi_{s^*})] \\
&= \beta E \Delta_i^2 V_{n-1}(i-\xi_{s^*}) \le 0,
\end{aligned}$$

because $V_{n-1}(i)$ is concave in i.

Case 2: $s_n^*(i) = s_n^*(i-1) = s_n^*(i-2) + 1 = s^*+1$. Then,

$$\begin{aligned}
\Delta_i^2 V_n(i) &= \beta E \Delta_i V_{n-1}(i-\xi_{s^*+1}) - \Delta r(s^*+1) \\
&\quad - \beta E[V_{n-1}(i-1-\xi_{s^*+1}) - V_{n-1}(i-2-\xi_{s^*})] \\
&= \beta E \Delta_i^2 V_{n-1}(i-\xi_{s^*+1}) - \Delta r(s^*+1) \\
&\quad - \beta E[V_{n-1}(i-2-\xi_{s^*+1}) - V_{n-1}(i-2-\xi_{s^*})] \\
&= \beta E \Delta_i^2 V_{n-1}(i-\xi_{s^*+1}) - \Delta_s V_n(i-2, s^*+1) \le 0,
\end{aligned}$$

where the equality follows Eq. (10.15) and the inequality follows a fact that $s^* = s_n^*(i-1)$ and the concavity of $V_{n-1}(i)$.

Case 3: $s_n^*(i) = s_n^*(i-1) + 1 = s_n^*(i-2) + 1 = s^* + 1$. Then,

$$\begin{aligned}
\Delta_i^2 V_n(i) &= \Delta r(s^*+1) + \beta \sum_{k=0}^{s^*} q_k \Delta_i V_{n-1}(i-k) \\
&\quad - \beta \sum_{k=0}^{s^*} q_k \Delta_i V_{n-1}(i-1-k) \\
&\quad - \beta \sum_{k=s^*+1}^{\infty} q_k \Delta_i V_{n-1}(i-1-s^*) \\
&= \Delta r(s^*+1) - \beta p_{s^*+1} \Delta_i V_{n-1}(i-s^*-1) \\
&\quad + \beta p_{s^*+1} \Delta_i V_{n-1}(i-s^*-1) \\
&\quad + \beta \sum_{k=0}^{s^*} q_k \Delta_i^2 V_{n-1}(i-k) - \beta p_{s^*+1} \Delta_i V_{n-1}(i-1-s^*) \\
&= \Delta_s V_{n-1}(i-1, s^*+1) + \beta \sum_{k=0}^{s^*} q_k \Delta_i^2 V_{n-1}(i-k) \ \leq \ 0,
\end{aligned}$$

where the inequality follows the definition of $s_n^*(i-1) = s^*$ and the concavity of $V_{n-1}(i)$ in i.

Case 4: $s_n^*(i) = s_n^*(i-1) + 1 = s_n^*(i-2) + 2 = s^* + 2$. Then,

$$\begin{aligned}
\Delta_i^2 V_n(i) &= \Delta^2 r(s^*+2) + \beta E[V_{n-1}(i - \xi_{s^*+2}) - V_{n-1}(i-1-\xi_{s^*+1})] \\
&\quad - \beta E_\xi[V_{n-1}(i-1-\xi_{s^*+1}) - V_{n-1}(i-2-\xi_{s^*})] \\
&= \Delta_s^2 r(s^*) + \beta \sum_{k=0}^{s^*} q_k \Delta_i^2 V_{n-1}(i-k) + \beta q_{s^*+1} \Delta_i V_{n-1}(i-s^*) \\
&\leq 0,
\end{aligned}$$

which can be proved as that for Proposition 10.4. □

The theorem above says that the greater the number of items held, the greater the number of items that will be allocated to each auction. The following theorem says that the more the horizons remain, the fewer the number of items allocated to each auction will be.

We let

$$i^* = \max\{i \mid \Delta r(i) \geq h\}.$$

Due to Proposition 10.2 we know that $\Delta r(i) \geq h$ if and only if $i \leq i^*$. In general, h is small and i^* may be large, whereas in yield management problems h is zero and i^* is infinite.

Theorem 10.3: *$s_n^*(i)$ is decreasing with n for each $i \leq i^*$.*

Proof: It suffices to show that $V_n(i,s)$ is submodular in (n,s); that is,

$$\Delta_n\Delta_s V_n(i,s) := \Delta_s V_n(i,s) - \Delta_s V_{n-1}(i,s) \le 0$$

for all n, i, s. With Eq. (10.15) we have

$$\begin{aligned} & \Delta_n\Delta_s V_n(i,s) \\ = \; & -\beta p_s \Delta_i V_{n-1}(i-s+1) + \beta p_s \Delta_i V_{n-2}(i-s+1) \\ = \; & -\beta p_s \Delta_n\Delta_i V_{n-1}(i-s+1). \end{aligned} \tag{10.19}$$

Thus it suffices to show that $\Delta_n\Delta_i V_n(i) \ge 0$ for all n, i, which is done by the inductive method in n in the following.

For $n = 1$, $\Delta_n\Delta_i V_1(i) = \Delta_i V_1(i) = \Delta r(i) - h \ge 0$ for $i \le i^*$ by Proposition 10.2.

Suppose that for some $n \ge 2$, $\Delta_n\Delta_i V_{n-1}(i) \ge 0$ for $i \le i^*$. Then from Eq. (10.19), $\Delta_n\Delta_s V_n(i,s) \le 0$ and so $s_n^*(i) \le s_{n-1}^*(i)$ for $i \le i^*$. Now we show that $\Delta_n\Delta_i V_n(i) \ge 0$ for $i \le i^*$. For any given $i \le i^*$, we denote $s_n^* = s_n^*(i-1)$ and $s_{n-1}^* = s_{n-1}^*(i-1)$ for convenience. From the inductive supposition, $s_n^* \le s_{n-1}^*$. We show $\Delta_n\Delta_i V_n(i) \ge 0$ by the following four cases.
Case 1: $s_n^*(i) = s_n^* + 1, s_{n-1}^*(i) = s_{n-1}^* + 1$. In this case, we can get from Eq. (10.18) that

