The Dissertation Committee for Shun-Pin Hsu
Certifies that this is the approved version of the following dissertation:

**DISCRETE-TIME PARTIALLY OBSERVED MARKOV DECISION PROCESSES: ERGODIC, ADAPTIVE, AND SAFETY CONTROL**

Committee:

Aristotle Arapostathis, Supervisor

Gustavo A. de Veciana

Baxter F. Womack

Takis Konstantopoulos

Klaus R. Bichteler

# DISCRETE-TIME PARTIALLY OBSERVED MARKOV DECISION PROCESSES: ERGODIC, ADAPTIVE, AND SAFETY CONTROL

**by**

**Shun-Pin Hsu, B.S., M.S.**

**DISSERTATION**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**DOCTOR OF PHILOSOPHY**

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2002

Dedicated to my parents *Mr. Ping-Chun Hsu* and *Mrs. Su-Yin Hsieh*, who never fail to back me up.

# Acknowledgments

It is my great honor to mention the following people who helped make the completion of this dissertation possible.

My supervisor Professor, Ari Arapostathis, who is often referred to by students in his class a genius, is not only highly regarded in academic research, but also committed and kind in his guidance and support of students. I give him my deepest appreciation for helping me many times as I sorted out my research and financial challenges of my doctoral program.

The committee members, Dr. Gustavo de Veciana, Dr. Baxter Womack, Dr. Takis Konstantopoulos, and Dr. Klaus Bichteler, have my sincere appreciation for the time they devoted and the constructive comments they provided.

Special thanks go to my research colleague, Dr. Dong-Ming Chuang, whose ideas inspired my work in several sections of this dissertation. Also, I would like to thank Mr. Mark Carpenter for his suggestions and help in revising the dissertation.

My graduate years at the University of Texas at Austin would not have been so enjoyable without the constant encouragement and prayers of my friends in the Austin Taiwanese Presbyterian Church. I would like to extend my heartfelt appreciation to Rev. John Yeh, who continuously confirms my faith by enlightening me with the wisdom of Bible and stories from his life

experience. Also, I want to thank wholeheartedly Elder Dr. Lu's family, Elder Hsu's family, Dr. Chow's family, and all my friends in the Young Adult Fellowship for their friendship, support, and care for me as their brother in the great family of Christ.

Finally, I want to express my profound gratitude to the Heavenly Father and my parents, to whom rightfully belongs the credit for my achievement and whose love taught me the true meaning of infinity.

Shun-Pin Hsu

December 2002

# DISCRETE-TIME PARTIALLY OBSERVED MARKOV DECISION PROCESSES: ERGODIC, ADAPTIVE, AND SAFETY CONTROL

Publication No. _____

Shun-Pin Hsu, Ph.D.
The University of Texas at Austin, 2002

Supervisor: Ari Arapostathis

In this dissertation we study stochastic control problems for systems modelled by discrete-time partially observed Markov decision processes. The issues we consider include ergodic control, adaptive control, and safety control. For ergodic control we propose a new condition that weakens the elegant *interior accessibility* assumption suggested recently. Using the standard procedure to transform the partially observed control problem to its completely observed equivalent, and then applying the *vanishing discount* method, we obtain *Bellman's ergodic optimality equation*, which characterizes the optimal policy. We also provide an example to compare our assumption with those of previous work.

When there are more than one decision maker in the system, we formulate our problem as a stochastic non-cooperative game where each decision maker seeks to minimize his or her own long-run average cost. A special class

of systems with two decision makers and mixed observation structure is considered, and the existence of a Nash equilibrium for the policies is proved.

In the study of adaptive control we extend settings of the ergodic control to the ones where the transition matrix is parameterized by a unknown vector. Motivated by notions of weak ergodicity, we propose a condition on the structure of the transition matrix that results in the ergodic behavior of the underlying controlled process. Under additional hypotheses, we show that the proposed adaptive policy is self-optimizing in appropriate sense.

A new concept designated *safety control* is introduced in our work where the notion of safety is specified in terms of membership in a set called *safe set*. We study the choices of an appropriate policy (called *safe policy*) and an initial state probability distribution such that a safety request, which asks the state probability distribution of the system to lie in a given convex set at each time step, is met. Since the choice of a safe policy is not unique in general, we apply techniques of constrained Markov decision processes to find an *optimal* policy in appropriate sense among the candidates. We also develop an algorithm to find the largest set of initial state probability distributions corresponding to a given safe policy to meet the safety request. The algorithm is proved to terminate in finite steps under reasonable assumptions. Finally we investigate the safety control under partial observations. A machine replacement problem is studied in detail and numerical simulations are presented.

# Table of Contents

**Acknowledgments** **v**

**Abstract** **vii**

**List of Figures** **xii**

**List of Notation and Terminology** **xiii**

**Chapter 1. Introduction** **1**

1.1 Motivation and Overview . . . . . 1
1.2 Review of Previous Work . . . . . 4
1.3 Markov Decision Process Model . . . . . 7
1.3.1 Model Description . . . . . 8
1.3.2 Canonical Sample Space . . . . . 9
1.3.3 Policy Space . . . . . 10
1.3.4 Performance Criterion . . . . . 11
1.4 Partially Observed Markov Decision Process Model . . . . . 11
1.4.1 Model Description . . . . . 11
1.4.2 Canonical Sample Space . . . . . 12
1.4.3 Policy Space . . . . . 13
1.4.4 Model Transformation . . . . . 14
1.5 Partially Observed Stochastic Game . . . . . 16
1.6 Problem Formulation . . . . . 18
1.6.1 Ergodic Control . . . . . 19
1.6.2 Stochastic Game . . . . . 19
1.6.3 Adaptive Control . . . . . 20
1.6.4 Safety Control . . . . . 20

**Chapter 2. Ergodic Control of Partially Observed Markov Decision Processes 22**
2.1 Introduction and Preliminaries . . . . . 22
2.2 Main Result . . . . . 24
2.3 Example and Concluding Remark . . . . . 28

**Chapter 3. Two-Person Zero-Sum Stochastic Games with Mixed Observation Structure 31**
3.1 Introduction and Preliminaries . . . . . 31
3.2 Main Result . . . . . 33

**Chapter 4. Adaptive Control of Partially Observed Markov Decision Processes 38**
4.1 Introduction and Preliminaries . . . . . 38
4.2 Main Result . . . . . 41

**Chapter 5. Safety Control 49**
5.1 Introduction . . . . . 49
5.2 Preliminaries and Notation . . . . . 51
5.3 Safety Enforcing Controller . . . . . 52
5.3.1 General Form . . . . . 52
5.3.2 Special Case . . . . . 53
5.4 Safe Policy . . . . . 58
5.4.1 Linear Programming Formulation . . . . . 59
5.4.2 Feasibility Analysis . . . . . 60
5.5 Supremal Invariant Safe Set . . . . . 64
5.5.1 Searching Algorithm . . . . . 66
5.5.2 Special Case as the size of $P_\pi$ is $2\times2$ . . . . . 69
5.6 Remark on the Assumption . . . . . 74
5.6.1 Spectral Representation of A Matrix Function . . . . . 74
5.6.2 Periodicity . . . . . 76
5.6.3 Scrambling Condition . . . . . 77

**Chapter 6. Safety Control with Partial Observations 79**
6.1 Introduction and Notation . . . . . 79
6.2 Almost Surely Safe Set . . . . . 80
6.3 Almost Surely Safety Enforcing Controller . . . . . 83
6.4 A Machine Replacement Example . . . . . 84

**Chapter 7. Conclusions and Future Work** **92**
7.1 Single and Dual Controllers Ergodic Control . . . . . 92
7.2 Adaptive and Safety Control . . . . . 93

**Appendix A. Proof of Lemma 2.2.1** **97**

**Appendix B. Proofs Regarding Chapter 4** **100**
B.1 Proof of Lemma 4.2.1 . . . . . 100
B.2 Proof of Lemma 4.2.3 . . . . . 100
B.3 Proof of Lemma 4.2.4 . . . . . 101
B.4 Proof of Lemma 4.2.5 . . . . . 102

**Appendix C. Proofs Regarding Chapter 5** **104**
C.1 Proof of Lemma 5.4.1 . . . . . 104
C.2 Proof of Theorem 5.6.2 . . . . . 105

**Appendix D. Set-Valued Function and Measurable Selectors** **108**

**Bibliography** **109**

**Vita** **118**

# List of Figures

6.1 the analytical relation for the priori probability $p$ and posteriori probability $T(p, y, u)$ for various $y$ and $u$ . . . . . 88

6.2 the simulated relation between the stationary policies and their incurred average costs. (The used parameters are $\eta = 0.2$, $q = 0.8$, $\alpha = 0.8$, $p = 0$, $R = 180$, and $C = 150$. The $MatLab^{©}$ random number generator is used and the time horizon is calculated up to 30,000 steps.) . . . . . 90

6.3 the relation between the upper bound $B$ and $T^{-1}(B, 1, 0)$ with the same parameters as those used for Figure 6.2 . . . . . 91

# List of Notation and Terminology

| | |
|---|---|
| w.p. | with probability |
| p.m. | probability measure |
| MDP | Markov Decision Process |
| POMDP | Partially Observed Markov Decision Process |
| POMG | Partially Observed Markov Game |
| $\mathbb{R}$ | set of real numbers |
| $\mathbb{R}^+$ | set of nonnegative numbers |
| $\mathbb{N}$ | set of positive integers |
| $\mathbb{N}_0$ | set of nonnegative integers |
| $\boldsymbol{X}$ | system space |
| $N_x$ | number of elements in $\boldsymbol{X}$ when $\boldsymbol{X}$ is finite |
| $\boldsymbol{Y}$ | observation space |
| $N_y$ | number of elements in $\boldsymbol{Y}$ |
| $\boldsymbol{U}$ | action space |
| $\mathcal{U}(i)$ | set of admissible actions when the system state is $i \in \boldsymbol{X}$ |
| $\boldsymbol{\mathcal{P}}(\boldsymbol{X})$ | space of probability distributions on $\boldsymbol{X}$ |
| $\boldsymbol{\Psi}$ | alternative symbol for $\boldsymbol{\mathcal{P}}(\boldsymbol{X})$ |
| $\boldsymbol{\Psi}_s$ | safety specification |
| $\boldsymbol{\Psi}_{ss}$ | supremal invariant safe set |
| $\Theta$ | compact parameter space |
| $\mathbb{P}^{\pi}_{\psi_0}$ | p.m. induced by the policy $\pi \in \Pi$ and the initial state $\psi_0 \in \boldsymbol{\Psi}$ |
| $\mathbb{E}^{\pi}_{\psi_0}$ | expectation operator corresponding to $\mathbb{P}^{\pi}_{\psi_0}$ |

| | |
|---|---|
| $\Pi$ | the set of admissible policies |
| $:=$ | equality by definition |
| $c$ | one-stage cost function |
| $\mathbf{1}$ | column vector of 1's |
| $I$ | identity matrix |
| $I_n$ | identity matrix with size $n \times n$ |
| $e^i$ | $i^{th}$ row of the identity matrix with size $N_x$ |
| $A_{\cdot j}$ | $j^{th}$ column of a matrix $A$ |
| $A_{i\cdot}$ | $i^{th}$ row of a matrix $A$ |
| $\mathbf{1}_{\{\cdot\}}$ | indicator function |
| $\{X_t\}_{t=0}^{\infty}$ | system state process |
| $\{Y_t\}_{t=0}^{\infty}$ | observation process |
| $\{U_t\}_{t=0}^{\infty}$ | action process |
| $\{\psi_t\}_{t=0}^{\infty}$ | information state process |
| $\{\hat{\theta}_t\}_{t=0}^{\infty}$ | sequence of estimates of $\theta$ |
| $\{\hat{\psi}_t\}_{t=0}^{\infty}$ | sequence of estimates of the information state process |

# Chapter 1

# Introduction

## 1.1 Motivation and Overview

The concept of Markov chains originates from the ingenious approach by Andrei Markov, a Russian Mathematician in the early twentieth century, to estimate the number of vowels and consonants in the poems of Pushkin. Markov derived what we term *conditional probability* that a certain character succeeds another character. After that the starting letter of a poem was enough to do the estimation. His idea of simplifying the estimation task is extended to the notion of conditional independence, called *Markovian*, in the time domain. Roughly speaking, if a system evolves along a discrete time horizon by taking a value out of a set of possible values as its status at each time step, then the system is said to generate a stochastic process. In particular, the stochastic process is called a Markov chain if it has the Markovian property. That is, if there exist some transition rules between statuses such that if the current status of the system is known, then the evolution to the next status of the system does not depend on the past history of the system status. On the other hand, if the evolution of the system can be altered by some other control, we say the process is a controlled Markov chain or Markov decision process. Thus the model suggested by a Markov decision process can be described as follows. At any time step, the system state is observed. A control is given to

decide which transition rule is to govern the evolution to the next step. A cost is incurred according to the current state and decision. As time moves on, the whole procedure repeats. The purpose of modelling the system is then to find the best policy to regulate the actions corresponding to the given criteria (minimizing the accumulated cost, for example).

Due to the model's simplification and generalization, a good number of applications to systems subject to random effects and with controllable natures have been reported in the past decades. Specialists in applied mathematics, engineering, operations research, and economics have been using this model as the foundation for various optimal solutions to many problems in their fields. The main research topics in Markov decision processes include existence theory, solution algorithms, and phenomenon modelling. In our work we focus on the exploration of conditions under which the existence of the optimal policy is ensured and its features are characterized. We also introduce and study the notion of safety control of stochastic discrete event systems modelled by Markov decision processes. For non-stochastic discrete event systems modelled by state machine or automata, safety is specified as a set of forbidden states that imposes bounds on the set of states allowed to be visited. This notion is generalized to the setting of stochastic systems in our work.

The content of the dissertation is organized as follows. In Chapter 1 the technical background is provided. In Chapter 2 the ergodic control of discrete-time partially observed Markov decision processes is studied. Here the ergodic cost, also called long-run average cost, is used as a performance index for the model running on the infinite time horizon; the partial observation, a

generalized version of the complete observation, means that the system state at each time step is not directly observable. Thus, a Bayesian type of estimation of the true state is necessary before each decision is made. We propose a very weak assumption to ensure the existence of an optimal policy and present a dynamic programming formulation to characterize that policy. An example is analyzed to compare our approach with those of previous studies.

We investigate in Chapter 3 a stochastic system with more than one controller. We assume that every controller pursues his or her own benefit and does not cooperate with others. A special class of systems modelled by a two-player zero-sum stochastic game with mixed observation feature is considered. Here by mixed observation we mean there exists a state that is complete observable and other state(s) partially observable. The existence of a *Nash equilibrium*, also called *saddle point equilibrium* under this setting, of the controllers' policies is proved.

In Chapter 4 we study the adaptive control. This subject arises naturally when the transition matrix of the Markov process is not completely known but depends on some unknown parameter. Supposing that a sequence of estimates converging in some appropriate sense to the true parameter is available, we develop a parameterized policy that makes use of these estimates. Under a set of conditions, the policy is proved to be self-optimizing in the sense that it achieves the same asymptotic average performance as if the true parameter were known. Among the set of conditions, we are particularly interested in the one which asks the estimation error between the true and estimated information state process to converge to zero as time goes to infinity.

Utilizing the notion of weak ergodicity, we impose a structural assumption on the transition matrix along with other hypotheses to imply this condition.

The concept of safety control is introduced in Chapter 5. We start our investigation with a system modelled by a completely observable Markov process with finite state and action spaces. Given a safety specification in the form of a convex set in which the state probability distribution of the system must lie at each time step, we study the choice of an appropriate policy and initial state probability distribution such that the system meets the safety requirement. That is, the system's state probability distribution remains in the convex set of safety specification at each time step. In Chapter 6 the study of safety control is extended to a system for which only partial observation is available. The safety specification is now defined as a convex set in which the information state of the system must lie with probability 1 at each time step. A machine replacement problem is studied in detail and some numerical simulations are implemented to contrast safety control with traditional optimal control. Finally some concluding remarks and future research possibilities are given in Chapter 7.

## 1.2 Review of Previous Work

In this section we briefly review the historical development of Markov decision processes (MDPs) and related research activities. The initiation of the MDP dated back to the late 1940s when the sequential decision problem was considered in [73]. In the 1950s the stochastic dynamic programming formulation was established in [8] based on the heuristic idea of *minimal cost*

*to go*. Various problems with basic settings (finite state and action space on the finite time horizon) were solved in [9].

In the 1960s progress was made due to the invention of algorithms of the policy iteration (see [42]) and value iteration (see [74]), which are still popular for deriving an optimal policy via stochastic dynamic programming. The linear programming interpretation of optimality was discovered in [52, 72]. Another achievement is the expansion of the time horizon from finite to infinite. The discounted cost (DC) criterion was considered and used as a platform to study the average cost (AC) criterion via Tauberian theory (see Hardy and Littlewood [33, Appendix H]) and the vanishing discount argument (see [12, 20]). As a result, the average cost optimality equation (ACOE), a stochastic dynamic programming equation characterizing the optimal policy for the AC criterion, was obtained.

A great amount of effort was made in the 1970s to generalize the setting of the model. In the study of a system with countable state space and arbitrary action space, many assumptions, which extend various forms of recurrence conditions and result in the existence of a unique invariant distribution of the underlying controlled Markov chain, were used to yield sufficient conditions for the existence of solutions to the ACOE (see [25, 26]). For a Borel state space with AC criterion, a methodology based on Ascoli's theory and vanishing discount argument was used in [65, 39, 11]. The partially observed Markov decision process model (POMDP) was also introduced around this era. It was treated as a model equivalent to the one with complete observation in a Borel state space. The Bayesian estimation scheme and the idea of separated policies

[48, Chapter 6] both play important roles in the theoretic development of the solution. It was proved in [6] that the value function of the discounted cost optimality equation (DCOE) for the POMDP is concave. In considering the existence of solutions to the ACOE for the POMDP, this property helps ease the challenge in the equicontinuity argument required to apply Ascoli's theory via the DCOE (see [30]). Several important conditions implying the existence of the ACOE were reported in [61]. An elegant assumption, which can be justified easily, was proposed in [19].

The study of stochastic games begins with the seminal paper of Shapley [71] and parallels the development of the MDP due to the similarity in their mathematical formulations. However, since the MDP can be viewed as a special case of the stochastic game that there exists only one controller, it is expected that problems in stochastic games are more challenging than those in MDPs. Ky Fan's [23] and Glicksberg's [35] fixed point theorems provide theoretic tools to prove the existence of a Nash equilibrium, which is characterized by an appropriate dynamic programming equation. Because of technical challenges, only limited results are available for systems with different settings (e.g., a system with countable state space running on the infinite time horizon [26], and a system with general state space running on the finite time horizon [14]). A system with slightly different settings can be seen in [54]. An important class of game models is the two-controller zero-sum game, which is mathematically much more related to the MDP model even though it seems conceptually closer to the game model. Problems in very general settings are solved in [43, 47, 55, 56, 57, 58]. In this dissertation we study a special class of games for which there exists a state that is completely observable but for

whose other state(s) only partial information is available.

Research in the adaptive control of various stochastic systems has been active for years (e.g., see [45, 46, 48]). In particular, a system modelled by the MDP is studied in [16, 36, 38]. Only little work, however, has gone into the POMDP model. Initial efforts were made in this topic in [5], followed by systematic work in [31], which proposed a methodology based on a set of assumptions. An example of binary machine replacement that satisfies these assumptions was reported in [28, 32]. Several justifiable conditions implying these assumptions were provided in [17]. Other special cases were treated in [21].

The exploration of the safety control of stochastic discrete event systems began in [4]. The concept derived from a natural generalization of the notion of safety in non-stochastic discrete event systems (e.g., see [63]). A research topic related to the safety control of stochastic systems is the traditional constrained MDP (see [1, 41] among others). A basic result was proposed in [4, 3], and more research is ongoing.