$$\begin{aligned} \Delta_n\Delta_i V_n(i) &= \Delta_i V_n(i) - \Delta_i V_{n-1}(i) \\ &= \Delta r(s_n^*+1) + \beta \sum_{k=0}^{s_n^*} q_k \Delta_i V_{n-1}(i-k) \\ &\quad - \Delta r(s_{n-1}^*+1) - \beta \sum_{k=0}^{s_{n-1}^*} q_k \Delta_i V_{n-2}(i-k) \\ &= -\sum_{k=s_n^*+2}^{s_{n-1}^*+1} \Delta^2 r(k) + \beta \sum_{k=0}^{s_n^*} q_k \Delta_n\Delta_i V_{n-1}(i-k) \\ &\quad - \beta \sum_{k=s_n^*+1}^{s_{n-1}^*} q_k \Delta_i V_{n-2}(i-k) \\ &= \beta \sum_{k=0}^{s_n^*} q_k \Delta_n\Delta_i V_{n-1}(i-k) - \sum_{k=s_n^*+2}^{s_{n-1}^*+1} \Delta^2 r(k) \\ &\quad - \beta \sum_{k=s_n^*+2}^{s_{n-1}^*+1} q_{k-1} \Delta_i V_{n-2}(i-k+1) \\ &= \beta \sum_{k=0}^{s_n^*} q_k \Delta_n\Delta_i V_{n-1}(i-k) \end{aligned}$$

$$- \sum_{k=s_n^*+2}^{s_{n-1}^*+1} \left\{ \Delta^2 r(k) + \beta q_{k-1} \Delta_i V_{n-2}(i-k+1) \right\},$$

which is nonnegative from inductive supposition and Eq. (10.16).
Case 2: $s_n^*(i) = s_n^* + 1, s_{n-1}^*(i) = s_{n-1}^*$. In this case, due to Eq. (10.17) and Eq. (10.18), we can know that

$$\begin{aligned}
&\Delta_n \Delta_i V_n(i) \\
&= \Delta r(s_n^*+1) + \beta \sum_{k=0}^{s_n^*} q_k \Delta_i V_{n-1}(i-k) \\
&\quad - \beta \sum_{k=0}^{s_{n-1}^*} q_k \Delta_i V_{n-2}(i-k) - \beta p_{s_{n-1}^*+1} \Delta_i V_{n-2}(i-s_{n-1}^*) \\
&= \Delta r(s_{n-1}^*+1) - \beta p_{s_{n-1}^*+1} \Delta_i V_{n-1}(i-s_{n-1}^*) \\
&\quad + \beta p_{s_{n-1}^*+1} \Delta_n \Delta_i V_{n-1}(i-s_{n-1}^*) + \Delta r(s_n^*+1) - \Delta r(s_{n-1}^*+1) \\
&\quad + \beta \sum_{k=0}^{s_n^*} q_k \Delta_i V_{n-1}(i-k) - \beta \sum_{k=0}^{s_{n-1}^*} q_k \Delta_i V_{n-2}(i-k) \\
&= \Delta_s V_n(i, s_{n-1}^*+1) + \beta p_{s_{n-1}^*+1} \Delta_n \Delta_i V_{n-1}(i-s_{n-1}^*) \\
&\quad + \Delta r(s_n^*+1) - \Delta r(s_{n-1}^*+1) \\
&\quad + \beta \sum_{k=0}^{s_n^*} q_k \Delta_i V_{n-1}(i-k) - \beta \sum_{k=0}^{s_{n-1}^*} q_k \Delta_i V_{n-2}(i-k).
\end{aligned}$$

In the right-hand side of the equation above, the first term $\Delta_s V_n(i, s_{n-1}^*+1)$ is nonnegative because $s_n^*(i) = s_n^*+1 \le s_{n-1}^*+1$, the second term $\beta p_{s_{n-1}^*+1} \Delta_n \Delta_i V_{n-1}(i-s_{n-1}^*)$ is also nonnegative due to the inductive supposition $\Delta_n \Delta_i V_{n-1}(i-s_{n-1}^*) \ge 0$, and that the remaining term

$$\begin{aligned}
&\Delta r(s_n^*+1) - \Delta r(s_{n-1}^*+1) \\
&+ \beta \sum_{k=0}^{s_n^*} q_k \Delta_i V_{n-1}(i-k) - \beta \sum_{k=0}^{s_{n-1}^*} q_k \Delta_i V_{n-2}(i-k)
\end{aligned}$$

is nonnegative can be proved exactly as that in Case 1. Thus $\Delta_n \Delta_i V_n(i) \ge 0$.
Case 3: $s_n^*(i) = s_n^*, s_{n-1}^*(i) = s_{n-1}^* + 1$. With Eq. (10.17) and Eq. (10.18),

$$\begin{aligned}
&\Delta_n \Delta_i V_n(i) \\
&= \beta \sum_{k=0}^{s_n^*} q_k \Delta_i V_{n-1}(i-k) + \beta p_{s_n^*+1} \Delta_i V_{n-1}(i-s_n^*)
\end{aligned}$$

$$
\begin{aligned}
& - \Delta r(s^*_{n-1}+1) - \beta \sum_{k=0}^{s^*_{n-1}} q_k \Delta_i V_{n-2}(i-k) \\
= \quad & - \Delta r(s^*_n+1) + \beta p_{s^*_n+1} \Delta_i V_{n-1}(i-s^*_n) + \Delta r(s^*_n+1) \\
& - \Delta r(s^*_{n-1}+1) + \beta \sum_{k=0}^{s^*_n} q_k \Delta_i V_{n-1}(i-k) - \beta \sum_{k=0}^{s^*_{n-1}} q_k \Delta_i V_{n-2}(i-k).
\end{aligned}
$$

For the first term in the right-hand side of the equation above, we have from Eq. (10.15) and the definition of $s^*_n(i)$ that

$$
\begin{aligned}
& - \Delta r(s^*_n+1) + \beta p_{s^*_n+1} \Delta_i V_{n-1}(i-s^*_n) + \Delta r(s^*_n+1) - \Delta r(s^*_{n-1}+1) \\
& \qquad = -\Delta_s V_n(i, s^*_n+1) > 0
\end{aligned}
$$

and the second term $\beta \sum_{k=0}^{s^*_n} q_k \Delta_i V_{n-1}(i-k) - \beta \sum_{k=0}^{s^*_{n-1}} q_k \Delta_i V_{n-2}(i-k)$is exactly the term $\Delta_n \Delta_i V_n(i)$ in Case 1 and so is nonnegative. Thus $\Delta_n \Delta_i V_n(i) \geq 0$.

Case 4: $s^*_n(i) = s^*_n, s^*_{n-1}(i) = s^*_{n-1}$. In this case, due to Eq. (10.17),

$$
\begin{aligned}
& \Delta_n \Delta_i V_n(i) \\
= \quad & \beta \sum_{k=0}^{s^*_n} q_k \Delta_i V_{n-1}(i-k) + \beta p_{s^*_n+1} \Delta_i V_{n-1}(i-s^*_n) \\
& - \beta \sum_{k=0}^{s^*_{n-1}} q_k \Delta_i V_{n-2}(i-k) - \beta p_{s^*_{n-1}+1} \Delta_i V_{n-2}(i-s^*_{n-1}) \\
= \quad & - \Delta r(s^*_n+1) + \beta p_{s^*_n+1} \Delta_i V_{n-1}(i-s^*_n) + \Delta r(s^*_{n-1}+1) \\
& - \beta p_{s^*_{n-1}+1} \Delta_i V_{n-1}(i-s^*_{n-1}) + \beta p_{s^*_{n-1}+1} \Delta_n \Delta_i V_{n-1}(i-s^*_{n-1}) \\
& + \Delta r(s^*_n+1) - \Delta r(s^*_{n-1}+1) \\
& + \beta \sum_{k=0}^{s^*_n} q_k \Delta_i V_{n-1}(i-k) - \beta \sum_{k=0}^{s^*_{n-1}} q_k \Delta_i V_{n-2}(i-k).
\end{aligned}
$$

Then, the nonnegativity of $\Delta_n \Delta_i V_n(i)$ can be proved similarly to that for Cases 2 and 3.