## 1.3 Markov Decision Process Model

In this section we present the technical background of the Markov decision process model in detail. First we list symbols and terminology that will be used throughout the dissertation:

- $\mathbb{R}$: the set of real numbers;
- $\mathbb{R}^+$: the set of nonnegative real numbers;

- $\mathbb{N}_0$: the set of nonnegative integers;
- $\mathbb{N}$: the set of positive integers;
- $\boldsymbol{\mathcal{B}}(\boldsymbol{V})$: the Borel $\sigma$-algebra of a given topological space $\boldsymbol{V}$;
- $\boldsymbol{\mathcal{P}}(\boldsymbol{X})$: the set of all probability measures on a Borel space $\boldsymbol{X}$ endowed with the topology of weak convergence (see [59]).

### 1.3.1 Model Description

A discrete time Markov decision process is described by a five-tuple $(\boldsymbol{X}, \boldsymbol{U}, \mathcal{U}, P, c)$ where

1. $\boldsymbol{X}$ is a Borel space, called *state space* with elements referred to as *states*;
2. $\boldsymbol{U}$ is a Borel space, called *control* (or *action*) *space*;
3. $\mathcal{U}$ is a set-valued map with nonempty compact values; $\mathcal{U}(x)$ denotes the set of feasible (or admissible) actions when the system is in state $x \in \boldsymbol{X}$ and has the property

$$\mathbb{K} := \mathit{Graph}(\mathcal{U}) = \{(x,u) : x \in \boldsymbol{X}, u \in \mathcal{U}(x)\} \in \boldsymbol{\mathcal{B}}(\boldsymbol{X} \times \boldsymbol{U})\,;$$

4. $P$ is a stochastic kernel on $\boldsymbol{X}$ given $\mathbb{K}$, called the *transition law*; $P(\cdot|\cdot)$ is a function such that $P(\cdot|k)$ is a probability measure on $\boldsymbol{X}$ for each fixed $k \in \mathbb{K}$, and $P(B|\cdot)$ is a measurable function on $\mathbb{K}$ for each fixed $B \in \boldsymbol{\mathcal{B}}(\boldsymbol{X})$;
5. $c$: $\mathbb{K} \to \mathbb{R}$ is a measurable function called (*one-stage*) *cost function*.

The time horizon of the MDP considered here is discrete and indexed by $\mathbb{N}_0$. Starting from $t = 0$, suppose $X_t$ is the state at time $t \in \mathbb{N}_0$. The decision maker assigns a control $U_t$ according to some criteria and this assignment incurs a cost $c(X_t, U_t)$. The system moves to $X_{t+1}$ according to the transition law $P(X_{t+1} \in B | X_t, U_t)$, for $B \in \mathcal{B}(\boldsymbol{X})$. The decision-and-evolution process repeats itself for all $t \in \mathbb{N}$ and yields a sample path in the canonical sample space explained in the following subsection.

### 1.3.2 Canonical Sample Space

The canonical sample space $\boldsymbol{\Omega}$ is defined as

$$\boldsymbol{\Omega} := (\boldsymbol{X} \times \boldsymbol{U})^{\infty} .$$

The history space up to time $t$ is denoted by $H_t$ and defined as

$$\begin{aligned} \boldsymbol{H}_0 &:= \boldsymbol{X}, \\ \boldsymbol{H}_t &:= (\boldsymbol{X} \times \boldsymbol{U})^t \times \boldsymbol{X} = (\boldsymbol{X} \times \boldsymbol{U}) \times \boldsymbol{H}_{t-1}, \quad t \in \mathbb{N} . \end{aligned}$$

Note that $\boldsymbol{\Omega}$ and $H_t$, $t \in \mathbb{N}_0$, are endowed with the respective product topologies and thus are Borel spaces. A sample point or sample path $\omega \in \boldsymbol{\Omega}$ is written as $\omega = (x_0, u_0, x_1, u_1, \dots)$. Given an $\omega$, the state process $\{X_t\}_{t \in N_0}$, action process $\{U_t\}_{t \in N_0}$ and history process $\{H_t\}_{t \in N_0}$ are defined on $(\boldsymbol{\Omega}, \mathcal{B}(\boldsymbol{\Omega}))$, with $\mathcal{B}(\boldsymbol{\Omega}) = (\mathcal{B}(\boldsymbol{X}) \times \mathcal{B}(\boldsymbol{U}))^{\infty}$, via the random variables $X_t$, $U_t$, and $H_t$, where

$$\begin{aligned} X_t(\omega) &:= x_t , \\ U_t(\omega) &:= u_t , \\ H_t(\omega) &:= (x_0, u_0, \dots, u_{t-1}, x_t) . \end{aligned}$$

### 1.3.3 Policy Space

An admissible *policy* (or *strategy*) $\pi$, is a sequence $\{\pi_t\}_{t\in\mathbb{N}_0}$ where each $\pi_t$ is a stochastic kernel on $\mathcal{U}(x_t)$ given $\boldsymbol{H}_t$ satisfying

$$\pi_t(\mathcal{U}(x_t)|h_t) = 1\,, \quad \forall h_t \in \boldsymbol{H}_t\,, \quad t \in \mathbb{N}_0\,.$$

The set of all admissible policies is denoted by $\Pi$. An admissible policy is *Markov randomized* if there exists a sequence $\{f_t\}_{t\in\mathbb{N}_0}$ of measurable maps $f_t : \boldsymbol{X} \to \boldsymbol{\mathcal{P}}(\boldsymbol{U})$, $t \in \mathcal{N}_0$, satisfying

$$\pi_t(\cdot|h_t) = f_t(x_t)(\cdot)\,.$$

Clearly $f_t(x)(\mathcal{U}(x)) = 1$. A Markov randomized policy $\{f_t\}_{t\in\mathbb{N}_0}$, is called *stationary* if there exists some $f : \boldsymbol{X} \to \boldsymbol{\mathcal{P}}(\boldsymbol{U})$ such that $f_t = f$, for all $t \in \mathbb{N}_0$. In particular, a stationary policy is called *deterministic* (or *pure*) if $f : \boldsymbol{X} \to \boldsymbol{\mathcal{P}}(\boldsymbol{U})$ assigns a point mass for each $x \in \boldsymbol{X}$. That is, $f$ can be expressed as the map: $\boldsymbol{X} \to \boldsymbol{U}$. The set of the Markov randomized, stationary, and deterministic policies are denoted by $\Pi_M$, $\Pi_S$ and $\Pi_D$, respectively, and clearly $\Pi_D \subset \Pi_S \subset \Pi_M \subset \Pi$. Given a policy $\pi \in \Pi$ and an initial distribution $\psi_0 \in \boldsymbol{\mathcal{P}}(\boldsymbol{X})$, there exists an unique probability measure $\mathbb{P}^\pi_{\psi_0}$ on $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{B}}(\boldsymbol{\Omega}))$ (see [11]) satisfying the following:

$$\begin{aligned}
&\mathbb{P}^\pi_{\psi_0}(X_0 \in B) = \psi_0(B)\,;\\
&\mathbb{P}^\pi_{\psi_0}(U_t \in A|H_t) = \pi_t(A|H_t), \quad \mathbb{P}^\pi_{\psi_0} \quad \text{a.s.}\,;\\
&\mathbb{P}^\pi_{\psi_0}(X_{t+1} \in B|H_t, U_t) = P(B|X_t, U_t)\,, \quad \mathbb{P}^\pi_{\psi_0} \quad \text{a.s.}\,.
\end{aligned}$$

Thus the probability space $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{B}}(\boldsymbol{\Omega}), \mathbb{P}^\pi_{\psi_0})$ is well defined. The corresponding expectation operator to $\mathbb{P}^\pi_{\psi_0}$ is denoted by $\mathbb{E}^\pi_{\psi_0}$.

### 1.3.4 Performance Criterion

Given an initial distribution $\psi_0 \in \boldsymbol{\mathcal{P}}(\boldsymbol{X})$ and a policy $\pi \in \Pi$, the most often used performance criterion for a system running on the finite time horizon with terminal cost $h$ is the total cost $J_N(\psi_0, \pi, h)$ defined by

$$J_N(\psi_0, \pi, h) = \mathbb{E}^{\pi}_{\psi_0}\left[\sum_{t=0}^{N-1} c(X_t, U_t) + h(X_N)\right]$$

where the time is indexed up to $N$ and $h : \boldsymbol{X} \to \mathbb{R}$ is a given terminal cost function. For a system running on the infinite time horizon, the discounted cost $J_\beta(\psi_0, \pi)$ and the average cost $J(\psi_0, \pi)$ are the preferred performance criteria where

$$J_\beta(\psi_0, \pi) = \mathbb{E}^{\pi}_{\psi_0}\left[\sum_{t=0}^{\infty} \beta^t \cdot c(X_t, U_t)\right],$$

$\beta \in (0, 1)$ is called *discount factor*, and

$$J(\psi_0, \pi) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^{\pi}_{\psi_0}\left[\sum_{t=0}^{T-1} c(X_t, U_t)\right].$$

## 1.4 Partially Observed Markov Decision Process Model

In this section definitions and concepts of the MDP model are extended to the POMDP model.

### 1.4.1 Model Description

A POMDP is specified by a six-tuple ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{U}, \mathcal{U}, Q, c$) where $\boldsymbol{X}$ is a finite space with cardinality $N_x \in \mathbb{N}$, $\boldsymbol{Y}$ is the finite observation space with cardinality $N_y \in \mathbb{N}$, and $\mathcal{U}(y)$ is the set of feasible action(s) as the observation is $y$. Here for each $y \in \boldsymbol{Y}$ we assume $\mathcal{U}(y) = \boldsymbol{U}$. $Q$ is the transition kernel,

written in the form of a matrix as

$$\begin{aligned} Q(y, U_t)_{X_t j} :&= \text{Prob}(X_{t+1} = j, Y_{t+1} = y | X_k, Y_k, U_k, k \leq t) \\ &= \text{Prob}(X_{t+1} = j, Y_{t+1} = y | X_t, U_t)\,, \end{aligned}$$

for all $t \in \mathbb{N}_0$, $j \in \boldsymbol{X}$, and $y \in \boldsymbol{Y}$. It is easy to see that each element of $Q$ is nonnegative. Moreover,

$$\sum_{y \in \boldsymbol{Y}} \sum_{j \in \boldsymbol{X}} [Q(y, u)]_{ij} = 1$$

for each $i \in \boldsymbol{X}$ and $u \in \boldsymbol{U}$.

### 1.4.2 Canonical Sample Space

The observation history space of a POMDP model up to time $t$ is denoted by $\tilde{\boldsymbol{H}}_t$ and defined by

$$\begin{aligned} \tilde{\boldsymbol{H}}_0 &= \boldsymbol{\Psi} = \boldsymbol{\mathcal{P}}(\boldsymbol{X})\,, \\ \tilde{\boldsymbol{H}}_t &= \tilde{\boldsymbol{H}}_{t-1} \times \boldsymbol{U} \times \boldsymbol{Y}, \quad t \in \mathbb{N}\,. \end{aligned}$$

The canonical sample space is defined by

$$\tilde{\boldsymbol{\Omega}} := (\boldsymbol{X} \times \boldsymbol{Y} \times \boldsymbol{U})^{\infty}\,.$$

These spaces are endowed with their respective product topologies. The state process $\{X_t\}_{t \in \mathbb{N}_0}$, action process $\{U_t\}_{t \in \mathbb{N}_0}$ and history process $\{\tilde{H}_t\}_{t \in \mathbb{N}_0}$ are all defined on $(\tilde{\boldsymbol{\Omega}}, \boldsymbol{\mathcal{B}}(\tilde{\boldsymbol{\Omega}}))$ where $\boldsymbol{\mathcal{B}}(\tilde{\boldsymbol{\Omega}}) = (\boldsymbol{\mathcal{B}}(\boldsymbol{X}) \times \boldsymbol{\mathcal{B}}(\boldsymbol{Y}) \times \boldsymbol{\mathcal{B}}(\boldsymbol{U}))^{\infty}$. Any sample point $\omega = (x_0, y_0, u_0, x_1, y_1, u_1, \dots) \in \tilde{\boldsymbol{\Omega}}$ can be projected onto its coordinates

to obtain

$$
\begin{aligned}
X_t(\omega) &:= x_t\,, \\
Y_t(\omega) &:= y_t\,, \\
U_t(\omega) &:= u_t\,, \\
\tilde{H}_t(\omega) &:= (y_0, u_0, \ldots, u_{t-1}, y_t)\,.
\end{aligned}
$$

### 1.4.3 Policy Space

For a POMDP model an admissible policy is a sequence $\{\pi_t\}_{t\in\mathbb{N}_0}$ where each $\pi_t$ is a stochastic kernel on $\boldsymbol{U}$ given $\tilde{\boldsymbol{H}}_t$ and satisfies

$$
\pi_t(\mathcal{U}(y_t)|h_t) = 1\,, \quad \forall h_t \in \tilde{\boldsymbol{H}}_t, \quad t \in \mathbb{N}_0\,.
$$

The set of $\Pi$, $\Pi_M$, $\Pi_s$ and $\Pi_D$ which corresponds to the set of admissible policies, Markov randomized policies, stationary policies, and deterministic policies, respectively, are defined analogously to the MDP model. Given an initial distribution $\mu \in \boldsymbol{X} \times \boldsymbol{Y}$ and an admissible policy $\pi \in \Pi$, there exists a unique probability measure $\mathbb{P}^\pi_\mu$ on $(\tilde{\boldsymbol{\Omega}}, \boldsymbol{\mathcal{B}}(\tilde{\boldsymbol{\Omega}}))$ where

$$
\begin{aligned}
&\mathbb{P}^\pi_\mu(dX_0, dY_0, dU_0, \ldots, dU_{t-1}, dX_t, dY_t) \\
&\qquad = \mu(dX_0, dY_0)\pi_0(dU_0|\psi_0, Y_0)Q(dX_1, dY_1|X_0, Y_0, U_0)\cdots \\
&\qquad\quad \cdots \pi_{t-1}(dU_{t-1}|\psi_0, Y_0, U_0, \ldots, Y_{t-1})Q(dX_t, dY_t|X_{t-1}, Y_{t-1}, U_{t-1})\,.
\end{aligned}
$$

Here $\psi_0$ is a version of the regular conditional law on $\boldsymbol{X}$ given $y \in \boldsymbol{Y}$ (see [2, section 7]). The expectation operator corresponding to $\mathbb{P}^\pi_\mu$ is denoted by $\mathbb{E}^\pi_\mu$.

### 1.4.4 Model Transformation

$\xi_t \in \boldsymbol{\Psi}$ is called an *information state* for a POMDP model if $\xi_t$ is a function of $(y_0, u_0, \ldots, u_{t-1}, y_t)$ and $\xi_{t+1}$ is a function of $\xi_t$, $y_{t+1}$, and $u_t$ for $t \in \mathbb{N}_0$. Given an initial distribution $\psi_0 \in \boldsymbol{\Psi}$ and an admissible $\pi \in \Pi$, the state process $\{X_t\}_{t\in\mathbb{N}_0}$ induces at time $t$ the conditional probability distribution expressed by

$$\mathbb{P}^{\pi}_{\psi_0}\{X_t = i | Y_t, \ldots, Y_1, U_{t-1}, \ldots, U_0\}, \quad \forall i \in \boldsymbol{X}, \quad t \in \mathbb{N}.$$

It is well know that this conditional probability distribution is an information state and constitutes a sufficient statistic for the history up to time $t$. It is also well known that a partially observed model can be transformed to a completely observable model by calculating these information states. Specifically, we can transform a POMDP model $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{U}, \mathcal{U}, Q, c)$ into the MDP model $(\boldsymbol{\Psi}, \boldsymbol{U}, \tilde{\mathcal{U}}, \mathcal{K}, \tilde{c})$, where the new state space $\boldsymbol{\Psi} = \boldsymbol{\mathcal{P}}(\boldsymbol{X})$ is the space of information states. For each $\psi \in \boldsymbol{\Psi}$, the set of admissible actions is $\tilde{\mathcal{U}} = \boldsymbol{U}$. The cost function $\tilde{c}$ is calculated by

$$\tilde{c}(\psi, u) = \sum_{x\in\boldsymbol{X}} c(x, u)\psi(x) \qquad \forall \psi \in \boldsymbol{\Psi}, u \in \boldsymbol{U}.$$

The information state process $\{\psi_t\}_{t\in\mathbb{N}}$ is generated by the following recursive computation:

$$\psi_{t+1} = T(\psi_t, y_{t+1}, u_t) := \sum_{y\in\boldsymbol{Y}} \frac{\psi_t \cdot Q(y_{t+1}, u_t)}{\psi_t \cdot Q(y_{t+1}, u_t) \cdot \mathbf{1}} \cdot \mathbf{1}_{\{y_{t+1}=y\}}$$

for $\psi_t \in \boldsymbol{\Psi}, u_t \in \boldsymbol{U}$, and $t \in \mathbb{N}_0$. Here $\mathbf{1}_{\{.\}}$ is the indicator function and $\mathbf{1}$ is a column vector of 1's with size $N_x$. Finally, the transition kernel $\mathcal{K} : \boldsymbol{\Psi} \times \boldsymbol{U} \to \boldsymbol{\Psi}$

is expressed by

$$\mathcal{K}(B|\psi,u) := \mathbb{P}^{\pi}_{\psi_0}\{\psi_{t+1} \in B|\psi_t = \psi, U_t = u\} \tag{1.1}$$

for each $t \in \mathbb{N}$, $B \in \mathcal{B}(\boldsymbol{\Psi})$, $\psi \in \boldsymbol{\Psi}$, and $u \in \boldsymbol{U}$. In the terminology of Bayesian estimation, we refer to $T(\psi_t, y_{t+1}, u_t)$ as the posteriori conditional distribution of the state $X_{t+1}$ given an action $u_t \in \boldsymbol{U}$, an observation $y_{t+1} \in \boldsymbol{Y}$, and a priori distribution $\psi_t \in \boldsymbol{\Psi}$. Denote

$$V(\psi_t, y_{t+1}, u_t) := \psi_t \cdot Q(y_{t+1}, u_t) \cdot \mathbf{1}\,,$$

then $V(\psi_t, \cdot, u_t)$ is interpreted as the conditional probability of the observation $y_{t+1}$ on $\boldsymbol{Y}$ given an action $u_t \in \boldsymbol{U}$ and a priori distribution $\psi_t \in \boldsymbol{\Psi}$. Hence, we can write the transition kernel $\mathcal{K}$ in the form

$$\mathcal{K}(B|\psi_t, u_t) = \sum_{y_{t+1} \in \boldsymbol{Y}} V(\psi_t, y_{t+1}, u_t) \cdot \mathbf{1}_{\{T(\psi_t, y_{t+1}, u_t) \in B\}}\,.$$

Define $\overline{\boldsymbol{H}_0} := \boldsymbol{\Psi}$, $\overline{\boldsymbol{H}_t} := (\boldsymbol{\Psi} \times \boldsymbol{U}) \times \overline{\boldsymbol{H}}_{t-1}$ for $t \geq 1$. Through the calculation of information states, we obtain a history

$$(\psi_0, u_0, \psi_1, u_1, \ldots, u_{t-1}, \psi_t) \in \overline{\boldsymbol{H}_t}$$

of the MDP model for a given history

$$\tilde{h}_t = (\psi_0, u_0, y_1, u_1, \ldots, u_{t-1}, y_t) \in \tilde{\boldsymbol{H}}_t$$

of the POMDP model by some correspondence $\zeta_t : \tilde{\boldsymbol{H}}_t \to \overline{\boldsymbol{H}_t}$. A policy $\{\pi_t\}_{t \in \mathbb{N}_0}$ is said to be *separated* if, for each $t$, $\pi_t$ depends on $(y_0, y_1, \ldots, y_{t-1})$ only through the information state $\psi_{t-1}$. Then for a policy $\tilde{\pi}_t$ in the POMDP

model, we can assign a corresponding separated policy $\overline{\pi}_t$ in the MDP model by

$$\overline{\pi}_t(\cdot|\zeta_t(h_t)) := \tilde{\pi}_t(\cdot|h_t)\,, \quad \forall h_t \in \tilde{\boldsymbol{H}}_t\,.$$

It is well known that separated policies are sufficient for optimality. That is, if an optimal policy exists in the POMDP model, there exists an optimal policy in the equivalent MDP model.

## 1.5 Partially Observed Stochastic Game

When a system has two decision makers (or called controllers or players), the POMDP model is extended naturally to the partially observed Markov game (POMG) model. In this case the system is described by a nine-tuple

$$(\boldsymbol{X},\, \boldsymbol{Y},\, \boldsymbol{U},\, \boldsymbol{V}, \mathcal{U}, \mathcal{V}, Q, c_1, c_2)$$

where $\boldsymbol{X} = \{1, \ldots, N_x\}$ is a finite state space and $Y = \{1, \ldots, N_y\}$ a finite observation space. $\boldsymbol{U}$ is the action space for controller 1 and $\boldsymbol{V}$ for controller 2. In addition, $\mathcal{U}(y)$ is the set of available actions for controller 1 and $\mathcal{V}(y)$ for controller 2 when the observation is $y \in \boldsymbol{Y}$. We assume for simplicity that $\mathcal{U}(y) = \boldsymbol{U}$ and $\mathcal{V}(y) = \boldsymbol{V}$ for each $y \in \boldsymbol{Y}$. The transition kernel is specified by

$$\begin{aligned} Q(y, U_t, V_t)_{X_t j} :&= \text{Prob}(X_{t+1} = j, Y_{t+1} = y | X_k, Y_k, U_k, V_k, k \le t) \\ &= \text{Prob}(X_{t+1} = j, Y_{t+1} = y | X_t, Y_t, U_t, V_t) \end{aligned}$$

for $t \in \mathbb{N}_0$, $j \in \boldsymbol{X}$, and $y \in \boldsymbol{Y}$. Elements of $Q$ are nonnegative and satisfy

$$\sum_{y \in \boldsymbol{Y}} \sum_{j \in \boldsymbol{X}} [Q(y, u, v)]_{ij} = 1\,,$$

for $i \in \boldsymbol{X}$, $u \in \boldsymbol{U}$, and $v \in \boldsymbol{V}$. $c_i : \boldsymbol{X} \times \boldsymbol{U} \times \boldsymbol{V} \to \mathbb{R}$ is the cost function for controller $i$, $i = 1, 2$. Specifically, $c_i(x, u, v)$ is the one-stage cost incurred for controller $i$, when the system is at $x \in \boldsymbol{X}$ and action $u, v$ is taken by controller 1, 2, respectively. The evolution of the POMG model is similar to that of the POMDP model except that in the game model the system depends on two controllers that make decisions according to their own criteria. So the canonical sample space becomes

$$\hat{\boldsymbol{\Omega}} = (\boldsymbol{X} \times \boldsymbol{Y} \times \boldsymbol{U} \times \boldsymbol{V})^{\infty}$$

and all the functions dependent on the control $u \in \boldsymbol{U}$ in the PMDDP model now depend on both $u \in \boldsymbol{U}$ and $v \in \boldsymbol{V}$. In particular, the policies $\pi^1$, $\pi^2$ of controller 1, 2, respectively, depend on the history process that is defined on $\hat{\boldsymbol{\Omega}}$. The policy spaces $\Pi$, $\Pi_M$, $\Pi_S$, $\Pi_D$ are defined analogously to those in the POMDP model, and it holds that given an initial distribution $\psi \in \boldsymbol{\Psi}$ and policies $\pi^1 \in \Pi^1$, $\pi^2 \in \Pi^2$, there exists an unique probability measure $\mathbb{P}_{\psi}^{\pi^1 \pi^2}$ on $(\hat{\boldsymbol{\Omega}}, \boldsymbol{\mathcal{B}}(\hat{\boldsymbol{\Omega}}))$. The associated expectation operator is denoted by $\mathbb{E}_{\psi}^{\pi^1 \pi^2}$.

In a way similar to the transformation of a POMDP model to its completely observable MDP model, we can also transform the POMG model to its completely observable equivalent. That is, we can obtain a eight-tuple $(\boldsymbol{\Psi}, \boldsymbol{U}, \boldsymbol{V}, \tilde{\mathcal{U}}, \tilde{\mathcal{V}}, \mathcal{K}, \tilde{c}_1, \tilde{c}_2)$ where for each $\psi \in \boldsymbol{\Psi}$ the set of admissible actions is $\tilde{\mathcal{U}}(\psi) = \boldsymbol{\mathcal{P}}(\boldsymbol{U})$ for controller 1 and $\tilde{\mathcal{V}}(\psi) = \boldsymbol{\mathcal{P}}(\boldsymbol{V})$ for controller 2. The cost function $\tilde{c}_i$ is defined by

$$\tilde{c}_i = \sum_{x \in \boldsymbol{X}} c_i(x, u, v)\psi(x)$$

for $i = 1, 2$. The posterior conditional distribution of the state $X_{t+1}$ given actions $u \in \boldsymbol{U}$, $v \in \boldsymbol{V}$, observation $y \in \boldsymbol{Y}$, and a priori distribution $\psi_t$ can be calculated by

$$T(\psi_t, y_{t+1}, u_t, v_t) = \frac{\psi_t \cdot Q(y_{t+1}, u_t, v_t)}{\psi_t \cdot Q(y_{t+1}, u_t, v_t) \cdot \mathbf{1}}$$

if $\psi_t \cdot Q(y_{t+1}, u_t, v_t) \cdot \mathbf{1}$ is not 0. Denote

$$V(\psi_t, y_{t+1}, u_t, v_t) := \psi_t \cdot Q(y_{t+1}, u_t, v_t) \cdot \mathbf{1}\,,$$

then $V(\psi_t, y_{t+1}, u_t, v_t)$ means the conditional probability of the observation $y_{t+1}$ on $\boldsymbol{Y}$ given actions $u_t \in \boldsymbol{U}$, $v_t \in \boldsymbol{V}$ and a priori distribution $\psi_t \in \boldsymbol{\Psi}$. The transition kernel $\mathcal{K}$ for the POMG model is thus expressed as

$$\begin{aligned}\mathcal{K}(B|\psi, u, v) : &= \mathbb{P}_{\psi_0}^{\pi^1\pi^2}\{\psi_{t+1} \in B | \psi_t = \psi, U_t = u, V_t = v\} \\ &= \sum_{y_{t+1} \in \boldsymbol{Y}} V(\psi, y_{t+1}, u, v) \cdot \mathbf{1}_{\{T(\psi, y_{t+1}, u, v) \in B\}}\end{aligned}$$

for $B \in \mathcal{B}(\boldsymbol{\Psi})$, $\psi \in \boldsymbol{\Psi}$, $u \in \boldsymbol{U}$, and $v \in \boldsymbol{V}$. Hence, given a history in the POMG model, we can obtain its corresponding history in the completely observable model. Furthermore, we can assign a separated policy for the completely observed game model if a policy in the POMG model is given.

## 1.6 Problem Formulation

After the detailed discussion of various models and construction of associated probability spaces as well as the classification of different policies, we formulate our problem for each model.

### 1.6.1 Ergodic Control

Recall that given an initial distribution $\psi_0 \in \boldsymbol{\Psi}$ and an admissible policy $\pi \in \Pi$, the average cost for the POMDP model is

$$J(\psi_0, \pi) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi}_{\psi_0} \left[ \sum_{t=0}^{T-1} \tilde{c}(\psi_t, U_t) \right]$$

where $\tilde{c}(\psi, u) = \sum_{x\in \boldsymbol{X}} c(x, u)\psi(x)$, for $\psi \in \boldsymbol{\Psi}, u \in \boldsymbol{U}$. The objective of ergodic control is to find an optimal control policy $\pi^* \in \Pi$ such that

$$J(\psi_0, \pi^*) = \inf_{\pi\in\Pi} J(\psi_0, \pi) \qquad \forall \psi_0 \in \boldsymbol{\Psi}\,.$$

Our challenge is to provide conditions for the existence of such $\pi^*$ and a method to characterize it.

### 1.6.2 Stochastic Game

For a two-controller POMG model, if an initial distribution $\psi_0 \in \boldsymbol{\Psi}$ and admissible policies $\pi^1 \in \Pi^1$, $\pi^2 \in \Pi^2$ of controller 1, 2, respectively, are given, then the incurred long-run average cost for controller $i$, $i = 1, 2$, respectively, is

$$J_i(\psi_0, \pi^1, \pi^2) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi^1\pi^2}_{\psi_0} \left[ \sum_{t=0}^{T-1} \tilde{c}_i(\psi_t, U_t, V_t) \right]$$

where $\tilde{c}_i(\psi, u, v) = \sum_{x\in \boldsymbol{X}} c_i(x, u, v)\psi(x)$, for $\psi \in \boldsymbol{\Psi}, u \in \boldsymbol{U}$, and $v \in \boldsymbol{V}$. Our goal is to provide conditions to imply the existence of a Nash equilibrium $(\pi^{1*}, \pi^{2*}) \in \Pi^1 \times \Pi^2$ for both controllers' policies where

$$J_1(\psi_0, \pi^{1*}, \pi^{2*}) = \inf_{\pi^1\in\Pi^1} J_1(\psi_0, \pi^1, \pi^{2*}) \qquad \forall \psi_0 \in \boldsymbol{\Psi}\,,$$

$$J_2(\psi_0, \pi^{1*}, \pi^{2*}) = \inf_{\pi^2\in\Pi^2} J_2(\psi_0, \pi^{1*}, \pi^2) \qquad \forall \psi_0 \in \boldsymbol{\Psi}\,.$$

### 1.6.3 Adaptive Control

Suppose that the transition matrix $Q$ depends on a parameter vector $\theta \in \Theta$ where $\Theta \in \mathbb{R}^{N_\theta}$ is the parameter space. Then the expectation operator will be parameterized by $\theta$ and written as $\mathbb{E}^{\pi}_{\psi_0,\theta}$ for a given $\psi_0 \in \boldsymbol{\Psi}$ and $\pi \in \Pi$. The corresponding average cost $J_\theta(\psi_0, \pi)$ is

$$J_\theta(\psi_0, \pi) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi}_{\psi_0,\theta} \left[ \sum_{t=0}^{T-1} \tilde{c}(\psi_t, U_t) \right] .$$

The objective of adaptive control is to design an optimal adaptive policy $\pi^a$ such that

$$J_\theta(\psi_0, \pi^a) = \inf_{\pi\in\Pi} J_\theta(\psi_0, \pi) .$$

### 1.6.4 Safety Control

For a system modelled by a MDP with finite $\boldsymbol{X} = \{1, 2, \cdots, N_x\}$, the safety specification $\boldsymbol{\Psi}_s$ is defined as the following convex set

$$\boldsymbol{\Psi}_s := \{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A} \leq \boldsymbol{b}\} \tag{1.2}$$

where $\boldsymbol{b} \in \mathbb{R}^{N_b}$, $N_b \in \mathbb{N}$, $\boldsymbol{A} \in \mathbb{R}^{N_x \times N_b}$. Let $P_\pi$ denote the transition matrix of the Markov decision process when the policy is $\pi$. Our objective in the safety control is to find an admissible policy $\pi \in \Pi$ and an associated set $\boldsymbol{\Psi}_{in} \subset \boldsymbol{\Psi}_s$ such that for any initial state probability distribution $\psi$ of the system we have

$$\psi \in \boldsymbol{\Psi}_{in} \quad \Rightarrow \quad \psi^{(k)} = \psi P^k_\pi \in \boldsymbol{\Psi}_s \qquad \forall k \in \mathbb{N} . \tag{1.3}$$

If the system is modelled by a POMDP, the safety specification $\boldsymbol{\Psi}_s$ is similarly defined as in (1.2). We aim at the search of an admissible policy $\pi \in \Pi$ and

an associated set $\boldsymbol{\Psi}_{in} \subset \boldsymbol{\Psi}_s$ such that for any initial information state $\psi$ of the system we have

$$\psi \in \boldsymbol{\Psi}_{in} \quad \Rightarrow \quad \mathbb{P}^{\pi}_{\psi}(\psi_k \in \boldsymbol{\Psi}) = 1 \qquad \forall k \in \mathbb{N} \tag{1.4}$$

where $\psi_k \in \boldsymbol{\Psi}_s$ is the $k$th step of the information state.

# Chapter 2

# Ergodic Control of Partially Observed Markov Decision Processes

## 2.1 Introduction and Preliminaries

Since the pioneering work of R. Bellman [8] in the 1950s, the controlled Markov process has been an important and active research topic due to its wide applications in operations research, communication networks, macroeconomics, and other fields. Special attention has been given to the study of the partially observable mode in which only imperfect information is available to the decision maker. This consideration provides potential applications to reliability-related problems, but mathematically this setting is more challenging, and answers to problems in this class are far from being complete. The standard approach to the control problem of the partially observed Markov decision process is to transform it into a fully observed separated control problem by considering the information states. The dynamic programming characterizing the optimal policy is usually derived from the *vanishing discount limit* of the corresponding discounted cost control problem. The main research issue focuses on the search of appropriate assumptions resulting in a bounded *differential discounted value function* (see [2, p301]), which is essential in justifying the argument of the vanishing discount limit. Classical assumptions include Ross's *renewability* condition [65] and Platzman's *reachability-*

*detectability* condition [61]. Recently, Chuang and Arapostathis [19] proposed a very weak assumption, which is satisfied by a benchmark machine replacement problem [17, Example 2.3.1] that can not be dealt with either by Ross's or Platzman's approach. Here we further weaken Chuang and Arapostathis's condition and give an example to show our work.

This chapter is organized as follows. For the remainder of this section we introduce the preliminaries for the ergodic control problem. Section 2.2 provides some notations and presents the main result of this chapter. Section 2.3 adds some remarks and presents an example to compare our work with the previous work mentioned above.

For a given initial state probability distribution $\psi_0 \in \boldsymbol{\Psi}$ and a given policy $\pi \in \Pi$, the incurred long-run average cost is expressed by

$$J(\psi_0, \pi) := \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}^{\pi}_{\psi_0} \left[ \sum_{t=0}^{T-1} \tilde{c}(\psi_t, U_t) \right] . \tag{2.1}$$

The classical approach, called *vanishing discount* method, is to start from the discounted cost model. Specifically, we consider the $\beta$-discounted model with the incurred cost

$$J_\beta(\psi_0, \pi) := \limsup_{T\to\infty} \mathbb{E}^{\pi}_{\psi_0} \left[ \sum_{t=0}^{T-1} \beta^t \tilde{c}(\psi_t, U_t) \right] . \tag{2.2}$$

The goal of optimization for the $\beta$-discounted model is also to find the minimizing policy $\pi_\beta$ and value function $h_\beta(\psi)$ such that

$$h_\beta(\psi) = \inf_{\pi_\beta \in \Pi} J_\beta(\psi, \pi), \qquad \psi \in \boldsymbol{\Psi}, \beta \in (0,1) . \tag{2.3}$$

**Assumption 2.1.1.** The one-stage cost function $c : \boldsymbol{X} \times \boldsymbol{U} \to \mathbb{R}^+$ is non-negative, bounded and continuous. Also, $U \to Q(y, U)$ is continuous for each $y \in \boldsymbol{Y}$.

**Lemma 2.1.1.** *[38] Suppose Assumption 2.1.1 holds. Then the value function $h_\beta(\psi)$ in (2.3) corresponding to the $\beta$-discounted cost model in (2.2) can be characterized by Bellman's $\beta$-discounted optimality equation:*

$$h_\beta(\psi) = \min_{u \in \boldsymbol{U}} \left\{ \tilde{c}(\psi, u) + \beta \int_{\boldsymbol{Y}} h_\beta(\psi') \mathcal{K}(d\psi' | \psi, u) \right\} \tag{2.4}$$

*for all $\psi \in \boldsymbol{\Psi}$ where $\mathcal{K}$ is defined in (1.1). Any policy resulting in the value function $h_\beta(\psi)$ for each $\psi \in \boldsymbol{\Psi}$ is optimal in the $\beta$-discounted sense.*

It is well known that $h_\beta(\psi)$ is the unique solution in $\boldsymbol{C}(\boldsymbol{\Psi})$ (the space of continuous functions on $\boldsymbol{\Psi}$) of Bellman's $\beta$-discounted optimality equation. It is also well known [61] that $h_\beta(\cdot)$ is concave. This is the main property that we employ for the development of our approach.

## 2.2 Main Result

Denote $\pi_\beta$ the optimal policy for the $\beta$-discounted model and $e^i$ the $i^{th}$ vertex of $\boldsymbol{\Psi}$, i.e., the $i^{th}$ row vector of the identity matrix with size $N_x \times N_x$. Also, define the following notation:

$$\psi^* := \arg\max_{\psi \in \boldsymbol{\Psi}} h_\beta(\psi)\,, \qquad \psi_* := \arg\min_{\psi \in \boldsymbol{\Psi}} h_\beta(\psi)\,,$$
$$\overline{h}_\beta(\psi) := h_\beta(\psi) - h_\beta(\psi_*)\,, \qquad \Delta h := h_\beta(\psi^*) - h_\beta(\psi_*)\,.$$

Here, $\overline{h}_\beta(\psi)$ is called *differential discounted value function*. Write $y^k := (y_1, y_2, \ldots, y_k)$ and $u^k := (u_0, u_1, \ldots, u_k)$ for the history of observations and

actions, respectively. Also,

$$\begin{aligned} Q(y^k, u^{k-1}) &:= Q(y_1, u_0) \cdots Q(y_k, u_{k-1}) \,, \\ V(\psi, y^k, u^{k-1}) &:= \psi Q(y^k, u^{k-1}) \mathbf{1} \,, \\ T(\psi, y^k, u^{k-1}) &:= \frac{\psi Q(y^k, u^{k-1})}{V(\psi, y^k, u^{k-1})} \,. \end{aligned}$$

Now we present the main assumption.