By the inductive method, we complete the proof. □

We have gotten two analytical results for the optimal allocation in the sequential Internet auctions. The first one is that there is no difference whether the reserve price is private or public, as shown in Theorem 10.1. The second result is about the monotone properties of the optimal policy as shown in Theorems 10.2 and 10.3. In the next section, we do some numerical analysis for illustrating the model and results obtained above and also analyzing the influence of some parameters on the results.

5. Numerical Results

Beam et al. [10] studied a similar optimal allocation problem to ours. But they assumed that all the items allocated to each auction will be auctioned off. Under this assumption, they formulated the problem as finite horizon deterministic dynamic programming and computed its solution for an example based on eBay. Surely, the assumption is not reasonable.

We consider also the data in Beam et al. [10], which include five consecutive auctions of CD Receivers (Item 2050), and the parameters, estimated in Beam et al. [10], are as follows.

1. The seller opens at most 5 auctions for each shipment, and the duration of each auction is 3.5 days. Initially, the total amount of the items is 35. That is, $W = 5, t_0 = 3.5, K = 35$, and $T = 17.5$.

2. The holding cost per item per auction period is \$0.13, say, $h = 0.13$.

3. The bidders' arrival rate is $\lambda = 13.6$.

4. The bids are distributed uniformly on $[75, 150]$, so $\underline{v} = 75, \overline{v} = 150$.

5. Beam et al. [10] have not considered the reserve price; that is, $v = 0$.

Substituting these parameters for Eq. (10.6) and Eq. (10.11), we get the transition probability $p_{ij}(s)$ and the reward function $r(s)$. Then solving the optimality Eq. (10.12) we can compute the optimality value function and solving Eq. (10.14) we can compute the optimal policy.

The maximal total profit computed here is $V_5(35) = \$5016.67$, which is $\$4454.35$ obtained in Beam et al. [10]. Thus, it is increased by 12.6%, which is a substantial rate for a company's profit. So it is not reasonable to assume as in Beam et al. [10] that all the items can be auctioned off at each auction. For the optimal policy, it is obtained in Beam et al. [10] to offer seven items for each of the five auctions. But here, we should give $s_n^*(i)$ for each n and i, which is shown in Figure 10.1. We can see from the figure that for period 1 (i.e., $n = 1$), the optimal number of offers $s_1^*(i)$ is exactly i. For later periods ($n \geq 2$), the optimal number of offers $s_n^*(i)$ is a step-increasing function with the length of each step n exactly. So we can simply write it as

$$s_n^*(i) = \lfloor (i-1)/n \rfloor + 1,$$

where $\lfloor x \rfloor$ is defined as the largest positive integer not greater than x. From this definition, we suppose that this formula is true for general cases. On the other hand, $s_n^*(i)$ is decreasing with n from Figure 10.1. These show the results from Theorems 10.2 and 10.3.

By the results we present in this chapter, we can further analyze the influence of the parameters on the maximal total profit for the seller.

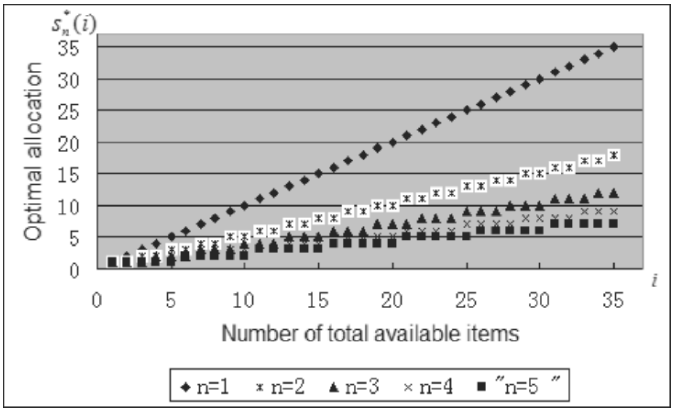

Figure 10.1. Optimal allocation $s_n^*(i)$ versus number of total available items with n.

First, the arrival rate λ means the expected number of arriving buyers per time. From this it is easy to suppose that $V_n(i)$ is increasing with λ. In Figure 10.2, we give $\{V_n(35), n = 1, 2, \ldots, 5\}$ for $\lambda = 3.6, 4.6, 7.6, 13.6$ respectively, where it is shown that $V_n(35)$ is increasing with λ.

Second, we consider the functions of the reserve price v set by the seller. In traditional auctions, the reserve price has two functions. One is to prevent a lower bidder from winning, and the other is to prohibit collusion by bidders. But in this chapter only the former one is considered.

Certainly, $V_5(35)$ is a function of v, which is shown in Figure 10.3.

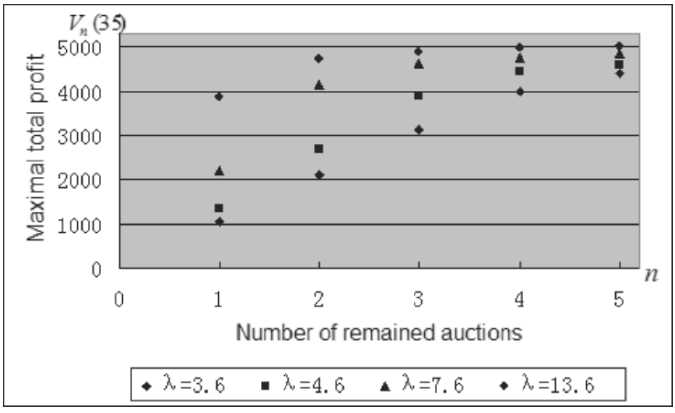

Figure 10.2. Maximal expected total profit $V_n(35)$ versus number of remained auctions with λ.