**Assumption 2.2.1.** (Achievability) There exist constants $\varepsilon > 0$, $N_0 \in \mathbb{N}$ and $\beta_0 < 1$ such that $\forall \beta \in [\beta_0, 1)$ we have

$$\begin{aligned} \max_{1 \le k \le N_0} \mathbb{P}^{\pi_\beta}_{\psi_*} \{ & T(\psi_*, Y^k, U^{k-1}) \ge \varepsilon \cdot T(\psi^*, Y^k, U^{k-1}) \,, \\ & V(\psi^*, Y^k, U^{k-1}) \ge \varepsilon \cdot V(\psi_*, Y^k, U^{k-1}) \} \ge \varepsilon \,. \end{aligned}$$

**Lemma 2.2.1.** *Under Assumption 2.2.1, let $\{\beta_n\}_{n=1}^{\infty} \subset [\beta_0, 1)$ be a sequence and $\beta_n \to 1$, then the sequence of differential discounted value functions $\{\overline{h}_{\beta_n}(\cdot)\}_{n=1}^{\infty}$ is uniformly bounded on $\boldsymbol{\Psi}$.*

*Proof.* See Appendix A. □

**Remark 2.2.2.** A stronger version of the above assumption is the following. There exist constants $\varepsilon > 0$, $N_0 \in \mathbb{N}$ and $\beta_0 < 1$ such that for each $\beta \in [\beta_0, 1)$ and $1 \le i, i_1, i_2, j \le N_x$ we have

$$\max_{1 \le k \le N_0} \mathbb{P}^{\pi_\beta}_{e^i} \{ Q(Y^k, U^{k-1})_{i_1 j} \ge \varepsilon \cdot Q(Y^k, U^{k-1})_{i_2 j} \} \ge \varepsilon \,.$$

If all the nonnegative parameters in all the transition kernels are lower-bounded by a positive constant, then Assumption 2.2.1 can be simplified as

$$\max_{1 \le k \le N_0} \mathbb{P}^{\pi_\beta}_{\psi_*} \{ T(\psi_*, Y^k, U^{k-1}) \ge \varepsilon \cdot T(\psi^*, Y^k, U^{k-1}) \} \ge \varepsilon \,.$$

**Theorem 2.2.3.** *Suppose Assumption 2.1.1 and 2.2.1 hold. Then there exist a constant $\rho$, which is the optimal ergodic cost, and a bounded, concave and continuous function $h$: $\boldsymbol{\Psi} \to \boldsymbol{R}$, such that $(\rho, h(\cdot))$ is a solution of the following dynamic programming equation:*

$$\rho + h(\psi) = \min_{u \in \boldsymbol{U}} \left\{ \tilde{c}(\psi, u) + \int_{\boldsymbol{Y}} h(\psi') \mathcal{K}(d\psi' | \psi, u) \right\} . \tag{2.5}$$

*Also, the following statements are equivalent.*

1. *$\pi^*$ is an optimal policy.*

2. *$\pi^*(\psi)$ assigns a minimizer $u$ for $\{\cdot\}$ in (2.5) for each $\psi \in \boldsymbol{\Psi}$.*

3. $$\lim_{t \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}^{\pi^*}_{\psi_0} \{ D(\psi_t, \pi^*(\psi_t)) \} = 0$$

   *where the discrepancy function $D : \boldsymbol{\Psi} \times \boldsymbol{U} \to \mathbb{R}$ is defined by*

$$D(\psi, u) := \tilde{c}(\psi, u) + \int_{\boldsymbol{Y}} h(\psi') \mathcal{K}(d\psi' | \psi, u) - \rho - h(\psi) .$$

*Proof.* Subtracting $h_\beta(\psi_*)$ at both sides of the equal sign of (2.4) we obtain

$$\overline{h}_\beta(\psi) + (1 - \beta) h_\beta(\psi_*) = \min_{u \in \boldsymbol{U}} \left\{ \tilde{c}(\psi, u) + \beta \int_{\boldsymbol{Y}} \overline{h}_\beta(\psi') \mathcal{K}(d\psi' | \psi, u) \right\} . \tag{2.6}$$

Since $0 \leq (1-\beta) h_\beta(\psi_*) \leq \|c\|_\infty$, $\{(1-\beta_n) h_{\beta_n}(\psi_*)\}_{n=0}^\infty$ is uniformly bounded in $n$. By the Bolzano-Weierstrass Theorem there exists a subsequence $\{\beta_{n'}\}_{n'=0}^\infty$ of $\{\beta_n\}_{n=0}^\infty$ and a constant $\rho$ such that

$$(1 - \beta_{n'}) h_{\beta_{n'}}(\psi_*) \to \rho \quad \text{as} \quad n' \to \infty .$$

On the other hand, due to the concavity of $\{h_{\beta_{n'}}(\cdot)\}$ on $\boldsymbol{\Psi}$, $\{\overline{h}_{\beta_{n'}}(\cdot)\}$ is also concave on $\boldsymbol{\Psi}$. This property along with the uniform boundedness of $\{\overline{h}_{\beta_{n'}}(\cdot)\}$

imply that it is locally equi-Lipschitzian, and in particular, equicontinuous [27]. Thus the Arzela-Ascoli Theorem can be applied to conclude that there exist a subsequence $\{\beta_n''\}_{n''=0}^{\infty}$ of $\{\beta_n'\}_{n'=0}^{\infty}$ and a continuous function $h(\psi)$ such that

$$\overline{h}_{\beta_{n''}}(\psi) \to h(\psi) \quad \text{as} \quad n'' \to \infty \quad \forall \psi \in \boldsymbol{\Psi}\,.$$

Apparently $h(\cdot)$ is concave on $\boldsymbol{\Psi}$. Due to the existence of $(\rho, h(\cdot))$ that solves (2.5), given any admissible $\pi \in \Pi$ and initial condition $\psi_t \in \boldsymbol{\Psi}$, we have for each $t \in \mathbb{N}_0$

$$D(\psi_t, u_t) = \tilde{c}(\psi_t, u_t) + \mathbb{E}_{\psi_t}^{\pi} h(\psi_{t+1}) - \rho - h(\psi_t) \geq 0\,.$$

Adding the terms from $t = 0$ to $t = T-1$ we obtain

$$\begin{aligned}\sum_{t=0}^{T-1} D(\psi_t, u_t) = \sum_{t=0}^{T-1} \tilde{c}(\psi_t, u_t) + \sum_{t=0}^{T-1} \mathbb{E}_{\psi_t}^{\pi} h(\psi_{t+1}) \\ - T\rho - \sum_{t=0}^{T-1} h(\psi_{t+1}) - h(\psi_0) + h(\psi_t) \geq 0\,.\end{aligned}$$

Now on both sides of the equality take the expectation $\mathbb{E}_{\psi_0}^{\pi}$, multiply $1/T$ and let $T \to \infty$. Note that $h(\cdot)$ is bounded due to its continuity on the compact $\boldsymbol{\Psi}$. We thus have

$$\begin{aligned}&\limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_{\psi_0}^{\pi} \left[ \sum_{t=0}^{T-1} D(\psi_t, U_t) \right] \\ =&\limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_{\psi_0}^{\pi} \left[ \sum_{t=0}^{T-1} \tilde{c}(\psi_t, U_t) \right] - \rho \geq 0\,. \end{aligned} \tag{2.7}$$

When $\pi = \pi^*$ the inequality in (2.7) becomes equality and the proof is completed. □

## 2.3 Example and Concluding Remark

Consider the following assumption recently proposed by Chuang and Arapostathis [19]:

**Assumption 2.3.1.** (Interior Accessibility) There exist constants $\varepsilon > 0$, $N_0 \in \mathbb{N}$ and $\beta_0 < 1$ such that $\forall \beta \in [\beta_0, 1)$ and $1 \leq i \leq N_x$ we have

$$\max_{1 \leq k \leq N_0} \mathbb{P}_{e^i}^{\pi_\beta} \{\psi_k \in \boldsymbol{\Psi}_\varepsilon\} \geq \varepsilon$$

where $\boldsymbol{\Psi}_\varepsilon := \{\psi \in \boldsymbol{\Psi} | \psi(i) \geq \varepsilon, \forall i = 1, 2, \ldots, N_x\}$.

Also consider a partially observable Markov process with the following property:

**Property 2.3.1.** Given any observation $y \in \boldsymbol{Y}$ there exists a nonempty set $\boldsymbol{X}_y \subset \boldsymbol{X}$ and $\boldsymbol{X}_y \neq \boldsymbol{X}$ such that all the possible states that the system lies in corresponding to the observation $y$ are contained in $\boldsymbol{X}_y$.

**Remark 2.3.2.** It is clear that if for all the transition kernels the ratios of their row sums are bounded, then *achievability* is a weaker assumption than *interior accessibility*. On the other hand, Assumption 2.3.1 excludes the class of problems with Property 2.3.1. The *mixed observation* process that there exists at least one completely observable state is an example owning this property. We will discuss this at the end of this section.

Now we study an example in detail to compare various assumptions. Suppose there is a machine whose performance is evaluated and classified into state 1, 2 and 3 representing *good, need maintenance*, and *down*, respectively.

Suppose also that the available actions are action 0 and 1, meaning *resume* and *replace*, respectively. We represent this system by the state space $\boldsymbol{X} = \{1, 2, 3\}$ and the action space $\boldsymbol{U} = \{0, 1\}$. It is natural to assume that the relation between the costs $c(x, u)$ for various kinds of combinations of state $x \in \boldsymbol{X}$ and action $u \in \boldsymbol{U}$ is

$$0 \leq c(1, 0) < c(2, 0) < c(3, 0) < c(x, 1) < \infty \,. \tag{2.8}$$

It is also natural to assume that the machine will deteriorate statistically over time if no maintenance is done at all. When the *replace* action is taken, the performance is improved. Suppose there exists a probability of erroneous observation between state 1, 2 and 3, then the process becomes only partially observable. Assume the observation space $\boldsymbol{Y} = \boldsymbol{X}$. We can thus write the transition probability matrix $Q_u^y = P_u O_y$ where

$$P_{u=0} = \begin{bmatrix} p_{11}^0 & p_{12}^0 & p_{13}^0 \\ 0 & p_{22}^0 & p_{23}^0 \\ 0 & 0 & 1 \end{bmatrix}, P_{u=1} = \begin{bmatrix} 1 & 0 & 0 \\ p_{21}^1 & p_{22}^1 & 0 \\ p_{31}^1 & p_{32}^1 & 0 \end{bmatrix},$$

and $P_u$ is a stochastic matrix for $u \in \boldsymbol{U} = \{0, 1\}$. Assume those $p_{ij}^k$'s are lower bounded by a constant. Also express $O_y$ as

$$O_y = \begin{bmatrix} q_{1y} & 0 & 0 \\ 0 & q_{2y} & 0 \\ 0 & 0 & q_{3y} \end{bmatrix}.$$

Note that $[q_{i1}, q_{i2}, q_{i3}] \in \boldsymbol{\mathcal{P}}(\boldsymbol{X})$ for $i \in \boldsymbol{X}$.

According to the setting, it is not difficult to deduce that $\psi_*$=[1 0 0] and $\pi_\beta(\psi_*) = 0$ for every $\beta \in$ (0,1), with an argument similar to that in [17, Example 2.3.1]. To compare with different assumptions proposed before, consider the following two cases:

*Case 1*: There exists an observation $y \in \boldsymbol{Y}$ such that $q_{1y}, q_{2y}$, and $q_{3y}$ are all lower-bounded by a positive constant. We can show that in this case Assumption 2.3.1 is satisfied but the *detectability* condition and *renewability* condition in [61] both fail.

*Case 2*: If there is no observation error when the state *down* is observed, then we can write

$$O_{y=1} = \begin{bmatrix} q & 0 & 0 \\ 0 & 1-q & 0 \\ 0 & 0 & 0 \end{bmatrix}, O_{y=2} = \begin{bmatrix} 1-q & 0 & 0 \\ 0 & q & 0 \\ 0 & 0 & 0 \end{bmatrix}, O_{y=3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where we assume $q \in (.5, 1)$. In this case state 1 serves as a recurrent state for the partially observable process so the *renewability* condition in [61] is satisfied. However, since the information state $\psi_t$ will never enter the interior of the simplex $\boldsymbol{\Psi}$, Assumption 2.3.1 fails. Finally we note that in *Case 1* Assumption 2.2.1 is satisfied from Remark 2.3.2. In *Case 2* Assumption 2.2.1 is satisfied with $N_0$=1 and

$$\varepsilon = \min\{\frac{p^0_{23}}{p^0_{13}}, p^0_{13}\}$$

for all $\beta \in$[0.5 1).

# Chapter 3

# Two-Person Zero-Sum Stochastic Games with Mixed Observation Structure

## 3.1 Introduction and Preliminaries

In this chapter we extend the ergodic control problem of a Markov decision process from one controller to two controllers. The new model is called *two-person stochastic game* and is introduced in Section 1.5. In our work we focus on the model with two features. The first one is called *zero-sum* feature since the controllers pursue benefits that are in conflict. The second one is called *mixed observation* feature in the sense that the system's states are only partially observable except for at least one state that is completely observable. Transforming the stochastic game model with mixed observation structure into its completely observable equivalent, and then applying the vanishing discount approach, we derive the Nash equilibrium for both controllers' policies in the form of a dynamic programming. Two-person zero-sum stochastic games with complete observation have been studied in [43, 47, 55, 56, 57, 58] where the generalized state space is considered. In our work, the concept of the recurrent state is used to derive the main result.

This chapter is organized as follows. For the remainder of this section we explain the notion of the Nash equilibrium for controllers' policies in a two-person zero-sum stochastic game. The main result of this chapter is presented

in Section 3.2.

As mentioned in Section 1.5, in the partially observed Markov game model, given an initial distribution $\psi_0 \in \boldsymbol{\Psi}$ and a pair of admissible policies $(\psi^1, \psi^2) \in \Pi^1 \times \Pi^2$, there exists a unique probability measure $\mathbb{P}_{\psi_0}^{\pi^1\pi^2}$ on its associated measurable space $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{B}}(\boldsymbol{\Omega}))$. We denote the expectation operator as $\mathbb{E}_{\psi_0}^{\pi^1\pi^2}$ and write the long-run average cost for controller $i$, $i = 1, 2$, as

$$J_i(\psi_0, \pi^1, \pi^2) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}_{\psi_0}^{\pi^1\pi^2} \left[ \sum_{t=0}^{T-1} \tilde{c}_i(\psi_t, U_t, V_t) \right]$$

where $\psi_t \in \boldsymbol{\Psi}, U_t \in \boldsymbol{U}, V_t \in \boldsymbol{V}$ and $\tilde{c}_i(\psi_t, U_t, V_t) = \sum_x c_i(x, U_t, V_t)\psi_t(x)$. The Nash equilibrium is a pair of policies $(\pi^{1*}, \pi^{2*}) \in \Pi^1 \times \Pi^2$ satisfying

$$J_1(\psi_0, \pi^{1*}, \pi^{2*}) \leq J_1(\psi_0, \pi^1, \pi^{2*}) \qquad \forall \pi^1 \in \Pi^1 \,,$$

$$J_2(\psi_0, \pi^{1*}, \pi^{2*}) \leq J_2(\psi_0, \pi^{1*}, \pi^2) \qquad \forall \pi^2 \in \Pi^2 \,.$$

In the case of a zero-sum game,

$$c_1(x, u, v) + c_2(x, u, v) = 0 \qquad \forall x \in \boldsymbol{X}, u \in \boldsymbol{U}, v \in \boldsymbol{V}.$$

Denote $J_2 = J$, then we have

$$J(\psi_0, \pi^{1*}, \pi^2) \geq J(\psi_0, \pi^{1*}, \pi^{2*}) \geq J(\psi_0, \pi^1, \pi^{2*})$$

for all $\psi_0, \pi \in \Pi^1, \pi \in \Pi^2$. That is

$$J(\psi_0, \pi^{1*}, \pi^2) \geq \sup_{\pi^1\in\Pi^1} J(\psi_0, \pi^1, \pi^{2*}) \geq \inf_{\pi^2\in\Pi^2} \sup_{\pi^1\in\Pi^1} J(\psi_0, \pi^1, \pi^2) \tag{3.1}$$

and

$$J(\psi_0, \pi^1, \pi^{2*}) \leq \inf_{\pi^2\in\Pi^2} J(\psi_0, \pi^{1*}, \pi^2) \leq \sup_{\pi^1\in\Pi^1} \inf_{\pi^2\in\Pi^2} J(\psi_0, \pi^1, \pi^2) \,. \tag{3.2}$$

We say a policy $\pi^{1*}$ satisfying (3.1) for each $\psi_0 \in \boldsymbol{\Psi}$ and $\pi^2 \in \Pi^2$ is optimal for controller 1 , and a policy $\pi^{2*}$ satisfying (3.2) for each $\psi_0 \in \boldsymbol{\Psi}$ and $\pi^1 \in \Pi^1$ is optimal for controller 2. Moreover, we obtain

$$\sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} J(\psi_0, \pi^1, \pi^2) \geq \inf_{\pi^2 \in \Pi^2} J(\psi_0, \pi^{1*}, \pi^2) \geq \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} J(\psi_0, \pi^1, \pi^2) \tag{3.3}$$

where the second inequality follows from (3.1). So we conclude that when there exists a Nash equilibrium $(\pi^{1*}, \pi^{2*}) \in \Pi^1 \times \Pi^2$ then $\pi^{i*}$ is optimal for controller $i$, $i = 1, 2$, and

$$\sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} J(\psi_0, \pi^1, \pi^2) = \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} J(\psi_0, \pi^1, \pi^2) = J(\psi_0, \pi^{1*}, \pi^{2*})$$

for all $\psi_0 \in \boldsymbol{\Psi}$. Here $J(\psi_0, \pi^{1*}, \pi^{2*})$ is called the value of the game and the Nash equilibrium is also called *saddle point equilibrium.*

## 3.2 Main Result

For the ease of the notation burden, we define the following with a little bit abuse of symbols. For $\psi \in \boldsymbol{\mathcal{P}}(\boldsymbol{X})$, $\tilde{u} \in \boldsymbol{\mathcal{P}}(\boldsymbol{U})$, and $\tilde{v} \in \boldsymbol{\mathcal{P}}(\boldsymbol{V})$

$$\begin{aligned}
\hat{c}(\psi, \tilde{u}, \tilde{v}) &:= \sum_{x \in \boldsymbol{X}} \sum_{u \in \boldsymbol{U}} \sum_{v \in \boldsymbol{V}} c(x, u, v) \cdot \psi(x) \cdot \tilde{u}(u) \cdot \tilde{v}(v)\,, \\
Q(y, \tilde{u}, \tilde{v}) &:= \sum_{u \in \boldsymbol{U}} \sum_{v \in \boldsymbol{V}} Q(y, u, v) \cdot \tilde{u}(u) \cdot \tilde{v}(v)\,, \\
V(\psi, y, \tilde{u}, \tilde{v}) &:= \psi \cdot Q(y, \tilde{u}, \tilde{v}) \cdot \mathbf{1}\,, \\
T(\psi, y, \tilde{u}, \tilde{v}) &:= \frac{\psi \cdot Q(y, \tilde{u}, \tilde{v})}{V(\psi, y, \tilde{u}, \tilde{v})}\,.
\end{aligned}$$

**Assumption 3.2.1.** Assume that the cost function $c : \boldsymbol{X} \times \boldsymbol{U} \times \boldsymbol{V}$ is non-negative and bounded, the transition law $Q(y, \cdot, \cdot)$ is continuous on $\boldsymbol{U} \times \boldsymbol{V}$ for

each $y \in \boldsymbol{Y}$. Furthermore, for each observation the available action space is $\boldsymbol{\mathcal{P}}(\boldsymbol{U})$ for player 1 and $\boldsymbol{\mathcal{P}}(\boldsymbol{V})$ for player 2.

**Lemma 3.2.1.** *For every $\beta \in$ (0,1), under Assumption 3.2.1, the partially observable zero-sum stochastic game with $\beta-$discounted criterion always has a value and both players have optimal stationary strategies which are characterized by Shapley's $\beta$-discounted equation:*

$$\begin{aligned} h_\beta(\psi) &= \max_{\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})} \min_{\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})} \left[\hat{c}(\psi,\tilde{u},\tilde{v}) + \beta \sum_{y\in\boldsymbol{Y}} V(\psi,y,\tilde{u},\tilde{v}) \cdot h_\beta(T(\psi,y,\tilde{u},\tilde{v}))\right] \\ &= \min_{\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})} \max_{\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})} \left[\hat{c}(\psi,\tilde{u},\tilde{v}) + \beta \sum_{y\in\boldsymbol{Y}} V(\psi,y,\tilde{u},\tilde{v}) \cdot h_\beta(T(\psi,y,\tilde{u},\tilde{v}))\right], \end{aligned} \tag{3.4}$$

*for each $\psi \in \boldsymbol{\Psi}$. That is, any pair of separated policies $(\pi^{1*}, \pi^{2*})$ satisfying (3.4) is a saddle-point equilibrium for the $\beta-$discounted cost model.*

**Assumption 3.2.2.** There exists an observation $y^*$ such that for all $u \in \boldsymbol{U}$ and $v \in \boldsymbol{V}$, $Q(y^*, u, v)$ is a matrix of rank 1 with at least one of its columns positive.

**Example 3.2.1.** Consider a system with mixed observation structure. Suppose the system has a completely observable state $x \in \boldsymbol{X}$ that is accessible by all other states. Also, when a state $y \in \boldsymbol{Y}$, $y \neq x$, is observed at some time step, the probability that the system is actually in state $x$ is 0, then this system satisfies Assumption 3.2.2.

**Theorem 3.2.2.** *Under Assumption 3.2.1 and Assumption 3.2.2, the two-person zero-sum stochastic game with average-cost criterion always has a value.*

*Both controllers have optimal stationary strategies which are characterized by Shapley's ergodic equation:*

$$\begin{aligned}\rho + h(\psi) &= \max_{\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})} \min_{\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})} \left[\hat{c}(\psi,\tilde{u},\tilde{v}) + \sum_{y\in\boldsymbol{Y}} V(\psi,y,\tilde{u},\tilde{v})\cdot h(T(\psi,y,\tilde{u},\tilde{v}))\right] \\ &= \min_{\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})} \max_{\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})} \left[\hat{c}(\psi,\tilde{u},\tilde{v}) + \sum_{y\in\boldsymbol{Y}} V(\psi,y,\tilde{u},\tilde{v})\cdot h(T(\psi,y,\tilde{u},\tilde{v}))\right] .\end{aligned} \tag{3.5}$$

*That is, any pair of separated policies $(\pi^{1*},\pi^{2*})$ satisfying (3.5) is a saddle-point equilibrium for the ergodic cost model.*

*Proof.* Suppose $y^*$ is the observation that the matrix $Q(y^*,u,v)$ is of rank 1. Under the assumption we can write $Q(y^*,u,v) = Q_a\cdot Q_b^T$, where $Q_a$ and $Q_b$ are both $N_x\times 1$ vectors and each element of $Q_a$ is positive. Also, there exists an $\alpha$ such that $0<\alpha\le V(\psi,y^*,\tilde{u},\tilde{v})$ for each $\psi\in\boldsymbol{\Psi}$, $\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})$ and $\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})$. Now consider a new process which has the same $T(\psi,y,\tilde{u},\tilde{v})$ but with new conditional probability of observation $y$:

$$V^*(\psi,y,\tilde{u},\tilde{v}) = \begin{cases} \frac{V(\psi,y,\tilde{u},\tilde{v})-\alpha}{1-\alpha} & \text{if} \quad y = y^*, \\ \frac{V(\psi,y,\tilde{u},\tilde{v})}{1-\alpha} & \text{if} \quad y\in\boldsymbol{Y}\setminus\{y^*\} \end{cases} .$$

Denote the value function with $(1-\alpha)-$discounted criterion for this new process by $\hat{h}_{1-\alpha}$, then by Lemma 3.2.1 the associated Shapley equation can be

written in the following:

$$\begin{aligned}
&\hat{h}_{1-\alpha}(\psi) \\
&= \max_{\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})} \min_{\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})} \left[\hat{c}(\psi,\tilde{u},\tilde{v}) + (1-\alpha)\sum_{y\in\boldsymbol{Y}} V^*(\psi,y,\tilde{u},\tilde{v})\hat{h}_{1-\alpha}(T(\psi,y,\tilde{u},\tilde{v}))\right] \\
&= \max_{\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})} \min_{\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})} \left[\hat{c}(\psi,\tilde{u},\tilde{v}) + (1-\alpha)\sum_{y\in\boldsymbol{Y}\setminus\{y^*\}} \frac{V(\psi,y,\tilde{u},\tilde{v})}{1-\alpha}\hat{h}_{1-\alpha}(T(\psi,y,\tilde{u},\tilde{v}))\right. \\
&\qquad\qquad \left. + (1-\alpha)\frac{V(\psi,y^*,\tilde{u},\tilde{v})-\alpha}{1-\alpha}\hat{h}_{1-\alpha}(\psi^*)\right] ..
\end{aligned}$$

Denote

$$\begin{aligned}
h^*(\psi) &:= \hat{h}_{1-\alpha}(\psi) - \hat{h}_{1-\alpha}(\psi^*)\,, \\
\rho^* &:= \alpha\cdot\hat{h}_{1-\alpha}(\psi^*)\,,
\end{aligned}$$

then we have

$$\begin{aligned}
\rho^* + h^*(\psi) &= \max_{\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})} \min_{\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})} \left[\hat{c}(\psi,\tilde{u},\tilde{v}) + \sum_{y\in\boldsymbol{Y}} V(\psi,y,\tilde{u},\tilde{v})h^*(T(\psi,y,\tilde{u},\tilde{v}))\right] \\
&= \min_{\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})} \max_{\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})} \left[\hat{c}(\psi,\tilde{u},\tilde{v}) + \sum_{y\in\boldsymbol{Y}} V(\psi,y,\tilde{u},\tilde{v})h^*(T(\psi,y,\tilde{u},\tilde{v}))\right] .
\end{aligned}$$

We now show that $\rho^*$ is the value of the game and any pair of policies $(\pi^{1*},\pi^{2*})$ satisfying

$$\begin{aligned}
&\rho^* + h^*(\psi) \\
&= \min_{\tilde{v}\in\boldsymbol{\mathcal{P}}(\boldsymbol{V})} \left[\hat{c}(\psi,\pi^{1*}(\psi),\tilde{v}) + \sum_{y\in\boldsymbol{Y}} V(\psi,y,\pi^{1*}(\psi),\tilde{v})\cdot h^*(T(\psi,y,\pi^{1*}(\psi),\tilde{v}))\right] \\
&= \max_{\tilde{u}\in\boldsymbol{\mathcal{P}}(\boldsymbol{U})} \left[\hat{c}(\psi,\tilde{u},\pi^{2*}(\psi)) + \sum_{y\in\boldsymbol{Y}} V(\psi,y,\tilde{u},\pi^{2*}(\psi))\cdot h^*(T(\psi,y,\tilde{u},\pi^{2*}(\psi)))\right]
\end{aligned} \tag{3.6}$$

is the saddle-point equilibrium. Since given any admissible $\pi^1 \in \Pi^1$ and initial condition $\psi_0 \in \boldsymbol{\Psi}$, from (3.6) we have for each $t \in \mathbb{N}_0$

$$\rho^* + h^*(\psi_t) - \hat{c}(\psi_t, \tilde{u}_t, \pi^{2*}(\psi_t)) \geq \mathbb{E}_{\psi_t}^{\pi^1 \pi^{2*}}[h^*(\psi_{t+1})]\,.$$

Adding the terms on both side of the inequality from $t = 0$ to $t = T-1$ we obtain

$$\begin{aligned} T\rho^* + \sum_{t=0}^{T-1} h^*(\psi_{t+1}) + h^*(\psi_0) - h^*(\psi_T) - \sum_{t=0}^{T-1} \hat{c}(\psi_t, \tilde{u}_t, \pi^{2*}(\psi_t)) \\ \geq \sum_{t=0}^{T-1} \mathbb{E}_{\psi_t}^{\pi^1, \pi^{2*}}[h^*(\psi_{t+1})]\,. \end{aligned}$$

Now on both sides we take the expectation $\mathbb{E}_{\psi_0}^{\pi^1, \pi^{2*}}$, multiply $1/T$ and let $T \to \infty$. Note that $h^*(\cdot)$ is bounded due to its continuity on the compact $\boldsymbol{\Psi}$. We thus have

$$\rho^* \geq \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_{\psi_0}^{\pi^1 \pi^{2*}} \left[ \sum_{t=0}^{T-1} \tilde{c}(\psi_t, \tilde{u}_t, \pi^{2*}(\psi_t)) \right] = J(\psi_0, \pi^1, \pi^{2*})\,. \tag{3.7}$$

That is

$$\rho^* \geq J(\psi_0, \pi^1, \pi^{2*})$$

and thus

$$\rho^* \geq \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^1} J(\psi_0, \pi^1, \pi^2)\,.$$

Similarly We can show

$$\rho^* \leq J(\psi_0, \pi^{1*}, \pi^2)$$

and thus

$$\rho^* \leq \sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} J(\psi_0, \pi^1, \pi^2)\,.$$

Hence $\rho^*$ is the value of the game and $\pi^{i*}$ is the optimal policy for controller $i$, $i = 1, 2$. That is, $(\pi^{1*}, \pi^{2*})$ is the saddle-point equilibrium. □

# Chapter 4

# Adaptive Control of Partially Observed Markov Decision Processes

## 4.1 Introduction and Preliminaries

During the past decade considerable effort has been invested in the study of stochastic adaptive control. Special attention has been paid to systems with incomplete or noisy state observation. This class of problems can be viewed as a generalization of discrete-time partially observable Markov decision processes (POMDP) in the sense that the transition probability matrix depends on some unknown parameter vector $\theta \in \Theta$, where $\Theta \in \mathbb{R}^{N_\theta}$ is the parameter space. The purpose of adaptive control is to optimally regulate the system in the presence of parameter uncertainty. To achieve this, a methodology involving a set of assumptions is proposed in [31]. Its idea can be outlined in the following two steps.

1. Make use of some identification scheme to obtain a sequence $\{\hat{\theta}_t\}_{t=0}^{\infty}$ of estimates of the true parameter $\theta$.

2. Compute the next information state $\hat{\psi}_{t+1}$ by using the most updated estimate $\hat{\theta}_t$ of the parameter $\theta$.

Finally, the policy suggests an action at time $t$ based on the estimates $\hat{\theta}_t$ and $\hat{\psi}_t$. It is shown in [27, Chapter 6] that this adaptive policy reaches some sense

of self optimization. The main assumptions mentioned in [31] include the existence of optimal policies for models parameterized by all $\theta \in \Theta$, and the convergence of the estimates of the information states in an appropriate sense. Since we are interested in the long-run average criterion, ergodic properties of the underlying controlled Markov chain are utilized to guarantee the validity of the above assumptions. In [32], controlled Markov processes satisfying the classical *renewability* condition are handled. In [21], controlled Markov processes with positive transition matrices are studied. Motivated by the conditions leading to a contracted mapping in the *Hilbert metric*, Chuang [17] proposed an assumption on the structure of the transition matrices to yield the desired ergodic property. In this chapter, we propose another assumption on the structure of transition matrices based on ideas of weak ergodicity and obtain similar results. An interesting point is that our assumption, which regulates column elements of products of transition matrices, can be viewed as a companion to Chuang's assumption, which regulates row elements of products of transition matrices. We note that either assumption is much weaker than those of other work in the literature and guarantees the existence of optimal policy. Under our ergodic assumption and a condition on the convergence speed of $\{\hat{\theta}_t\}_{t=0}^{\infty}$ to $\theta$, we are able to justify the convergence of the estimates of the information states. Hence, the self-optimizing property of the adaptive policy is obtained.

For the remainder of this section we introduce notation and concepts that will be used later. The main results are presented in Section 4.2.

When the transition matrix $Q$ is not perfectly known but instead parameterized by a vector $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^{N_\theta}$ is a compact space, a stochastic approximation-type estimation algorithm can be designed to form a sequence $\{\hat{\theta}_t\}_{t=0}^{\infty}$ such that $\{\hat{\theta}_t\}_{t=0}^{\infty} \to \theta$ w.p.1. Related to this convergent sequence we denote the parameterized transition matrix

$$Q(y_t, u_{t-1}) = Q(y_t, u_{t-1}, \theta), \quad \hat{Q}(y_t, u_{t-1}) = Q(y_t, u_{t-1}, \hat{\theta}_t)\,.$$

Suppose an optimal stationary policy corresponding to each parameter $\theta \in \Theta$ exists and is denoted by $\pi^*(\cdot, \theta)$. Define the adaptive policy $\pi^a$ which generates a sequence of actions $\{u_t\}_{t=0}^{\infty}$ according to

$$u_t = \pi^*(\hat{\psi}_t, \hat{\theta}_t) \tag{4.1}$$

where

$$\hat{\psi}_{t+1} := \sum_{y \in \mathbf{Y}} \frac{\hat{\psi}_t \hat{Q}(y, u_t)}{\hat{\psi}_t \hat{Q}(y, u_t)\mathbf{1}} \cdot \mathbf{1}_{\{Y_{t+1}=y\}}\,, \qquad \forall \hat{\psi}_t \in \boldsymbol{\Psi}, u_t \in \boldsymbol{U}, t \in \mathbb{N}_0\,.$$

Note that $\{\hat{\psi}_{t+1}\}_{t=0}^{\infty}$ is the sequence of estimates of the information states. Following the methodology in [31], for $\pi^a$ to be a self-optimizing policy in the sense that

$$J_\theta(\psi_0, \pi^a) = \inf_{\pi \in \Pi} J_\theta(\psi_0, \pi) = \inf_{\pi \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi}_{\psi_0, \theta} \left\{ \sum_{t=0}^{T-1} c(X_t, U_t) \right\}\,,$$

the sequence of estimation errors between the estimated and the true information states should converge to 0. That is

$$\|\hat{\psi}_t - \psi_t\|_1 \longrightarrow 0 \qquad \text{in} \quad \mathbb{P}^{\pi^a}_{\psi_0} \quad \text{as} \quad t \to \infty\,,$$

where $\|V\|_1 := \sum_{i=1}^{N_v} |V(i)|$ for a vector $V \in \mathbb{R}^{N_v}$. We study the conditions that imply this convergence in the next section. Specifically we show under

our assumption that

$$\mathbb{E}^{\pi^a}_{\psi_0}\left[\|\psi_t - \hat{\psi}_t\|_1\right] \longrightarrow 0 \quad \text{as} \quad t \to \infty, \quad \forall \psi_0 \in \boldsymbol{\Psi}\,. \tag{4.2}$$

In order to obtain our results, properties for products of nonnegative matrices need to be explored. Here we collect some concepts and definitions to be used later. A *nonnegative matrix* is a square matrix with all of its elements nonnegative. A *row-allowable matrix* is a nonnegative matrix with all of its row sums positive. A *substochastic matrix* is a nonnegative matrix with all of its row sums no greater than 1. In particular, if every row sum of a substochastic matrix equals 1, it is called a *stochastic matrix*. The following definitions are used throughout this section. For a sequence $\{B_k\}_{k=1}^{\infty}$ of row-allowable $N_x \times N_x$ matrices, $B_m^{m+n} := B_m B_{m+1} \cdots B_{m+n}$, and $B^k := B_1 B_2 \cdots B_k$. $B_{i\cdot}$, $B_{\cdot j}$ represent the $i^{th}$ row and $j^{th}$ column of $B$, respectively. $\varepsilon \in (0,1)$ is some positive number.

## 4.2 Main Result

Our assumption dealing with the convergence issue of the information state process is mainly based on the concept of the weak ergodicity for products of nonnegative matrices (see [70]). In this section we first show some properties for two classes of row-allowable matrices, and then apply these properties to the derivation of our main result.

**Lemma 4.2.1.** *If $B_m$ is row-allowable, $(B_m)_{i_1 j} \geq \varepsilon \cdot (B_m)_{i_2 j}$ for each $i_1, i_2, j \in \{1, 2, \ldots, N_x\}$, and $m \in \mathbb{N}$, then we have*

$$\frac{1}{(B_m^{m+r})_{i_2 \cdot}\mathbf{1}} \geq \varepsilon \cdot \frac{1}{(B_m^{m+r})_{i_1 \cdot}\mathbf{1}}$$

*for each* $i_1, i_2 \in \{1, 2, \ldots, N_x\}$ *and* $m, r \in \mathbb{N}$.

*Proof.* See Appendix B.1. □

**Corollary 4.2.2.** *Under the assumption of Lemma 4.2.1 we have*

$$\frac{(B_m)_{i_2 j}(B_{m+1}^{m+n})_{j\cdot}\mathbf{1}}{(B_m^{m+r})_{i_2\cdot}\mathbf{1}} \geq \varepsilon^2 \frac{(B_m)_{i_1 j}(B_{m+1}^{m+n})_{j\cdot}\mathbf{1}}{(B_m^{m+r})_{i_1\cdot}\mathbf{1}},$$

*and for each* $\psi_1, \psi_2 \in \boldsymbol{\Psi}$

$$\frac{\psi_1 B^n \mathbf{1}}{\psi_2 B^n \mathbf{1}} \geq \varepsilon.$$

**Lemma 4.2.3.** *If* $B_k$ *is row-allowable with* $(B_k)_{i_1\cdot}\mathbf{1} \geq \varepsilon \cdot (B_k)_{i_2\cdot}\mathbf{1}$ *for each* $i_1, i_2 \in \{1, 2, \ldots, N_x\}$ *and* $k \in \mathbb{N}$, *then we have*

$$\frac{B_{i_1\cdot}^k \mathbf{1}}{B_{i_2\cdot}^k \mathbf{1}} \geq \varepsilon^k$$

*for each* $i_1, i_2 \in \{1, 2, \ldots, N_x\}$ *and* $k \in \mathbb{N}$.

*Proof.* See Appendix B.2. □

**Lemma 4.2.4.** *If* $(B_m)_{i_1 j} \geq \varepsilon \cdot (B_m)_{i_2 j}$ *for each* $i_1, i_2, j \in \{1, 2, \ldots, N_x\}$ *and* $m \in \mathbb{N}$, *then we have*

$$\left| \frac{(B_m^{m+n})_{i_1 j}}{(B_m^{m+n})_{i_1\cdot}\mathbf{1}} - \frac{(B_m^{m+n})_{i_2 j}}{(B_m^{m+n})_{i_2\cdot}\mathbf{1}} \right| \leq (1 - \varepsilon^2)^n$$

*for each* $i_1, i_2, j \in \{1, 2, \ldots, N_x\}$ *and* $m, n \in \mathbb{N}$.

*Proof.* See Appendix B.3. □

It is shown in [5] that if $A$ is a nonnegative matrix, $\psi_1, \psi_2 \in \boldsymbol{\Psi}$, and

$$\tau_1(A) := \frac{1}{2} \max_{i,j} \|A_{i\cdot} - A_{j\cdot}\|_1,$$

then we have

$$\|(\psi_1 - \psi_2)A\|_1 \leq \tau_1(A) \|\psi_1 - \psi_2\|_1 .$$

Apparently if $A$ is a stochastic matrix, then $\tau(A) \in [0, 1]$. If $\tau_1(A) \in (0, 1)$ then $A$ has a contraction property.

**Lemma 4.2.5.** *For each $i_1, i_2, j \in \boldsymbol{X}$ and $k \in \mathbb{N}$,*

1. *If $\psi_1, \psi_2 \in \boldsymbol{\Psi}$ and $B_k$ is row-allowable with $(B_k)_{i_1 \cdot}\mathbf{1} \geq \varepsilon \cdot (B_k)_{i_2 \cdot}\mathbf{1}$ then*

$$\left\| \frac{\psi_1 B^n}{\psi_1 B^n \boldsymbol{1}} - \frac{\psi_2 B^n}{\psi_2 B^n \boldsymbol{1}} \right\|_1 \leq \frac{2}{\varepsilon^n} \|\psi_1 - \psi_2\|_1 .$$

2. *If $\psi_1, \psi_2 \in \boldsymbol{\Psi}$, and $B_k$ is row-allowable with $(B_k)_{i_1 j} \geq \varepsilon \cdot (B_k)_{i_2 j}$ then*

$$\left\| \frac{\psi_1 B^n}{\psi_1 B^n \boldsymbol{1}} - \frac{\psi_2 B^n}{\psi_2 B^n \boldsymbol{1}} \right\|_1 \leq \frac{N_x}{\varepsilon}(1 - \varepsilon^2)^{n-1} \|\psi_1 - \psi_2\|_1 .$$

*Proof.* See Appendix B.4. □

Now we propose the main assumption based on Lemma 4.2.5.

**Assumption 4.2.1.** For each parameter $\theta \in \Theta$, $Q(y, u, \theta)$ is row-allowable for each $y \in \boldsymbol{Y}$ and $u \in \boldsymbol{U}$. Also, there exist constants $\varepsilon > 0$, $N_0 \in \mathbb{N}$ such that

$$\max_{1 \leq k \leq N_0} \mathbb{P}_{\psi_0}^{\pi^a}\{Q(Y^k, U^{k-1}, \theta)_{i_1 j} \geq \varepsilon \cdot Q(Y^k, U^{k-1}, \theta)_{i_2 j}, 1 \leq i_1, i_2, j \leq N_x\} = 1$$

holds for all $\psi_0 \in \boldsymbol{\Psi}$, where $\pi^a$ is the adaptive strategy defined in (4.1) and $Q(Y^k, U^{k-1}, \theta)$ is a $k$-step transition kernel parameterized by $\theta$.

Suppose Assumption 4.2.1 holds. Then there exists an increasing sequence of integers $\{m_l\}_{l=0}^n \subset \mathbb{N}_0$ satisfying $m_0$=0, $m_l - m_{l-1} \leq N_0$ for $l =$

$1, 2, \ldots, n$, $m_n \leq t$, such that

$$\mathbb{P}^{\pi^a}_{\psi_0}\{[Q(Y_{m_{l-1}+1}, U_{m_{l-1}}) \cdots Q(Y_{m_l}, U_{m_l-1})]_{i_1 j} \geq \varepsilon[Q(Y_{m_{l-1}+1}, U_{m_{l-1}}) \cdots Q(Y_{m_l}, U_{m_l-1})]_{i_2 j}\} = 1 \tag{4.3}$$

for each $i_1, i_2, j \in \boldsymbol{X}$. Let $\{\hat{\theta}_t\}_{t=0}^{\infty}$, $\hat{\theta}_t \in \Theta$, be a sequence of estimates of $\theta$ and satisfies $\hat{\theta}_t = \hat{\theta}_{m_l+1}$ for $m_l + 1 \leq t \leq m_{l+1}$. That is, $\hat{\theta}_t$ is updated only at time $t = m_l + 1$, $l \in \mathbb{N}_0$. The following two conditions similar to those in [17] on the properties of the sequence $\{\hat{\theta}_t\}_{t=0}^{\infty}$ are necessary to obtain our result.