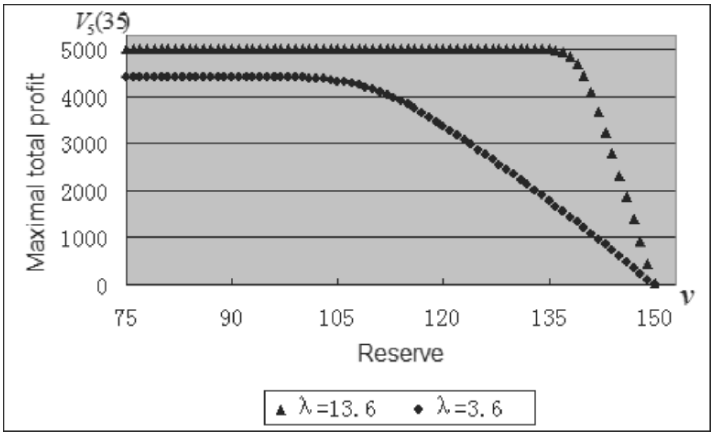

Figure 10.3. Maximal expected total profit $V_5(35)$ versus reserve with λ.

From Figure 10.3, we can see that $V_5(35)$ is a constant 5016.67 for $75 \leq v \leq 117$ and will decrease slowly when $v > 117$ and quickly when $v > 135$. For $v = 135$, $V_3(35) = 5007.46$, which is only a little lower than the maximal value 5016.67. So only a larger reserve price will influence the total expected profit of the seller. This is because the number of arriving bidders is larger. In fact, the mean number of arriving bidders during the whole period is $\lambda \cdot T = 13.6 \times 17.5 = 238$, although there are only 35 items. This can also be seen from Figure 10.3 where the arrival rate $\lambda = 3.6$ is smaller. In this case, a smaller reserve price will influence the total expected profit of the seller. Moreover, if the seller must pay costs for items when she receives shipments, then her maximal total profit will increase with the reserve price v initially and after achieving the maximal value will decrease with v. This reveals a problem associated with the optimal reserve price, which is a topic for further research.

Finally, when $\lambda = 13.6$, we have approximately $V_3(35) = V_5(35)$, which implies that three periods of auctions are enough, and the remaining two auctions can be used to auction other items. But we may need additional conditions to consider the problem of the optimal number W^* of auctions. A dual problem to the optimal number W^* is to consider the optimal quantity K^* of items.

6. Notes and References

In modern information technology, end-to-end quality of service guarantees for multi-media services have become more desirable with the increase in popularity of Internet auctions. This has resulted in several research studies on Internet auctions. For example, based on eBay, Wilcox [150] focused on the impact of the auction experience on bidding behavior. Ockenfels and Roth [99] discussed

bidding strategies such as late bidding and incremental bids in Internet auctions, and pointed out that late bidding is the best response in many environments.

Beam et al. [10] studied an optimal allocation problem in a sequential Internet auction, where a seller holds a given amount of items and wants to sell them in sequential auctions (one after another). At each auction, the buyers arrive according to a Poisson process and bid honestly, whereas the seller determines a quantity of items to be auctioned for each auction. Under a condition that all the items allocated to each auction will be auctioned off, they formulated this case as finite horizon deterministic dynamic programming and computed its solution for an example based on eBay. Vulcano et al. [141] studied a general problem (where they called the problem "yield management") on the auction mechanism under a condition that the number of items auctioned off at each auction will be determined after all buyers' bids are submitted, that being, at the end of the auction. Vulcano et al. [142] studied the same problem for an infinite horizon model. Segev et al. [117] presented a queueing model to approximate multi-item Internet auctions.

In general, the total amount of items offered at an auction is determined before the beginning of the auction, and the number of items auctioned off in an auction is essentially random. In fact, even the number of arriving customers is random in Internet auctions. Thus the conditions introduced in the above papers are not practical. At the same time, the seller often sets a reserve price on the item to ensure his profit and prohibit collusion from customers. However, this is not considered in the above papers. These two aspects are considered in the problem of this chapter.

In the Internet auctions, the buyers generally arrive randomly, one after another. So their arrivals form a stochastic point process. For simplicity, it is often assumed to be a Poisson process, which was checked statistically in Beam et al. [10] and is true in many cases.

This chapter is from Du et al. [30].

Further research may include the optimality problems of the initial quantity K, of the number of auctions W, and the infinite horizon problem when those sequential auctions are repeated again and again.

Problems

1. For the model discussed in this chapter, suppose that all items offered for each auction will be sold out. Set this problem up as a dynamic programming model and write the optimality equation.

2. For the model discussed in this chapter, suppose that the arrival rate of customers in one auction depends on the items offered. Let $\lambda(s)$ be the arrival rate if s items offered. Set the problem up as a Markov decision process model

and write the optimality equation. Whether or not the monotone properties are still true?

3. For the model discussed in this chapter, study further the optimality problems of the initial quantity K and that of the number of auctions W.

4. Suppose that there is only one auction, then how to get the optimal reserve price?

5. Suppose that there are infinite horizons and at the beginning of each auction the seller can order from her supplier with unit cost c. Set this problem up as a Markov decision process model and write the optimality equation.

References

[1] A. Araposthasis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh and S. I. Marcus, "Discrete-time controlled Markov processes with average cost criterion: A survey," *SIAM J. Control Optim.,* Vol. 31, No. 2, pp. 282–334, 1993.

[2] D. P. Berovic and R. B. Vinter, "The application of dynamic programming to optimal inventory control," *IEEE Trans. Autom. Control*, Vol. 49, No. 5, pp. 676–685, 2004.

[3] F. Beutler and K. W. Ross, "Uniformization for semi-Markov decision processes under stationary policies," *J. Appl. Prob.*, Vol. 24, pp. 644–650, 1987.

[4] A. Benveniste, E. Fabre, S. Haar and C. Jard, "Diagnosis of asynchronous discrete-event systems: A net unfolding approach," *IEEE Trans. Autom. Control*, Vol. 48, No. 5, pp. 714–727, 2003.

[5] D. Blackwell, "Discrete dynamic programming," *Ann. Math. Stat.*, Vol. 33, pp. 719–726, 1962.

[6] D. Blackwell, "Discounted dynamic programming," *Ann. Math. Stat.*, Vol. 36, pp. 226–235, 1965.

[7] D. Blackwell, "Positive dynamic programming," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 415–418, 1967, University of California Press, Berkeley.

[8] B. A. Brandin and W. M. Wonham, "Supervisory control of timed discrete event systems," *IEEE Trans. Autom. Control*, Vol. 39, No. 2, pp. 329–342, 1994.

[9] Y. Brave and M. Heymann, "On optimal attraction in discrete event systems," *Inf. Sci.*, Vol. 67, No. 3, pp. 245–276, 1993.

[10] C. Beam, A.J. Segev and G. Shanthikumar, "Electronic negotiation through Internet-based auctions," *CITM Working Paper 96-WP-1019*, 1996.

[11] A. Budhiraja, "An ergodic control problem for constrained diffusion processes: existence of optimal Markov control," *SIAM J. Control Optim.,* Vol. 42, No. 2, pp. 532–558, 2003.