**Assumption 4.2.2.** The parameter space $\Theta$ is compact and the transition matrix $Q(y, u)$ is continuously differentiable on $\Theta$ for every $y \in \boldsymbol{Y}$ and $u \in \boldsymbol{U}$.

**Assumption 4.2.3.** The sequence $\{\hat{\theta}_t\}_{t=0}^{\infty}$ of estimates of $\theta$ satisfies:

1. $\hat{\theta}_t$ is $\sigma(Y_0, \ldots, Y_t)-$measurable .
2. $\hat{\theta}_t \to \theta$ as $t \to \infty$ in $\mathbb{P}^{\pi^a}_{\psi_0}$ .
3. there exists a constant $M$ such that $\|\hat{\theta}_{m_{l+1}+1} - \hat{\theta}_{m_l+1}\|_1 \leq \frac{M}{l+1}$ for every $l \in \mathbb{N}_0$ .

Now we are ready for the following result.

**Theorem 4.2.6.** *Suppose Assumption 2.1.1 and Assumption 4.2.1 $\sim$ 4.2.3 are satisfied, then for each $\psi_0 \in \boldsymbol{\Psi}$*

$$\mathbb{E}^{\pi^a}_{\psi_0}\left[\|\psi_t - \hat{\psi}_t\|_1\right] \longrightarrow 0 \qquad as \quad t \to \infty .$$

*Proof.* By Assumption 4.2.1 there exists a set of integers $\{m_0, m_1, \ldots, m_n\}$ such that (4.3) holds. Define products of the transition kernels:

$$B_i^{y^t} := \begin{cases} I & i = 0 \\ Q(Y_{m_{i-1}+1}, U_{m_{i-1}}) \cdots Q(Y_{m_i}, U_{m_i-1}) & i = 1, \ldots, n \\ Q(Y_{m_{i-1}+1}, U_{m_{i-1}}) \cdots Q(Y_t, U_{t-1}) & i = n+1 \end{cases} \tag{4.4}$$

where $I$ is the identity matrix with size $N_x \times N_x$. $\hat{B}_i^{y^t}$ is defined similarly to (4.4) with $Q(Y,U)$ replaced by $\hat{Q}(Y,U)$ for $i = 1, \ldots, n+1$. Also, $\hat{B}_0^{y^t} = I$. For $l = 1, \ldots, n+1$ new information states are denoted by

$$\tilde{\psi}_l := \frac{\psi_0 \hat{B}_1^{y^t} \cdots \hat{B}_{l-1}^{y^t}}{\psi_0 \hat{B}_1^{y^t} \cdots \hat{B}_{l-1}^{y^t} \mathbf{1}}, \quad \hat{\tilde{\psi}}_l := \frac{\tilde{\psi}_l \hat{B}_l^{y^t}}{\tilde{\psi}_l \hat{B}_l^{y^t} \mathbf{1}}, \quad \overline{\hat{\tilde{\psi}}}_l := \frac{\tilde{\psi}_l B_l^{y^t}}{\tilde{\psi}_l B_l^{y^t} \mathbf{1}}.$$

Then by triangular inequality we have with probability 1

$$\begin{aligned} &\left\| \psi_t - \hat{\psi}_t \right\|_1 \\ &\leq \sum_{l=1}^{n+1} \left\| \frac{\psi_0 \hat{B}_0^{y^t} \cdots \hat{B}_{l-1}^{y^t} B_l^{y^t} \cdots B_{n+1}^{y^t}}{\psi_0 \hat{B}_0^{y^t} \cdots \hat{B}_{l-1}^{y^t} B_l^{y^t} \cdots B_{n+1}^{y^t} \mathbf{1}} - \frac{\psi_0 \hat{B}_1^{y^t} \cdots \hat{B}_l^{y^t} B_{l+1}^{y^t} \cdots B_{n+1}^{y^t}}{\psi_0 \hat{B}_1^{y^t} \cdots \hat{B}_l^{y^t} B_{l+1}^{y^t} \cdots B_{n+1}^{y^t} \mathbf{1}} \right\|_1 \\ &= \sum_{l=1}^{n} \left\| \frac{\overline{\hat{\tilde{\psi}}}_l B_{l+1}^{y^t} \cdots B_{n+1}^{y^t}}{\overline{\hat{\tilde{\psi}}}_l B_{l+1}^{y^t} \cdots B_{n+1}^{y^t} \mathbf{1}} - \frac{\hat{\hat{\tilde{\psi}}}_l B_{l+1}^{y^t} \cdots B_{n+1}^{y^t}}{\hat{\hat{\tilde{\psi}}}_l B_{l+1}^{y^t} \cdots B_{n+1}^{y^t} \mathbf{1}} \right\|_1 + \left\| \overline{\hat{\psi}}_{n+1} - \hat{\hat{\psi}}_{n+1} \right\|_1 . \end{aligned} \tag{4.5}$$

If Assumption 4.2.1 is satisfied, then by Lemma 4.2.5 we have

$$\begin{aligned} (4.5) &\leq \frac{2}{\varepsilon^{N_0}} \left\{ \sum_{l=1}^{n-1} \left\| \frac{\overline{\hat{\tilde{\psi}}}_l B_{l+1}^{y^t} \cdots B_n^{y^t}}{\overline{\hat{\tilde{\psi}}}_l B_{l+1}^{y^t} \cdots B_n^{y^t} \mathbf{1}} - \frac{\hat{\hat{\tilde{\psi}}}_l B_{l+1}^{y^t} \cdots B_n^{y^t}}{\hat{\hat{\tilde{\psi}}}_l B_{l+1}^{y^t} \cdots B_n^{y^t} \mathbf{1}} \right\|_1 + \sum_{l=n}^{n+1} \left\| \overline{\hat{\tilde{\psi}}}_l - \hat{\hat{\tilde{\psi}}}_l \right\|_1 \right\} \\ &\leq \frac{2}{\varepsilon^{N_0}} \left\{ \sum_{l=1}^{n-1} \frac{N_x}{\varepsilon} (1-\varepsilon^2)^{n-1-l} \left\| \overline{\hat{\tilde{\psi}}}_l - \hat{\hat{\tilde{\psi}}}_l \right\|_1 + \sum_{l=n}^{n+1} \left\| \overline{\hat{\tilde{\psi}}}_l - \hat{\hat{\tilde{\psi}}}_l \right\|_1 \right\} \\ &= \frac{2N_x}{\varepsilon^{N_0+1}(1-\varepsilon^2)} \left\{ \sum_{l=1}^{n-1} (1-\varepsilon^2)^{n-l} \left\| \overline{\hat{\tilde{\psi}}}_l - \hat{\hat{\tilde{\psi}}}_l \right\|_1 + \sum_{l=n}^{n+1} \left\| \overline{\hat{\tilde{\psi}}}_l - \hat{\hat{\tilde{\psi}}}_l \right\|_1 \right\} \\ &= \frac{2N_x}{\varepsilon^{N_0+1}(1-\varepsilon^2)} \left\{ \sum_{l=1}^{n} (1-\varepsilon^2)^{n-l} \left\| \overline{\hat{\tilde{\psi}}}_l - \hat{\hat{\tilde{\psi}}}_l \right\|_1 + \left\| \overline{\hat{\tilde{\psi}}}_{n+1} - \hat{\hat{\tilde{\psi}}}_{n+1} \right\|_1 \right\} . \end{aligned} \tag{4.6}$$

Due to the differentiability of $Q(y,u,\cdot)$ on $\Theta$ for each $y \in \boldsymbol{Y}$ and $u \in \boldsymbol{U}$, and, for $k = m_l + 1, \ldots, m_{l+1}$ $\hat{\theta}_k = \hat{\theta}_{m_l+1}$, we can apply the mean value theorem and write that for $l = 1, \ldots, n$ there exists a finite positive constant $M$ such that

$$\left\|\overline{\hat{\psi}}_l - \hat{\bar{\psi}}_l\right\|_1 \leq M \sum_{k=m_{l-1}+1}^{m_l} \left\|\hat{\theta}_k - \theta\right\|_1 \leq MN_0 \left\|\hat{\theta}_{m_{l-1}+1} - \theta\right\|_1 .$$

Also,

$$\left\|\overline{\hat{\psi}}_{n+1} - \hat{\bar{\psi}}_{n+1}\right\|_1 \leq M \sum_{k=m_n+1}^{t} \left\|\hat{\theta}_k - \theta\right\|_1 \leq MN_0 \left\|\hat{\theta}_{m_n+1} - \theta\right\|_1 .$$

Therefore, following (4.6) there exist finite numbers $M_1 > 0$, $M_2 > 0$ and $\alpha \in (0,1)$ such that

$$\begin{aligned}\left\|\psi_t - \hat{\psi}_t\right\|_1 &\leq M_1 \sum_{l=1}^{n} \alpha^{n-l} \left\|\hat{\theta}_{m_{l-1}+1} - \theta\right\|_1 + M_2 \left\|\hat{\theta}_{m_n+1} - \theta\right\|_1 \\ &= M_1 \sum_{l=0}^{n-1} \alpha^{n-1-l} \left\|\hat{\theta}_{m_l+1} - \theta\right\|_1 + M_2 \left\|\hat{\theta}_{m_n+1} - \theta\right\|_1 . \qquad (4.7)\end{aligned}$$

Applying the triangular inequality again we obtain for $l = 0, 1, \ldots, n-1$

$$\begin{aligned}\left\|\hat{\theta}_{m_l+1} - \theta\right\|_1 &\leq \left\|\hat{\theta}_{m_n+1} - \theta\right\|_1 + \left\|\hat{\theta}_{m_{n-1}+1} - \hat{\theta}_{m_n+1}\right\|_1 + \cdots + \left\|\hat{\theta}_{m_l+1} - \hat{\theta}_{m_{l+1}+1}\right\|_1 \\ &= \left\|\hat{\theta}_{m_n+1} - \theta\right\|_1 + \sum_{i=l}^{n-1} \left\|\hat{\theta}_{m_{i+1}+1} - \hat{\theta}_{m_i+1}\right\|_1 \\ &\leq \left\|\hat{\theta}_{m_n+1} - \theta\right\|_1 + \sum_{i=l}^{n-1} \frac{\overline{M}}{i+1} \qquad (4.8)\end{aligned}$$

where the last inequality follows from Assumption 4.2.3-(3). On the other hand,

$$\begin{aligned}\sum_{l=0}^{n-1} \sum_{i=l}^{n-1} \frac{\alpha^{n-1-l}}{i+1} &= \frac{1}{1-\alpha} \left\{ \sum_{i=0}^{n-1} \frac{\alpha^i}{n-i} - \alpha^n \sum_{i=1}^{n} \frac{1}{i} \right\} \leq \frac{1}{1-\alpha} \sum_{i=0}^{n-1} \frac{\alpha^i}{n-i} \\ &\leq \frac{1}{1-\alpha} \sum_{i=0}^{n-1} \frac{1+i}{n} \alpha^i \leq \frac{1}{n(1-\alpha)^3} \qquad (4.9)\end{aligned}$$

where the second inequality follows from that for $0 \leq i \leq n-1$ we have

$$\frac{n}{n-i} = 1 + \frac{i}{n-i} \leq 1 + i\,.$$

Finally we obtain from (4.7) $\sim$ (4.9) that there exist finite numbers $M_3 > 0$, $M_4 > 0$ such that

$$\mathbb{E}^{\pi^a}_{\psi_0}\left[\|\psi_t - \hat{\psi}_t\|_1\right] \leq M_3 \mathbb{E}^{\pi^a}_{\psi_0}\left[\|\hat{\theta}_{m_n+1} - \theta\|_1\right] + \frac{M_4}{n}\,. \tag{4.10}$$

As $t \to \infty$, $n \geq t/N_0 \to \infty$, we conclude that for each $\psi_0 \in \boldsymbol{\Psi}$

$$\mathbb{E}^{\pi^a}_{\psi_0}\left[\|\psi_t - \hat{\psi}_t\|_1\right] \longrightarrow 0 \qquad \text{as} \quad t \to \infty$$

by Assumption 4.2.3-(2), the compactness of $\Theta$, and inequality (4.10). □

In the following, we will show the self-optimizing property of $\pi^a$.

**Theorem 4.2.7.** *For the parameterized POMDP model, suppose the action space $\boldsymbol{U}$ is finite. Also, Assumption 2.1.1, Assumption 4.2.1 $\sim$ 4.2.3 hold. Then for a given unknown true parameter vector $\theta \in \Theta$ the adaptive policy $\pi^a$ defined in (4.1) is self-optimizing with respect to the long-run average cost criterion.*

*Proof.* By Theorem 2.2.3, under the assumptions we have for each $\theta \in \Theta$ a bounded solution $(\rho_\theta, h_\theta)$ for equation (2.5) in parameterized form. Furthermore $h_\theta(\psi)$ is continuous and bounded both in $\psi \in \boldsymbol{\Psi}$ and $\theta \in \Theta$. Define the discrepancy function

$$D_\theta(\psi, u) := \tilde{c}(\psi, u) + \int_{\boldsymbol{Y}} h_\theta(\psi')\mathcal{K}(d\psi'|\psi, u, \theta) - \rho_\theta - h_\theta(\psi)\,.$$

It is not difficult to see that $D_\cdot(\cdot, u)$ is uniformly continuous and bounded in $\Theta \times \boldsymbol{\Psi}$ for each $u \in \boldsymbol{U}$. Assumption 4.2.3 together with Theorem 4.2.6 imply for each $u \in \boldsymbol{U}$

$$D_{\hat{\theta}_t}(\hat{\psi}_t, u) \longrightarrow D_\theta(\psi_t, u) \quad \text{in} \quad \mathbb{P}_{\psi_0}^{\pi^a} \quad \text{as} \quad t \to \infty\,.$$

Due to the finiteness of $\boldsymbol{U}$ we can write

$$D_{\hat{\theta}_t}(\hat{\psi}_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t)) \longrightarrow D_\theta(\psi_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t))) \quad \text{in} \quad \mathbb{P}_{\psi_0}^{\pi^a} \quad \text{as} \quad t \to \infty\,. \tag{4.11}$$

Since $\pi^*(\cdot, \theta)$ is optimal for $\theta \in \Theta$ we have

$$D_{\hat{\theta}_t}(\hat{\psi}_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t)) = 0\,.$$

Define for arbitrary $\varepsilon > 0$ and $t \in \mathbb{N}$

$$\begin{aligned}\Omega_t(\varepsilon) &:= \{\omega : |D_\theta(\psi_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t)) - D_{\hat{\theta}_t}(\hat{\psi}_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t))|(\omega) > \varepsilon\} \\ &= \{\omega : D_\theta(\psi_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t))(\omega) > \varepsilon\}\,.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}_{\psi_0}^{\pi^a}\{D_\theta(\psi_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t))\} &= \int_{\Omega_t(\varepsilon)} D_\theta(\psi_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t)) d\mathbb{P}_{\psi_0}^{\pi^a} \\ &\qquad + \int_{\Omega \setminus \Omega_t(\varepsilon)} D_\theta(\psi_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t)) d\mathbb{P}_{\psi_0}^{\pi^a} \\ &\leq K\mathbb{P}_{\psi_0}^{\pi^a}(\Omega_t(\varepsilon)) + \varepsilon\end{aligned}$$

for some finite $K > 0$. By (4.11) and letting $\varepsilon \to 0$, we have

$$\mathbb{E}_{\psi_0}^{\pi^a}\{D_\theta(\psi_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t))\} \longrightarrow 0 \qquad \text{as} \quad t \to \infty\,.$$

Therefore,

$$\frac{1}{T}\sum_{t=0}^{N-1} \mathbb{E}_{\psi_0}^{\pi^a}\{D_\theta(\psi_t, \pi^*(\hat{\psi}_t, \hat{\theta}_t))\} \longrightarrow 0 \qquad \text{as} \quad t \to \infty$$

and the result follows from Theorem 2.2.3. □

# Chapter 5

# Safety Control

## 5.1 Introduction

In this chapter we study the concept of safety control of stochastic discrete event systems (DESs), which is an extension of the idea of safety control of non-discrete DESs [63]. A non-stochastic DES is often modelled by a state machine or an automaton (see [37, 40, 49] for other models) that evolves in response to the occurrence of events. Events are categorized into a controllable class, which can be controlled by an external agent, and a uncontrollable class. The controller dynamically disables controllable events so that closed-loop behavior satisfies the control goal. The goal of safety control is normally specified by a set of forbidden states that the system must avoid. Thus, a controller performing a safety control must prevent the system from visiting those pre-specified states. The concept of safety control of non-stochastic DESs is generalized naturally to the safety control of stochastic DESs modelled by discrete-time complete observed Markov decision processes introduced in Section 1.3. The first result with this setting is in [4], where the safety is specified by an upper bound on the system's state probability distribution vector. Hence, under the safety control, the system's probability in visiting each of its state at each time step is bounded above. If the safety specification assigns 1 as an upper bound for some state, then it is equivalent

to saying that the system's probability in visiting that state is not limited. Since the objective of safety control in stochastic DESs requires the system to meet the safety specification at each time step, the concept of the safety control is much more conservative (restrictive) than the traditional *constrained Markov decision process*, which only regulates the system's long-term behavior (e.g., see [1]).

We extend in this chapter the concept of safety control of stochastic DESs by specifying the safety control objective as a convex set in which the system's state probability distribution vector must lie at each time step. To do this, an appropriate choice of a policy and an initial state probability distribution is necessary. A probability distribution is called *safe* if it is in the convex set. When some admissible policy is performed, if there exists a subset of that convex set such that any element that is inside the subset and serves as an initial probability distribution of the system induces a sequence of state probability distributions that are always safe, then that admissible policy is called a *safe policy*. In the following sections we first analyze conditions for an admissible policy such that the controlled system meets the safety specification; that is, if the initial state probability distribution is safe, the state probability distribution under that policy remains safe at each time step. A state feedback controller that performs this kind of policy is called a *safety enforcing controller*. The necessary and sufficient condition for a controller to be safety enforcing is obtained for a special case of safety specification. Next, since the safe policy, if it exists, is not unique in general, we apply linear programming techniques to search for a safe policy that is optimal in terms of the given cost function and in the sense of a long-run average. The feasibility

of our linear programming formulation is also discussed. When a safe policy is available, we study the set of initial probability distributions corresponding to that safe policy such that the system meets the safety specification as a result of starting the system from an element in that set and performing the safe policy. The set owning that property is normally not unique. We provide algorithms to characterize the largest one and show that these algorithms terminate in finite steps under mild conditions. In particular, we prove that the algorithms terminate in one iteration when the system has only two states. Finally we make remarks on the main assumption that implies the existence of safe policies and the finite termination of our algorithms.

## 5.2 Preliminaries and Notation

We consider a system modelled by a discrete-time completely observable Markov decision process with finite system space $\boldsymbol{X} = \{1, 2, \ldots, N_x\}$. Let $\boldsymbol{\Psi} = \boldsymbol{\mathcal{P}}(\boldsymbol{X})$ be the space of probability distributions on $\boldsymbol{X}$. $P_\pi$ is the transition matrix of the system when the policy is $\pi$. $P_\pi$ is called *irreducible* if for each $i, j \in \boldsymbol{X}$ there exists a $n_{ij} \in \mathbb{N}$ such that $(P_\pi^{n_{ij}})_{ij} > 0$. The *period* of state $i \in \boldsymbol{X}$ under policy $\pi$ is the greatest common divider of the set $\{n|(P_\pi^n)_{ii} > 0\}$. $P_\pi$ is called *aperiodic* if the period for all of its states is 1. $\hat{\boldsymbol{\Psi}} \subset \boldsymbol{\Psi}$ is called an *invariant set of distribution* of $P_\pi$ if $\psi \in \hat{\boldsymbol{\Psi}}$ implies $\psi P_\pi \in \hat{\boldsymbol{\Psi}}$. In particular, $\psi^* \in \boldsymbol{\Psi}$ is called an *invariant distribution* of $P_\pi$ if $\psi^* P_\pi = \psi^*$. It is well known that an aperiodic and irreducible transition matrix has a unique invariant distribution. A safety specification $\boldsymbol{\Psi}_s$ is a subset of $\boldsymbol{\Psi}$ where

$$\boldsymbol{\Psi}_s := \{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A} \leq \boldsymbol{b}\}, \tag{5.1}$$

$\boldsymbol{b} \in \mathbb{R}^{N_b}$, $N_b \in \mathbb{N}$, and $\boldsymbol{A} \in \mathbb{R}^{N_x \times N_b}$. We say that a probability distribution $\psi \in \boldsymbol{\Psi}$ is safe if $\psi \in \boldsymbol{\Psi}_s$. For an admissible policy $\pi \in \Pi$, if there exists an associated set $\boldsymbol{\Psi}_{in} \subset \boldsymbol{\Psi}_s$ such that

$$\psi \in \boldsymbol{\Psi}_{in} \Rightarrow \psi^{(k)} = \psi P_\pi^k \in \boldsymbol{\Psi}_s \qquad \forall k \in \mathbb{N}\,, \tag{5.2}$$

then this policy $\pi$ is called a safe policy corresponding to the $\boldsymbol{\Psi}_s$. The objective of safety control is to find an safe policy $\pi \in \Pi$ and an associated set $\boldsymbol{\Psi}_{in} \subset \boldsymbol{\Psi}_s$ such that (5.2) holds.