[12] R. Cavazos-Cadena, "Necessary conditions for the optimality equation in average reward Markov decision processes," *Appl. Math. Optim.*, Vol. 19, No. 1, pp. 97–112, 1989.

[13] D. Cansever and R. A. Milito, "Optimal hop-by-hop flow control policies with multiple hetrogeneous transmitters," *Proc. Conf. on Decision and Control,* pp. 1291–1296, 1988.

[14] K. Y. Cai, Y. C. Li and K. Liu, "Optimal and adaptive testing for software reliability assessment," *Inf. Softw. Technol.,* Vol. 46, No. 15, pp. 989–1000, 2004.

[15] J. H. Cao, "Stochastic behavior of a man-machine system operating under changing environment subject to a Markov process with two states," *Microelectron. Reliab.*, Vol. 29, pp. 529–531, 1989.

[16] X. R. Cao, "From perturbation analysis to Markov decision processes and reinforcement learning," *Discrete Event Dynamic Syst.,* Vol. 13, Nos. 1–2, pp. 9–39, 2003.

[17] X. R. Cao, "Semi-Markov decision problems and performance sensitivity analysis," *IEEE Trans. Autom. Control,* Vol. 48, No. 5, pp. 758–769, 2003.

[18] H. S. Chang, R. Givan, and E. K. P. Chong, "Parallel rollout for online solution of partially observable Markov decision processes," *Discrete Event Dynamic Syst. Theor. Appl.,* Vol. 14, No. 3, pp. 309–341, 2004.

[19] Z. Chao and Y. Xi, "Necessary conditions for control consistency in hierarchical control of discrete-event systems," *IEEE Trans. Autom. Control*, Vol. 48, No. 3, pp. 465–468, 2003.

[20] J. Chiang and J. Yuan, "Optimal maintenance policy for a Markovian system under periodic inspection," *Reliab. Eng. Syst. Safety*, Vol. 71, No. 2, pp. 165–172, 2001.

[21] D. Cho and M. Parlar, "A survey of maintenance models for multiunit systems," *Eur. J. Oper. Res.*, Vol. 51, No. 1, pp. 1–23, 1991.

[22] Y. S. Chow and H. Teicher, *Probability Theorem*, Springer-Verlag, New York, 1978.

[23] K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, Springer-Verlag, New York, 1960.

[24] O. L. V. Costa and F. Dufour, "On the poisson equation for piecewise-deterministic Markov processes," *SIAM J. Control Optim.,* Vol. 42, No. 3, pp. 985–1001, 2003.

[25] S. P. Coraluppi and S. I. Marcus, "Mixed risk-neutral/minimax control of discrete time, finite state Markov decision processes," *IEEE Trans. Autom. Control*, Vol. 45, No. 3, pp. 528–532, 2000.

[26] C. G. Cassandras, D. L. Pepyne and Y. Wardi, "Optimal control of a class of hybrid systems," *IEEE Trans. Autom. Control*, Vol. 46, No. 3, pp. 398–415, 2001.

[27] R. Dawen, "Pointwise and uniformly good stationary strategies for dynamic programming models," *Math. Oper. Res.*, Vol. 11, No. 3 pp. 521–535, 1986.

[28] Z. Dong and K. Liu, "The structure of optimal policies in discounted semi-Markov decision processes with unbounded rewards," *Sci. Sinica*, Ser. A, No. 11, pp. 975–985, 1987.

[29] B. T. Doshi, "Continuous time control of Markov processes on an arbitrary state space: discounted rewards," *Ann. Statist.*, Vol. 4, No. 6, pp. 1219–1235, 1976.

[30] L. Du, Q. Hu and W. Yue, "Analysis and evaluation for optimal allocation in sequential Internet auction systems with reserve price," *Dynamics Continuous, Discrete Impulsive Syst. Ser. B,* Vol. 12, No. 4, pp. 617–632, 2005.

[31] D. Duffie, *Dynamic Asset Pricing Theory*, Princeton University Press, Princeton, NJ, 1996.

[32] R. T. Dunn and K. D. Glazebrook, "Discounted multiarmed bandit problems on a collection of machines with varying speeds," *Math. Oper. Res.,* Vol. 29, No. 2, pp. 266-279, 2004.

[33] A. Ephremides and S. Verdu, "Control and optimization methods in communication network problems," *IEEE Trans. Autom. Control,* Vol. 34, No. 9, pp. 930–942, 1989.

[34] A. Federgruen and A. Heching, "Combined pricing and inventory control under uncertainty," *Oper. Res.,* Vol. 47, No. 3, pp. 454–475, 1999.

[35] A. Federgruen, P. J. Schweitzer and H. C. Tijms, "Denumerable undiscounted sem-Markov decision processes with unbounded rewards," *Math. Oper. Res.,* Vol. 8, No. 2, pp. 298–313, 1983.

[36] A. Federgruen and H. M. Tijms, "The optimality equations in average cost denumerable state semi-Markov decision processes: Recurrency conditions and algorithms," *J. Appl. Prob.*, Vol. 15, pp. 356–373, 1978.

[37] E. A. Feinberg, "Controlled Markov processes with arbitrary numerical criteria," *Theory Probab. Appl.*, Vol. 27, pp. 486–503, 1982.

[38] E. A. Feinberg, "Total reward criteria," Ch. 6 in *Handbook of Markov Decision Processes*, Academic Press, New York, 2000.

[39] E. A. Feinberg and A. Shwartz, *Handbook of Markov Decision Processes*, Kluwer Academic, Boston, 2002.

[40] R. M. Feldman, "Optimal replacement with semi-Markov shock models," *J. Appl. Prob.*, Vol. 13, pp. 108–117, 1976.

[41] A. A. Fernandez-Gaucher and S. I. Marcus, "Remarks on the existence of solutions to the average cost optimality equation in Markov decision processes," *Syst. Control Lett.,* Vol. 15, No. 5, pp. 425–432, 1990.

[42] A. Federgruen, A. Hordijk, and H. C. Hijms, "A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices," *J. Appl. Prob.*, Vol. 15, pp. 842–847, 1978.

[43] W. H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions,* Springer-Verlag, New York, 1993.

[44] J. Fu, A. Ray and C. M. Lagoa, "Unconstrained optimal control of regular languages," *Automatica*, Vol. 40, No. 4, pp. 639–646, 2004.

[45] C. H. Golaszewski and P. J. Ramadge, "Control of discrete event processes with forced events," *Proceedings of the 26th IEEE Conference on Decision and Control*, Los Angeles, IEEE, New York, pp. 247–251, 1987.

[46] S. Guo, "Optimal policies problems in discounted Markov decision processes," *Economical Math.*, Vol. 1, No. 1, pp. 109–120, 1984 (in Chinese).