## 5.3 Safety Enforcing Controller

Given a safety specification $\boldsymbol{\Psi}_s = \{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A} \leq \boldsymbol{b}\}$ as defined in (5.1), a controller is safety enforcing if it is induced by a safe policy $\pi$ such that

$$\psi \in \boldsymbol{\Psi}_s \quad \Rightarrow \quad \psi^{(k)} = \psi P_\pi^k \in \boldsymbol{\Psi}_s \qquad \forall k \in \mathbb{N}\,. \tag{5.3}$$

That is, $\boldsymbol{\Psi}_{in}$ in (5.2) equals $\boldsymbol{\Psi}_s$ under a safety enforcing controller. We analyze in Section 5.3.1 the condition for the existence of a safety enforcing controller. Let $I$ be the identity matrix with size $N_x \times N_x$. A special case when $\boldsymbol{A}$ in the safety specification (5.1) equals $[I \ -I]^T$ is discussed in Section 5.3.2.

### 5.3.1 General Form

In the consideration of the existence and characterization of a safety enforcing controller, the following analysis provides a dual problem. Suppose there exist a nonnegative $\boldsymbol{U} \in \mathbb{R}^{N_b \times N_b}$ and a row vector $\boldsymbol{v} \in \mathbb{R}^{N_b}$ such that

$$P_\pi \boldsymbol{A} \leq \boldsymbol{A}\boldsymbol{U} + \boldsymbol{1}\boldsymbol{v}\,, \quad \boldsymbol{b}\boldsymbol{U} + \boldsymbol{v} \leq \boldsymbol{b}\,, \tag{5.4}$$

where $\mathbf{1} \in \mathbb{R}^{N_x}$ is a column vector of 1's. By the nonnegativity of $\boldsymbol{U}$ and $\psi$, as well as inequalities in (5.4), we have

$$\psi \in \boldsymbol{\Psi}_s \quad \Rightarrow \quad \psi P_\pi \boldsymbol{A} \leq \psi \boldsymbol{A}\boldsymbol{U} + \psi \mathbf{1}\boldsymbol{v} \leq \boldsymbol{b}\boldsymbol{U} + \boldsymbol{v} \leq \boldsymbol{b}.$$

Therefore (5.3) holds under the policy $\pi$.

We can also formulate a set of linear programming problems to discuss the sufficient and necessary conditions for (5.3) to hold. Consider

$$\begin{aligned} \zeta_i := \max_{\psi \in \boldsymbol{\Psi}} \quad & \psi (P_\pi \boldsymbol{A})_{\cdot i} \\ \text{subject to} \quad & \psi \boldsymbol{A} \leq \boldsymbol{b} \end{aligned} \tag{5.5}$$

where $(\boldsymbol{D})_{\cdot j}$ is the $j^{th}$ column of matrix $\boldsymbol{D}$. It is easy to see that (5.3) holds if and only if $\zeta_i \leq \boldsymbol{b}(i)$ for every $i \in \boldsymbol{X}$.

### 5.3.2 Special Case

In this section we consider a special case of $\boldsymbol{A}$. Let $I$ be the identity matrix with size $N_x \times N_x$. If

$$\boldsymbol{A} = \begin{bmatrix} I \\ -I \end{bmatrix}, \ \boldsymbol{b} = \begin{bmatrix} \overline{b} \\ -\underline{b} \end{bmatrix},$$

then we write

$$\boldsymbol{\Psi}_s = \boldsymbol{\Psi}(\underline{b}, \overline{b}) := \{\psi \in \boldsymbol{\Psi} | \underline{b} \leq \psi \leq \overline{b}\},$$

where we assume that the components of $\underline{b}$ and $\overline{b}$ lie in $[0\ 1]$. In this case, $\zeta_i$ in (5.5) can be obtained by identifying the $\psi$ in extreme points of the feasible region $\boldsymbol{\Psi}(\underline{b}, \overline{b})$ with a permutation argument (see the proof of Theorem 5.3.1). So the necessary and sufficient condition for (5.3) to hold can be easily

characterized. Specifically, let

$$p^j = [p^{(j)}(1) \quad p^{(j)}(2) \quad \cdots \quad p^{(j)}(N_x)]^T$$

be the $j$th column of $P_\pi$ and $\sigma_j$ be a permutation of $\{1, 2, \ldots, N_x\}$ such that for each $j \in \boldsymbol{X}$, $p^{(j)}(\sigma_j(i))$ is decreasing as $i$ increases. That is

$$p^{(j)}(\sigma_j(1)) \geq p^{(j)}(\sigma_j(2)) \geq \cdots \geq p^{(j)}(\sigma_j(N_x)) \qquad \forall j \in \boldsymbol{X}.$$

Define $\overline{n}_j \in \boldsymbol{X}$ and $\underline{n}_j \in \boldsymbol{X}$ to be the smallest integers satisfying

$$\sum_{i=1}^{\overline{n}_j} \overline{b}(\sigma_j(i)) + \sum_{i=\overline{n}_j+1}^{N_x} \underline{b}(\sigma_j(i)) \geq 1$$

and

$$\sum_{i=1}^{\underline{n}_j} \underline{b}(\sigma_j(i)) + \sum_{i=\underline{n}_j+1}^{N_x} \overline{b}(\sigma_j(i)) \leq 1$$

for each $j \in \boldsymbol{X}$. Then we have the following results:

**Theorem 5.3.1.** *For all $\psi \in \boldsymbol{\Psi}_s = \boldsymbol{\Psi}(\underline{b}, \overline{b})$, $\psi P_\pi \in \boldsymbol{\Psi}_s$ if and only if for all $j \in \boldsymbol{X}$,*

$$\sum_{i=1}^{\overline{n}_j-1} \overline{b}(\sigma_j(i))p^{(j)}(\sigma_j(i)) + \left(1 - \sum_{i=1}^{\overline{n}_j-1} \overline{b}(\sigma_j(i)) - \sum_{i=\overline{n}_j+1}^{N_x} \underline{b}(\sigma_j(i))\right) p^{(j)}(\sigma_j(\overline{n}_j)) + \sum_{i=\overline{n}_j+1}^{N_x} \underline{b}(\sigma_j(i))p^{(j)}(\sigma_j(i)) \leq \overline{b}(j)$$

*and*

$$\sum_{i=1}^{\underline{n}_j-1} \underline{b}(\sigma_j(i))p^{(j)}(\sigma_j(i)) + \left(1 - \sum_{i=1}^{\underline{n}_j-1} \underline{b}(\sigma_j(i)) - \sum_{i=\underline{n}_j+1}^{N_x} \overline{b}(\sigma_j(i))\right) p^{(j)}(\sigma_j(\underline{n}_j)) + \sum_{i=\underline{n}_j+1}^{N_x} \overline{b}(\sigma_j(i))p^{(j)}(\sigma_j(i)) \geq \underline{b}(j)\,.$$

*Proof.* By the definition of $\underline{n}_j$ and $\overline{n}_j$ we have for $j \in \boldsymbol{X}$

$$\sum_{i=1}^{\overline{n}_j-1} \overline{b}(\sigma_j(i)) + \sum_{i=\overline{n}_j}^{N_x} \underline{b}(\sigma_j(i)) < 1 \leq \sum_{i=1}^{\overline{n}_j} \overline{b}(\sigma_j(i)) + \sum_{i=\overline{n}_j+1}^{N_x} \underline{b}(\sigma_j(i))$$

and

$$\sum_{i=1}^{\underline{n}_j} \underline{b}(\sigma_j(i)) + \sum_{i=\underline{n}_j+1}^{N_x} \overline{b}(\sigma_j(i)) \leq 1 < \sum_{i=1}^{\underline{n}_j-1} \underline{b}(\sigma_j(i)) + \sum_{i=\underline{n}_j}^{N_x} \overline{b}(\sigma_j(i)) \,.$$

Equivalently, we have

$$\underline{b}(\sigma_j(\overline{n}_j)) < 1 - \sum_{i=1}^{\overline{n}_j-1} \overline{b}(\sigma_j(i)) - \sum_{i=\overline{n}_j+1}^{N_x} \underline{b}(\sigma_j(i)) \leq \overline{b}(\sigma_j(\overline{n}_j)) \,, \tag{5.6}$$

$$\underline{b}(\sigma_j(\underline{n}_j)) \leq 1 - \sum_{i=1}^{\underline{n}_j-1} \underline{b}(\sigma_j(i)) - \sum_{i=\underline{n}_j+1}^{N_x} \overline{b}(\sigma_j(i)) < \overline{b}(\sigma_j(\underline{n}_j)) \,. \tag{5.7}$$

For $j \in \boldsymbol{X}$, consider $\overline{\psi}^{(j)}$ and $\underline{\psi}^{(j)}$ defined by

$$\overline{\psi}_k^{(j)} := \begin{cases} \overline{b}(k) & \text{if} \quad k \in \{\sigma_j(1), \sigma_j(2), \ldots, \sigma_j(\overline{n}_j - 1)\} \\ 1 - \sum_{i=1}^{\overline{n}_j-1} \overline{b}(\sigma_j(i)) - \sum_{i=\overline{n}_j+1}^{N_x} \underline{b}(\sigma_j(i)) & \text{if} \quad k = \overline{n}_j \\ \underline{b}(k) & \text{otherwise} \end{cases}$$

and

$$\underline{\psi}_k^{(j)} := \begin{cases} \underline{b}(k) & \text{if} \quad k \in \{\sigma_j(1), \sigma_j(2), \ldots, \sigma_j(\underline{n}_j - 1)\} \\ 1 - \sum_{i=1}^{\underline{n}_j-1} \underline{b}(\sigma_j(i)) - \sum_{i=\underline{n}_j+1}^{N_x} \overline{b}(\sigma_j(i)) & \text{if} \quad k = \underline{n}_j \\ \overline{b}(k) & \text{otherwise} \end{cases} .$$

Apparently $\overline{\psi}_k^{(j)}, \underline{\psi}_k^{(j)} \in \boldsymbol{\Psi}_s$ by (5.6) and (5.7). Since for each $\psi \in \boldsymbol{\Psi}_s$, $\psi P_\pi \in \boldsymbol{\Psi}_s$ if and only if for each $j \in \boldsymbol{X}$

$$\max_{\psi \in \boldsymbol{\Psi}_s} (\psi P_\pi)_j \leq \overline{b}(j) \quad \text{and} \quad \min_{\psi \in \boldsymbol{\Psi}_s} (\psi P_\pi)_j \geq \underline{b}(j) \,.$$

It is enough to show that for each $j \in \boldsymbol{X}$

$$\overline{\psi}^{(j)} = \arg\max_{\psi \in \boldsymbol{\Psi}_s} (\psi P_\pi)_j \quad \text{and} \quad \underline{\psi}^{(j)} = \arg\min_{\psi \in \boldsymbol{\Psi}_s} (\psi P_\pi)_j \,.$$

Consider

$$\begin{aligned}
(\psi P_\pi)_j &= \sum_{i=1}^{N_x} \psi(\sigma_j(i)) p^{(j)}(\sigma_j(i)) \\
&= \sum_{i=1}^{\overline{n}_j - 1} \psi(\sigma_j(i)) p^{(j)}(\sigma_j(i)) \qquad (5.8) \\
&\qquad + \left[ 1 - \sum_{i \neq \overline{n}_j} \psi(\sigma_j(i)) \right] p^{(j)}(\sigma_j(\overline{n}_j)) + \sum_{i=\overline{n}_j+1}^{N_x} \psi(\sigma_j(i)) p^{(j)}(\sigma_j(i)) \\
&= \sum_{i=1}^{\underline{n}_j - 1} \psi(\sigma_j(i)) p^{(j)}(\sigma_j(i)) \qquad (5.9) \\
&\qquad + \left[ 1 - \sum_{i \neq \underline{n}_j} \psi(\sigma_j(i)) \right] p^{(j)}(\sigma_j(\underline{n}_j)) + \sum_{i=\underline{n}_j+1}^{N_x} \psi(\sigma_j(i)) p^{(j)}(\sigma_j(i)) \,.
\end{aligned}$$

Since for each $\psi \in \boldsymbol{\Psi}_s$,

$$\begin{aligned}
(5.8) &= p^{(j)}(\sigma_j(\overline{n}_j)) + \sum_{i=1}^{\overline{n}_j - 1} \psi(\sigma_j(i)) \left[ p^{(j)}(\sigma_j(i)) - p^{(j)}(\sigma_j(\overline{n}_j)) \right] \\
&\qquad + \sum_{i=\overline{n}_j+1}^{N_x} \psi(\sigma_j(i)) \left[ p^{(j)}(\sigma_j(i)) - p^{(j)}(\sigma_j(\overline{n}_j)) \right] \\
&\leq p^{(j)}(\sigma_j(\overline{n}_j)) + \sum_{i=1}^{\overline{n}_j - 1} \overline{b}(\sigma_j(i)) \left[ p^{(j)}(\sigma_j(i)) - p^{(j)}(\sigma_j(\overline{n}_j)) \right] \\
&\qquad + \sum_{i=\overline{n}_j+1}^{N_x} \underline{b}(\sigma_j(i)) \left[ p^{(j)}(\sigma_j(i)) - p^{(j)}(\sigma_j(\overline{n}_j)) \right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{\overline{n}_j-1} \overline{b}(\sigma_j(i))p^{(j)}(\sigma_j(i)) + \sum_{i=\overline{n}_j+1}^{N_x} \underline{b}(\sigma_j(i))p^{(j)}(\sigma_j(i)) \\
&\qquad + \left(1 - \sum_{i=1}^{\overline{n}_j-1} \overline{b}(\sigma_j(i)) - \sum_{i=\overline{n}_j+1}^{N_x} \underline{b}(\sigma_j(i))\right) p^{(j)}(\sigma_j(\overline{n}_j)) \\
&= \max_{\psi\in\boldsymbol{\Psi}_s} (\psi P_\pi)_j = \sum_{k=1}^{N_x} \overline{\psi}_k^{(j)} p^{(j)}(k)\,.
\end{aligned}
$$

Similarly for each $\psi \in \boldsymbol{\Psi}_s$,

$$
\begin{aligned}
(5.9) &= p^{(j)}(\sigma_j(\underline{n}_j)) + \sum_{i=1}^{\underline{n}_j-1} \psi(\sigma_j(i)) \left[p^{(j)}(\sigma_j(i)) - p^{(j)}(\sigma_j(\underline{n}_j))\right] \\
&\qquad + \sum_{i=\underline{n}_j+1}^{N_x} \psi(\sigma_j(i)) \left[p^{(j)}(\sigma_j(i)) - p^{(j)}(\sigma_j(\underline{n}_j))\right] \\
&\geq p^{(j)}(\sigma_j(\underline{n}_j)) + \sum_{i=1}^{\underline{n}_j-1} \underline{b}(\sigma_j(i)) \left[p^{(j)}(\sigma_j(i)) - p^{(j)}(\sigma_j(\underline{n}_j))\right] \\
&\qquad + \sum_{i=\underline{n}_j+1}^{N_x} \overline{b}(\sigma_j(i)) \left[p^{(j)}(\sigma_j(i)) - p^{(j)}(\sigma_j(\underline{n}_j))\right] \\
&= \sum_{i=1}^{\underline{n}_j-1} \underline{b}(\sigma_j(i))p^{(j)}(\sigma_j(i)) + \sum_{i=\underline{n}_j+1}^{N_x} \overline{b}(\sigma_j(i))p^{(j)}(\sigma_j(i)) \\
&\qquad + \left(1 - \sum_{i=1}^{\underline{n}_j-1} \underline{b}(\sigma_j(i)) - \sum_{i=\underline{n}_j+1}^{N_x} \overline{b}(\sigma_j(i))\right) p^{(j)}(\sigma_j(\underline{n}_j)) \\
&= \min_{\psi\in\boldsymbol{\Psi}_s} (\psi P_\pi)_j = \sum_{k=1}^{N_x} \underline{\psi}_k^{(j)} p^{(j)}(k)\,.
\end{aligned}
$$

So the proof is complete. □

**Example 5.3.2.** *(Special Case) When the size of the transition matrix is 2×2, express the transition matrix as*

$$
P_\pi = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \tag{5.10}
$$

*where $p \geq 1-q$ and $p, q \in (0,1)$. Since there is no need to consider both the lower and upper bound in this case, we write the safety specification*

$$\boldsymbol{\Psi}_s := \{\psi \in \boldsymbol{\Psi} | \psi \leq \boldsymbol{b} = [B_1 \; B_2]\}\,. \tag{5.11}$$

To prevent trivialities, we assume the unique existence of an invariant distribution $\psi^*$ for $P_\pi$. Also, assume $\boldsymbol{b} \geq \psi^*$. So we have

$$\psi^* P_\pi = \psi^* = [\frac{1-q}{2-p-q} \quad \frac{1-p}{2-p-q}] \leq \boldsymbol{b} = [B_1 \quad B_2]\,. \tag{5.12}$$

From Theorem 5.3.1 and condition (5.12) it is easy to conclude that (5.3) holds since

$$B_1 p + (1-B_1)(1-q) \leq B_1$$

and

$$(1-B_2)(1-p) + B_2 q \leq B_2\,.$$

## 5.4 Safe Policy

As defined in Section 5.2, if there exists an associated set $\boldsymbol{\Psi}_{in} \subset \boldsymbol{\Psi}_s$ for an admissible policy $\pi \in \Pi$ such that

$$\psi \in \boldsymbol{\Psi}_{in} \Rightarrow \psi^{(k)} = \psi P_\pi^k \in \boldsymbol{\Psi}_s \qquad \forall k \in \mathbb{N}\,, \tag{5.13}$$

then this policy $\pi$ is called a safe policy corresponding to the $\boldsymbol{\Psi}_s$. In the following we apply linear programming techniques to identify a safe policy. The feasibility of our linear programming formulation is also studied.

### 5.4.1 Linear Programming Formulation

Consider any policy $\pi$ that induces an irreducible and aperiodic $P_\pi$. If its unique invariant distribution $\psi^*$ satisfies $\psi^* \boldsymbol{A} \leq \boldsymbol{b}$, then $\pi$ is a safe policy with $\boldsymbol{\Psi}_{in}$ equal to, for example, the set containing the only element $\psi^*$. According to this observation, the safe policy for a given $\boldsymbol{\Psi}_s$ might not be unique, if exists. Thus we introduce a cost function $c : \boldsymbol{X} \times \boldsymbol{U} \to \mathbb{R}$ as defined in the MDP model and have the following hypothesis.

**Assumption 5.4.1.** Suppose for every stationary policy $\pi$ (defined in Section 1.3.3) the induced transition matrix $P_\pi$ is irreducible and aperiodic.