[47] X. P. Guo and W. P. Zhu, "Denumerable-state dontinuous-time Markov decision processes with unbounded transition and reward rates under the discounted criterion," *J. Appl. Prob.*, Vol. 39, No. 2, pp. 233–250, 2002.

[48] X. P. Guo and W. P. Zhu, "Denumerable-state continuous-time Markov decision processes with unbounded cost and transition rates under average criterion," *ANZIAM J.*, Vol. 43, No. 4, pp. 541–557, 2002.

[49] J. M. Harrison, "Discrete dynamic programming," *Ann. Math. Statist.*, Vol. 43, No. 2, pp. 636–644, 1972.

[50] O. Hernandez-Lerma, J. Gonzalez-Hernandez, and R. R. Lopez-Martuez, "Constrained average cost Markov control processes in Borel spaces," *SIAM J. Control Opt.*, Vol. 42, No. 2, pp. 442–468, 2003.

[51] O. Hernandez-Lerma and J. B. Lasserre, "Weak conditions for average optimality in Markov control processes," *Syst. Control Lett.*, Vol. 22, No. 4, pp. 287–291, 1994.

[52] K. Hinderer, *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Lecture Notes in Operations Research and Mathematical Systems, No. 33, Springer-Verlag, New York, 1970.

[53] B. Hou, *Continuous Time Markov Decision Programming with Polynomial Rewards*, MS Thesis, Institute of Applied Mathematics, Academia Sinica, Bejing, 1986.

[54] A. Hordijk, P. J. Schweitzer, and H. C. Tijms, "The asymptotic behavior of the minimal total expected cost for the denumerable state Markov decision model," *J. Appl. Prob.*, Vol. 12, pp. 298–305, 1975.

[55] R. Howard, *Dynamic Programming and Markov Decision Processes*, MIT Press, Cambridge, MA, 1960.

[56] R. Howard, "Semi-Markovian decision processes," *Proc. Intern. Statist. Inst.*, Ottawa, Canada, pp. 625–652, 1963.

[57] Q. Hu, "CTMDPs and its relationship with DTMDPs," *Chinese Sci. Bull.*, Vol. 35, No. 9, pp. 710–714, 1990.

[58] Q. Hu, "Discounted and average Markov decision processes with unbounded rewards: new conditions," *J. Math. Anal. Appl.*, Vol. 171, No. 1, pp. 111–124, 1992.

[59] Q. Hu, "Contiunous time shock Markov decision processes with discounted criterion," *Optimization*, Vol. 25, pp. 271–283, 1992.

[60] Q. Hu, "Nonstationary continuous time Markov decision processes with discounted criterion," *J. Math. Anal. Appl.*, Vol. 180, No. 1, pp. 60–70, 1993.

[61] Q. Hu, "Discrete type shock semi-Markov decision processes with Borel state space," *Optimization*, Vol. 28, pp. 367–382, 1994.

[62] Q. Hu, "Continuous time discounted Markov decision process in a semi-Markov environment with its approximating problem," *Optimization*, Vol. 30, pp. 163–176, 1994.

[63] Q. Hu, "Nonstationary continuous time Markov decision process in a semi-Markov environment with discounted criterion," *J. Math. Anal. Appl.*, Vol. 194, No. 3, pp. 640–659, 1995.

[64] Q. Hu, "The optimal replacement of a Markovian deteriorative system under stochastic shocks," *Microelectron. Reliab.*, Vol. 35, No. 1, pp. 27–31, 1995.

[65] Q. Hu, "Continuous time Markov decision processes with discounted moment criterion," *J. Math. Anal. Appl.*, Vol. 203, No. 1, pp. 1–12, 1996.

[66] Q. Hu, "Nonstationary continuous time Markov decision processes with the expected total rewards criterion," *Optimization*, Vol. 36, pp. 181–189, 1996.

[67] Q. Hu, "Discounted semi-Markov decision process in a semi-Markov environment," *Optimization*, Vol. 39, pp. 367–382, 1997.

[68] Q. Hu, "Average optimality in Markov decision processes with unbounded reward," *OR Trans.*, Vol. 6, No, 1, pp. 1–8, 2002.

[69] Q. Hu and J. Liu, *An Introduction to Markov Decision Processes*, Press of Xidian University, Xian, China, 2000.

[70] Q. Hu and Y. Liu, "Markov decision processes methods in static stability of discrete event systems", *Acady Acta Mathemeticae Applicate Sinica*, Vol. 24, No. 3, pp. 377–383, 2001 (in Chinese series).

[71] Q. Hu, J. Liu, and W. Yue, "Necessary conditions for continuous time Markov decision processes with expected discounted total rewards," *Intern. J. Pure Appl. Math.*, Vol. 7, No. 2, pp. 147–175, 2003.

[72] Q. Hu and J. Wang, "Mixed Markov decision processes in a semi-Markov environment with discounted criterion," *J. Math. Anal. Appl.*, Vol. 219, No. 1, pp. 1–20, 1998.

[73] Q. Hu and J. Wang, "Continuous time Markov decision process with nonuniformly bounded transition rate: Expected total rewards," *Optimization*, Vol. 43, pp. 219–233, 1998.

[74] Q. Hu and C. Xu, "The finiteness of the reward function and the optimal value function in Markov decision processes," *Math. Meth. Oper. Res.*, Vol. 49, No. 2, pp. 255–266, 1999.

[75] Q. Hu and W. Yue, "Optimal replacement of a system according to a semi-Markov decision process in a semi-Markov environment," *Optim. Meth. Softw.*, Vol. 18, No. 2, pp. 181–196, 2003.

[76] Q. Hu and W. Yue, "Analysis for some properties of discrete time Markov decision processes", *Optimization*, Vol. 52, Nos. 4, 5, pp. 495–505, 2003.

[77] Q. Hu and W. Yue,"Two new optimal models for controlling discrete event systems," *J. Industr. Manage. Optim.*, Vol. 1, No. 1, pp. 65–80, 2005.

[78] Q. Hu and W. Yue,"Optimal control for resource allocation in discrete event systems," *J. Industr. Manage. Optim.*, Vol. 2, No. 1, pp. 63–80, 2006.

[79] Q. Hu and W. Yue,"Optimal control for discrete event systems with arbitrary control pattern," *Discrete Contin. Dynam. Syst., Ser. B*, Vol. 6, No. 3, pp. 535–558, 2006.

[80] S. Hu and Q. Hu, "Markov decision programming with generalized unbounded reward function,", *Math. Statist. Appl. Prob.*, Vol. 4, No. 3, pp. 327–335, 1989 (in Chinese).

[81] W. S. Jewell, "Markov-renewal programming 1: Formulation, finite return models," *Oper. Res.*, Vol. 11, 938–948, "2: Infinite return models, example," Vol. 11, No. 6, pp. 949–971, 1963.

[82] P. V. Kakumanu, "Continuous time Markov decision models with applications to optimization problems," *Technical Report 63*, Dept. Oper. Res., Cornell Univ., 1969.

[83] P. V. Kakumanu, "Continuously discounted Markov decision model with countable state and action space," *Ann. Math. Statist.*, Vol. 42, pp. 665–670, 1971.

[84] G. M. Koole, *Stochastic Scheduling and Dynamic Programming*, CWI, Amsterdam, 1995.

[85] A. Kuczura, "Piecewise Markov processes," *SIAM J. Appl. Math.*, Vol. 24, No. 2, pp. 169–181, 1973.

[86] R. Kumar and V. K. Garg, "Optimal supervisory control of discrete event dynamic systems," *SIAM J. Control Optim.*, Vol. 33, No. 2, pp. 419–439, 1995.

[87] M. E. Lewis and M. L. Puterman, "A probabilistic analysis of bias optimality in unichain Markov decision processes," *IEEE Trans. Autom. Control*, Vol. 46, No. 1, pp. 96–100, 2001.

[88] J. Liu, Q. Hu, and J. Wang, "The basic assumption in continuous time Markov decision processes," *Acta Math. Appl. Sinica*, Vol. 27, No. 4, pp. 756–759, 2004.

[89] Y. Li, F. Lin, and Z. H. Lin, "A generalized framework for supervisory control of discrete event systems," *Intern. J. Intell. Control Syst.*, Vol. 2, pp. 139–160, 1998.

[90] Y. Li and W. M. Wonham, "Controllability and observability in the state-feedback control of discrete-event systems," *Proc. of IEEE Conf. Decision & Control*, pp. 203–208, 1989.

[91] Y. Lin and W. M. Wonham, "On observability of discrete-event systems," *Inf. Sci.*, Vol. 44, No. 3, pp. 173–198, 1989.

[92] S. A. Lippman, "On dynamic programming with unbounded rewards," *Manage. Sci.*, Vol. 21, No. 11, pp. 1225–1233, 1975.

[93] Y. Mao and Q. Hu, *Stochastic Processes*, Xidian Press, Xian, China, 1998 (in Chinese).

[94] A. Martin-Lof, "Optimal control of a continuous-time Markov chain with periodic transition probabilities," *Oper. Res.*, Vol. 15, No. 5, pp. 872–881, 1967.

[95] J. Moon and Y. Wardi, "Optimal control of processing times in single-stage discrete event dynamic systems with blocking," *IEEE Trans. Autom. Control*, Vol. 50, No. 6, pp. 880–884, 2005.

[96] J. Mosely and R. A. Humblet, "A class of efficient contention resolution algorithms for multiple access channels," *IEEE Trans. Autom. Control*, Vol. 33, No. 2, pp. 145–151,1985.

[97] A. Muller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*, Wiley, New York, 2002.

[98] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach*, Johns Hopkins University Press, Baltimore, MD, 1981.

[99] A. Ockenfels and A. E. Roth, "The timing bids in Internet auctions: Market design, bidder behavior, and artificial agents," *AI Magazine*, pp. 79–88, 2002.

[100] K. M. Passino and P. J. Antsaklis, "On the optimal control of discrete event systems," *Proc. of IEEE Conf. Decision & Control*, pp. 2713–2718, 1989.

[101] K. M. Passino and K. L. Burgess, "Lagrange stability and boundedness of discrete event systems," *Discrete Event Dynam. Syst. Theor. Appl.*, Vol. 5, No. 4, pp. 383–403, 1995.

[102] N.C. Petruzzi and M. Dada, "Pricing the newsvendor problem: A review with extensions," *Oper. Res.,* Vol. 47, No. 2, pp. 183–194, 1999.

[103] G. Quelle, "Dynamic programming of expectation and variance," *J. Math. Anal. Appl.*, Vol. 55, No. 1, pp. 239–252, 1976.

[104] P. J. Ramadge and W. M. Wonham, "Supervisory control of a class of discrete event systems," *SIAM J. Control Optim.*, Vol. 25, No. 1, pp. 206–230, 1987.

[105] P. J. Ramadge and W. M. Wonham, "Modular feedback logic for discrete event systems," *SIAM J. Control Optim.*, Vol. 25, No. 5, pp. 1202–1218, 1987.

[106] P. J. Ramadge and W. M. Wonham, "The control of discrete event systems," *Proc. of the IEEE*, Vol. 77, No. 1, pp. 81–98, 1989.

[107] A. Ray, J. Fu and C. Lagoa, "Optimal supervisory control of finite state automata," *Int. J. Control*, Vol. 77, No. 12, pp. 1083–1100, 2004.

[108] D. Reetz, "Punctuated and truncated annuities for expanding Markovian decision processes," in *Recent Developments in Markov Decision Processes* (R. Hartley, L. C. Thomas, and D. J. White, Eds.), Academic, New York, pp. 35–56, 1980.

[109] Z. Ren and B.H. Krogh, "Adaptive control of Markov chains with average cost," *IEEE Trans. Autom. Control*, Vol. 46, No. 4, pp. 613–617, 2001.

[110] S. A. Reveliotis and J. Y. Choi, "On the optimality of randomized deadlock avoidance policies," *Discrete Event Dynam. Syst. Theor. Appl.*, Vol. 13, No. 4, pp. 303–320, 2003.

[111] R. K. Ritt and L. I. Sennott, "Optimal stationary policies in general state space Markov decision chains with finite action sets," *Math. Oper. Res.*, Vol. 17, No. 4, pp. 901–909, 1992.

[112] Z. Rosberg and I. S. Gopal, "Optimal hop-by-hop flow control in computer networks," *IEEE Trans. Autom. Control,* Vol. 31, No. 9, pp. 813–822, 1986.

[113] S. M. Ross, "Non-discounted denumerable Markovian decision model," *Ann. Math. Statist.*, Vol. 39, No. 2, pp. 412–423, 1968.

[114] S. M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1971.

[115] S. M. Ross, *Introduction to Stochastic Dynamic Programming*, Academic, New York, 1983.

[116] K. Rudie, S. Lafortune, and F. Lin, "Minimal communication in a distributed discrete-event system," *IEEE Trans. Autom. Control*, Vol. 48, No. 6, pp. 957–975, 2003.

[117] A. Segev, C. Beam, and J. G. Shanthikumar, "Optimal design of Internet-based auctions," *Inf. Technol. Manage.*, Vol. 2, No. 2, pp. 121–163, 2001.

[118] M. Schal, "Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal," *Z. Wahrscheinlichkeitsth*, Vol. 32, pp. 179–196, 1975.

[119] M. Schal, "Average optimality in dynamic programming with general state space," *Math. Oper. Res.*, Vol. 18, No. 1, pp. 163–172, 1993.

[120] P. J. Schweitzer, "Iterative solution of the functional equations of undiscounted Markov renewal programming," *J. Math. Anal. Appl.*, Vol. 34, No. 3, pp. 494–501, 1971.

[121] L. I. Sennott, "Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs," *Oper. Res.*, Vol. 37, No. 4, pp. 626–633, 1989.

[122] L. I. Sennott, "Average cost semi-Markov decision processes and the control of queueing systems," *Prob. Eng. Inform. Sci.*, Vol. 3, pp. 247–272, 1989.

[123] L. I. Sennott, "Another set of conditions for average optimality in Markov control processes," *Syst. Control Lett.*, Vol. 24, No. 2, pp. 147–151, 1995.

[124] L. I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley, New York, 1999.

[125] R. F. Serfozo, "Monotone optimal policies for Markov decision processes," *Math. Program. Study*, Vol. 6, pp. 202–215, 1976.

[126] R. F. Serfozo, "An equivalence between continuous and discrete time Markov decision processes," *Oper. Res.*, Vol. 27, No. 3, pp. 60–70, 1979.

[127] S. Sheu, "A generalized age and block replacement of a system subject to shocks," *Eur. J. Oper. Res.*, Vol. 108, No. 2, pp. 345–362, 1998.

[128] S. Sheu and T. Chang, "Generalized sequential preventive maintenance policy of a system subject to shocks," *Int. J. Syst. Sci.*, Vol. 33, No. 4, pp. 267–276, 2002.

[129] J. Song, "Continuous time Markov decision processes with nonuniformly bounded transition rate family," *Scientia Sinica Series A*, Vol. 11, pp. 1281–1290, 1988.

[130] T. Satow, K. Teramoto, and T. Nakagawa, "Optimal replacement policy for a cumulative damage model with time deterioration," *Math. Comput. Model.*, Vol. 31, pp. 313–319, 2000.

[131] S. E. Shrev and D. P. Bertsekas, "Universally measurable policies in dynamic programming," *Math. Oper. Res.*, Vol. 4, No. 1, pp. 15–30, 1979.

[132] R. Strauch, "Negative dynamic programming," *Ann. Math. Stat.*, Vol. 37, pp. 871–890, 1966.

[133] R. Su and W.M. Wonham, "Supervisor reduction for discrete-event systems," *Discrete Event Dynam. Syst. Theor. Appl.*, Vol. 14, No. 1, pp. 31–53, 2004.

[134] S. Takai, "Supervisory control of partially observed discrete event systems with arbitrary control patterns," *Int. J. Syst. Sci.*, Vol. 31, No. 5, pp. 649–656, 2000.

[135] S. Takai, "Maximizing robustness of supervisors for partially observed discrete event systems," *Automatica*, Vol. 40, pp. 531–535, 2000.

[136] K. T. Talluri and G. J. V. Ryzin, *Revenue Management*, Kluwer Academic, Boston, 2004.

[137] H. M. Taylor, "Markovian sequential replacement processes," *Ann. Math. Statist.*, Vol. 36, pp. 1677–1694, 1965.

[138] V. Thangaraj and A. D. J. Stanly, "Optimum replacement policies for systems subject to shocks," *Optimization*, Vol. 23, pp. 139–154, 1992.

[139] L. C. Thomas, "Connectedness conditions for denumerable state Markovian decision processes," in *Recent Developments in Markov Decision Processes* (R. Hartley, L. C. Thomas, and D. J. White Eds.), Academic, New York, pp. 181–204, 1980.

[140] J. N. Tsitsiklis, "On the control of discrete event dynamic systems," *Math. Control Signals Syst.*, Vol. 2, pp. 95–107, 1989.

[141] G. Vulcano, G. van Ryzin, and C. Maglaras, "Optimal dynamic auctions for revenue management," *Manufact. Serv. Oper. Manage.*, Vol. 4, No. 1, pp. 7–11, 2002.

[142] G. Vulcano, G. van Ryzin, and C. Maglaras, "Optimal dynamic auctions for revenue management," *Manage. Sci.*, Vol. 38, No. 11, pp. 1388–1407, 2002.

[143] J. Wal, "On stationary strategies in countable state total reward Markov decision processes," *Math. Oper. Res.*, Vol. 9, No. 2, pp. 290–300, 1984.

[144] H. Wang, "A survey of maintenance policies of deteriorating systems," *Eur. J. Oper. Res.*, Vol. 139, No. 3, pp. 469–489, 2002.

[145] J. Wessels, "Markov programming by successive approximations with respect to weighted supremum norms," *J. Math. Anal. Appl.*, Vol. 58, No. 2, pp. 326–335, 1977.

[146] C. C. White, "The optimality of isotone strategies for Markov decision problems with utility criterion," *Oper. Res.*, Vol. 20, pp. 261–276, 1980.

[147] D. J. White, "A survey of applications of Markov decision processes," *J. Oper. Res. Soc.*, Vol. 44, No. 11, pp. 1073–1096, 1993.

[148] C. C. White and D. J. White, "Markov decision processes," *Europ. J. Oper. Res.*, Vol. 39, No. 1, pp. 1–16, 1988.

[149] D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1946.

[150] R. T. Wilcox, "Experts and amateurs: the role of experience in Internet auctions," *Market. Lett.*, Vol. 11, pp. 363–374, 2000.

[151] W. M. Wonham and P. J. Ramadge, "On the supremal controllable sublanguage of a given language," *SIAM J. Control Optim.*, Vol. 25, No. 3, pp. 637–659, 1987.

[152] C. Xu and Q. Hu, "The Borel state space semi-Markov decision process with expected total rewards in a semi-Markov environment," *Syst. Sci. Math. Sci.*, Vol. 12, No. 1, pp. 82–91, 1999.

[153] T. Yamasaki and T. Ushio, "Supervisory control of partially observed discrete event systems based on reinforcement learning," *Proc. of the 2003 IEEE International Conference on System, Man, and Cybernetics*, pp. 2956–2961, 2003.

[154] R. Yeh, "Optimal inspection and replacement policies for multi-state deteriorating systems," *Eur. J. Oper. Res.*, Vol. 96, No. 2, pp. 248–259, 1997.

[155] S. H. Zad, R. H. Kwong and W. M. Wonham, "Fault diagnosis in discrete-event systems: Framework and model reduction," *IEEE Trans. Autom. Control*, Vol. 48, No. 7, pp. 1199–1212, 2003.

[156] Z. Zhang and C. Love, "A simple recursive Markov chain model to determine the optimal replacement policies under general repairs," *Comput. Oper. Res.*, Vol. 27, pp. 321–333, 2000.

[157] W. H. Zheng and S. W. Wang, *An Outline of Real and Functional Analysis*, People Education Press, Bejing, 1980 (in Chinese).

Index