Our goal is to find the safe policy that incurs the minimal long-run average cost under Assumption 5.4.1 by applying linear programming techniques. To achieve this, denote

$$p_{ij}^u = \text{Prob}(X_{n+1} = j | X_n = i, u_n = u) \qquad \forall n \in \mathbb{N}\,.$$

Let $\mathcal{U}(i)$ be the set of available action(s) when the system is at state $i$. Suppose $\beta_{iu}$ defines a randomized policy which assigns the probability $\beta_{iu}$ to the action $u \in \mathcal{U}(i)$ when the system is at state $i \in \boldsymbol{X}$. Thus, $\beta_{iu} \geq 0$ for each $i \in \boldsymbol{X}$ and $u \in \mathcal{U}(i)$. Also,

$$\sum_{i \in \boldsymbol{X}} \sum_{u \in \mathcal{U}(i)} \beta_{iu} = 1\,. \tag{5.14}$$

The average probability that the system visits state $i$ is $\sum_{u \in \mathcal{U}(i)} \beta_{iu}$. So the identity for the invariant distribution $\psi^* = \psi^* P_\pi$ suggests

$$\sum_{u \in \mathcal{U}(i)} \beta_{iu} = \sum_{j \in \boldsymbol{X}} \sum_{u \in \mathcal{U}(j)} \beta_{ju} p_{ji}^u \qquad \forall i \in \boldsymbol{X}\,, \tag{5.15}$$

and the safety constraint $\psi^* \boldsymbol{A} \leq \boldsymbol{b}$ implies

$$\sum_j \sum_{u \in \mathcal{U}(j)} \beta_{ju} \boldsymbol{A}_{ji} \leq \boldsymbol{b}(i) \qquad \forall i \in \boldsymbol{X}. \tag{5.16}$$

Finally, the *optimal* safe policy is obtained by putting (5.14) $\sim$ (5.16) all together and solving the following linear programming:

$$\begin{aligned} \min_{\beta_{iu}} \quad & \sum_{i \in \boldsymbol{X}} \sum_{u \in \mathcal{U}} c(i,u) \beta_{iu} \\ \text{subject to} \quad & \sum_{i \in \boldsymbol{X}} \sum_{u \in \mathcal{U}(i)} \beta_{iu} = 1\,, \\ & \sum_{u \in \mathcal{U}(i)} \beta_{iu} = \sum_{j \in \boldsymbol{X}} \sum_{u \in \mathcal{U}(j)} \beta_{ju} p_{ji}^u \qquad \forall i \in \boldsymbol{X}, \\ & \sum_j \sum_{u \in \mathcal{U}(j)} \beta_{ju} \boldsymbol{A}_{ji} \leq \boldsymbol{b}(i) \qquad \forall i \in \boldsymbol{X}, \\ & \beta_{iu} \geq 0 \qquad \forall i \in \boldsymbol{X}, u \in \mathcal{U}(i)\,. \end{aligned} \tag{5.17}$$

### 5.4.2 Feasibility Analysis

In this section we discuss the feasibility of the linear programming formulation (5.17). Some general properties regarding this formulation are also presented. Denote $\mathcal{U}(i) = \{1, \ldots, n_i\}$ for each $i \in \boldsymbol{X}$, $\sum_{i=1}^{N_x} n_i = m$, $I_m$ is the identity matrix with size $m \times m$, and $e^i$ the $i^{th}$ column of the identity matrix with size $N_x \times N_x$. $\mathbf{0}_m = [0 \ \cdots \ 0]^T \in \mathbb{R}^m$. $\mathbf{1}_n := [1 \ \cdots \ 1]^T \in \mathbb{R}^n$. Moreover, let $A_{i\cdot}$ be the $i^{th}$ row of $\boldsymbol{A}$ and $A_{i\cdot}^T$ the transpose of $A_{i\cdot}$. We write

(5.17) in the following form:

$$\begin{aligned} \min_{\beta} \quad & c^T\beta \\ \text{subject to} \quad & R\beta = S\,, \\ & W\beta \le Z\,, \\ & \beta \ge \mathbf{0}_m\,, \end{aligned} \tag{5.18}$$

where

$$p^u_{i\cdot} = [p^u_{i1} \quad p^u_{i2} \quad \cdots \quad p^u_{iN_x}] \qquad \forall i \in \boldsymbol{X}, u \in \mathcal{U}(i)\,,$$

$$\begin{aligned} \beta &= [\beta_{11} \quad \beta_{12} \quad \cdots \quad \beta_{1n_1} \quad \beta_{21} \quad \cdots \quad \beta_{N_x n_{N_x}}]^T\,, \\ c &= [c(1,1) \quad c(1,2) \quad \cdots \quad c(1,n_1) \quad c(2,1) \quad \cdots \quad c(N_x, n_{N_x})]^T\,, \\ R &= \begin{bmatrix} p^{1T}_{1\cdot} - e^1 & p^{2T}_{1\cdot} - e^1 & \cdots & p^{n_1 T}_{1\cdot} - e^1 & p^{1T}_{2\cdot} - e^2 & \cdots & p^{{n_{N_x}}^T}_{N_x\cdot} - e^{N_x} \\ 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \end{bmatrix}. \end{aligned}$$

Note that $R \in \mathbb{R}^{(N_x+1)\times m}$. Furthermore,

$$\begin{aligned} S &= [0 \quad \cdots \quad 0 \quad 1]^T \in \mathbb{R}^{N_x+1}\,, \\ W &= [A^T_{1\cdot}\mathbf{1}^T_{n_1} \quad A^T_{2\cdot}\mathbf{1}^T_{n_2} \quad \cdots \quad A^T_{N_x\cdot}\mathbf{1}^T_{N_x}] \in \mathbb{R}^{N_b\times m}\,, \\ Z &= [\boldsymbol{b}(1) \quad \boldsymbol{b}(2) \quad \ldots \quad \boldsymbol{b}(N_b)]^T \in \mathbb{R}^{N_b}\,. \end{aligned}$$

The linear programming problem (5.17) is feasible if the set

$$\beta^{\le} := \{\beta | R\beta = S, W\beta \le Z, \beta \ge \mathbf{0}\}$$

is nonempty. Consider the set

$$\theta^{\ge} := \{\theta^T = [\theta_1^T \quad \theta_2^T] | \theta_1^T S + \theta_2^T Z < 0, \theta_1^T R + \theta_2^T W \ge \mathbf{0}^T, \theta_2 \ge \mathbf{0}\}\,.$$

If both $\beta^{\leq}$ and $\theta^{\geq}$ are nonempty, then there exist $\beta^*$ and $\theta^*$ such that

$$\theta_1^{*T} S = \theta_1^{*T} R\beta^* \,, \qquad \theta_2^{*T} Z \geq \theta_2^{*T} W\beta^* \,.$$

Hence

$$\theta_1^{*T} S + \theta_2^{*T} Z \geq \theta_1^{*T} R\beta^* + \theta_2^{*T} W\beta^* = (\theta_1^{*T} R + \theta_2^{*T} W)\beta^* \geq 0$$

and a contradiction results, which means that $\beta^{\leq}$ and $\theta^{\geq}$ can not be both nonempty.

The question of which set is nonempty is answered by considering the following quadratic programming:

$$\begin{aligned} \min_{\beta} \quad & \|R\beta - S\|^2 \\ \text{subject to} \quad & W\beta \leq Z \,, \\ & \beta \geq \mathbf{0} \,. \end{aligned} \tag{5.19}$$

where $\|\cdot\|$ denotes the Euclidean norm. Denote $r_\beta := R\beta - S$ and we have the following result:

**Lemma 5.4.1.** *In (5.19), for a feasible $\beta^*$, that is, $\beta^* \geq \mathbf{0}_{N_b}$ and $W\beta^* \leq Z$, if there exists $v \geq \mathbf{0}$ such that $r_{\beta^*}^T R + v^T W \geq \mathbf{0}^T$, $(r_{\beta^*}^T R + v^T W)\beta^* = 0$, and $v^T(W\beta^* - Z) = 0$, then $\beta^*$ solves (5.19).*

*Proof.* See Appendix C.1. □

**Remark 5.4.2.** Define $L_\beta$ the index set of *active constraints* at $\beta$. That is $L_\beta := \{r | W_{r\cdot}\beta = Z(r)\}$ where $W_{r\cdot}$ is the $r^{th}$ row of $W$ and $Z(r)$ is the $r^{th}$ element of $Z$. If $\beta^*$ is a *regular point*, that is, if the rows $W_{r\cdot}$'s are linearly

independent for $r \in L_{\beta^*}$, then the condition in Lemma 5.4.1 is the necessary and sufficient condition due to the convexity of the objective function. See Remark C.1.1 in Appendix C.1 also.

According to the definition of $r_\beta$,

$$r_{\beta^*}^T S + v^T Z = -\|r_{\beta^*}\|^2 + r_{\beta^*}^T R\beta^* + v^T Z\,.$$

By the conditions of Lemma 5.4.1, $v^T Z = v^T W\beta^*$, and $(r_{\beta^*}^T R + v^T W)\beta^*$=0. So the following result is ready.

**Theorem 5.4.3.** *Suppose that $\beta^*$ solves (5.19) and is a regular point. The residual vector $r_{\beta^*} = R\beta^* - S$. If $r_{\beta^*} = \mathbf{0}$ then $\beta^* \in \beta^{\leq}$. Otherwise there exists $v \geq \mathbf{0}_{N_b}$ such that if $r^{*T} := [r_{\beta^*}^T \; v^T]$ then $r^* \in \theta^{\geq}$. Furthermore, $r_{\beta^*}^T S + v^T Z = -\|r_{\beta^*}\|^2 < 0$.*

$r^*$ in Theorem 5.4.3 has the following property.

**Theorem 5.4.4.** *If in Theorem 5.4.3 $r_{\beta^*} \neq \mathbf{0}$ then $r^*/\|r_{\beta^*}\|$ solves the following problem*

$$\begin{aligned} min_{\theta_1,\theta_2} \quad & \theta_1^T S + \theta_2^T Z \\ subject\ to \quad & \theta_1^T R + \theta_2^T W \geq 0\,, \\ & \|\theta_1\| = 1\,, \\ & \theta_2 \geq \mathbf{0}\,. \end{aligned} \tag{5.20}$$

*Proof.*

$$\begin{aligned} & \theta_1^T S + \theta_2^T Z \\ = \; & \theta_1^T (R\beta^* - r_{\beta^*}) + \theta_2^T (Z - W\beta^*) + \theta_2^T W\beta^* \\ = \; & -\theta_1^T r_{\beta^*} + (\theta_1^T R + \theta_2^T W)\beta^* + \theta_2^T (Z - W\beta^*)\,. \end{aligned}$$

Since $W\beta^* \leq Z$, $\theta_2 \geq \mathbf{0}$, $\beta^* \geq \mathbf{0}$, and $\theta_1, \theta_2$ satisfy the constraints in (5.20), we have

$$\theta_1^T S + \theta_2^T Z \geq -\theta_1^T r_{\beta^*} \geq -\|r_{\beta^*}\|$$

where the second inequality follows from the Cauchy inequality. So by Theorem 5.4.3 the minimizer $[\theta_1^* \ \theta_2^*]$ solve (5.20) where $\theta_1^* = r_{\beta^*}/\|r_{\beta^*}\|$ and $\theta_2^* = v/\|r_{\beta^*}\|$ . □

## 5.5 Supremal Invariant Safe Set

For a safety specification $\boldsymbol{\Psi}_s = \{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A} \leq \boldsymbol{b}\}$ as defined in (5.1), we mentioned in Section 5.4 that an admissible policy $\pi \in \Pi$ is safe if there exists an associated set $\boldsymbol{\Psi}_{in} \subset \boldsymbol{\Psi}_s$ such that

$$\psi \in \boldsymbol{\Psi}_{in} \Rightarrow \psi^{(k)} = \psi P_\pi^k \in \boldsymbol{\Psi}_s \qquad \forall k \in \mathbb{N}\,. \tag{5.21}$$

Here $\boldsymbol{\Psi}_{in}$ is called an safe set under $\pi$. In this section we study the characterization of a safe set for a given safe policy, in particular, the largest safe set when it is not unique. To achieve this, we first introduce the concept called *invariant safe set*. A set $\hat{\boldsymbol{\Psi}} \subset \boldsymbol{\Psi}_s$ is an *invariant safe set* if $\psi \in \hat{\boldsymbol{\Psi}}$ implies $\psi P_\pi \in \hat{\boldsymbol{\Psi}}$. Thus, for a policy $\pi$, if it induces a safety enforcing controller mentioned in Section 5.3, then its invariant safe set can be as large as the entire safety specification $\boldsymbol{\Psi}_s$, which is the possibly largest. If $\pi$ induces an irreducible and aperiodic $P_\pi$ with the unique invariant distribution $\psi^*$ being safe, then its invariant safe set can be as small as the set containing the only element $\psi^*$. Other nontrivial examples in an easy form for invariant safe sets can be written as

$$\{\psi \in \boldsymbol{\Psi} | \underline{\boldsymbol{d}} \leq \psi \leq \overline{\boldsymbol{d}}\}$$

where

1. $\underline{\boldsymbol{d}} = (1-\varepsilon_1)\psi^*$, $\overline{\boldsymbol{d}} = \underline{\boldsymbol{d}} + \varepsilon_1 \mathbf{1}^T$, and

$$0 \leq \varepsilon_1 \leq \min_{j \in \boldsymbol{X}} \frac{\boldsymbol{b}(j) - \psi^* A_{\cdot j}}{\overline{A}_{\cdot j} - \psi^* A_{\cdot j}} \,. \tag{5.22}$$

2. $\overline{\boldsymbol{d}} = (1+\varepsilon_2)\psi^*$, $\underline{\boldsymbol{d}} = \overline{\boldsymbol{d}} - \varepsilon_2 \mathbf{1}^T$ and

$$0 \leq \varepsilon_2 \leq \min\{\min_{j \in \boldsymbol{X}} \frac{\boldsymbol{b}(j) - \psi^* A_{\cdot j}}{\psi^* A_{\cdot j} - \underline{A}_{\cdot j}} \,, \min_{j \in \boldsymbol{X}} \frac{\psi^*(j)}{1 - \psi^*(j)} \,, \min_{j \in \boldsymbol{X}} \frac{1 - \psi^*(j)}{\psi^*(j)}\} \,. \tag{5.23}$$

Here $\overline{A}_{\cdot j}$, $\underline{A}_{\cdot j}$ is the maximal, minimal value in the $j^{th}$ column of $A$, respectively, and the upper bounds for $\varepsilon_1$ and $\varepsilon_2$ in (5.22) and (5.23) are assumed positive. These examples are motivated by the fact that if

$$\Delta := (1-\varepsilon_1)\psi^* + \varepsilon_1 \cdot \boldsymbol{\Psi}$$

or

$$\Delta := (1+\varepsilon_2)\psi^* - \varepsilon_2 \cdot \boldsymbol{\Psi}$$

then $\Delta P_\pi \subset \Delta$.

The above observation shows the fact that given a policy and a safety specification, the associated invariant safe set, if exists, might not be unique. We call the largest one *supremal safe set* and write it as $\boldsymbol{\Psi}_{ss}$. As we will see later, to search the largest safe set is equal to search the largest invariant safe set. Thus, in the following section we introduce algorithms to characterize $\boldsymbol{\Psi}_{ss}$ for a given safety specification and a policy. To make our problem more tractable, we make the following *interior assumption* on the given policy.

**Assumption 5.5.1.** Given a safety specification $\boldsymbol{\Psi}_s$ as defined in (5.1), an admissible policy $\pi \in \Pi$ is called to satisfy the *interior assumption* if (a) the induced transition matrix $P_\pi$ is irreducible and aperiodic and (b) there exists an $\varepsilon > 0$ such that $\psi^* \boldsymbol{A} \leq \boldsymbol{b} - \varepsilon \mathbf{1}_{N_b}^T$ where $\mathbf{1}_{N_b}^T = [1 \cdots 1]$ is a row vector of 1's with size $N_b$, and $\psi^*$ is the unique invariant distribution for $P_\pi$.

### 5.5.1 Searching Algorithm

Consider the following algorithm to compute $\boldsymbol{\Psi}_{ss}$.

**Algorithm 5.5.1.** (Forward Algorithm) Let $\boldsymbol{\Psi}_s = \{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A} \leq \boldsymbol{b}\}$ as defined in (5.1) be a safety specification. $P_\pi$ is a transition probability matrix with an invariant distribution $\psi^*$. Define

$$\Delta_{\varepsilon_1} := (1 - \varepsilon_1)\psi^* + \varepsilon_1 \boldsymbol{\Psi}$$

where

$$\varepsilon_1 = \min_j \frac{\boldsymbol{b}(j) - \psi^* A_{\cdot j}}{\overline{A}_{\cdot j} - \psi^* A_{\cdot j}} \,.$$

The Forward Algorithm is to calculate the following recursively.

$$\begin{aligned}
\boldsymbol{\Psi}^{(0)} :=& \Delta_{\varepsilon_1} \,, \\
\boldsymbol{\Psi}^{(k)} :=& \{\psi \in \boldsymbol{\Psi}_s | \psi P_\pi \in \boldsymbol{\Psi}^{(k-1)}\} \\
=& \{\psi \in \boldsymbol{\Psi}_s | \psi P_\pi^k \in \Delta_{\varepsilon_1},\ \psi P_\pi^j \in \boldsymbol{\Psi}_s,\ 1 \leq j \leq k-1\} \qquad \forall k \in \mathbb{N} \,.
\end{aligned}$$

Some properties regarding the algorithm are explored in the following. Suppose $P_\pi$ is irreducible and aperiodic. It is shown in [70, Theorem 2.9] that

$$\rho := \min_{i,j} \left\{ [P_\pi^r]_{ij} \right\} > 0 \qquad \text{as} \quad r = N_x^2 - 2N_x + 2 \,. \tag{5.24}$$

We use this property to prove the following lemma.

**Lemma 5.5.2.** *Let $\psi^*$ be the invariant distribution of an irreducible and aperiodic $P_\pi$. $\rho$ and $r$ are defined as in (5.24). If $\lambda_0 \geq 0$ is the smallest non-negative real number such that*

$$(1-\lambda_0)\psi^*(i) \leq \rho \qquad \forall i = 1, 2, \ldots, n$$

*then*

$$(\boldsymbol{\Psi} - \psi^*)P_\pi^r \subset \lambda_0(\boldsymbol{\Psi} - \psi^*)\,.$$

*Proof.* It is easy to see that for all $\psi \in \boldsymbol{\Psi}$

$$\sum_{i=1}^{n}((\psi P_\pi^r)(i) - (1-\lambda_0)\psi^*(i)) = \lambda_0\,.$$

By the definition of $\rho$ we have

$$\psi P_\pi^r - (1-\lambda_0)\psi^* \geq \rho - (1-\lambda_0)\psi^* \geq 0 \qquad \forall \psi \in \boldsymbol{\Psi}\,,$$

so

$$\psi P_\pi^r - (1-\lambda_0)\psi^* \in \lambda_0\boldsymbol{\Psi} \qquad \forall \psi \in \boldsymbol{\Psi}\,.$$

Thus

$$\psi P_\pi^r \in \psi^* + \lambda_0(\boldsymbol{\Psi} - \psi^*) \qquad \forall \psi \in \boldsymbol{\Psi}\,,$$

and the proof is completed. □

**Theorem 5.5.3.** *Suppose a policy $\pi$ satisfies the interior assumption as defined in Assumption 5.5.1 with its induced transition matrix $P_\pi$. Let $\varepsilon_1$, $\Delta_{\varepsilon_1}$, $r$, and $\lambda_0$ be defined as in Algorithm 5.5.1 and Lemma 5.5.2. Let $k_0$ be the smallest positive integer such that $\lambda_0^{k_0} \leq \varepsilon_1$. Then the Forward Algorithm terminates in finite steps with an upper bound $rk_0$. That is, $\boldsymbol{\Psi}^{(rk_0)} = \boldsymbol{\Psi}_{ss}$.*

*Proof.* It is easy to see that

$$\boldsymbol{\Psi}^{(0)} \subset \boldsymbol{\Psi}^{(1)} \subset \cdots \subset \boldsymbol{\Psi}_{ss}.$$

Hence, to show $\boldsymbol{\Psi}^{(rk_0)} = \boldsymbol{\Psi}_{ss}$, it is enough to show $\boldsymbol{\Psi}^{(rk_0+n)} = \boldsymbol{\Psi}^{(rk_0)}$ for $n \in \mathbb{N}_0$. By Lemma 5.5.2, for each initial distribution $\psi \in \boldsymbol{\Psi}$ and $n \in \mathbb{N}_0$

$$\begin{aligned}
\psi^{(rk_0+n)} &= \psi P_\pi^{rk_0+n} \\
&= \psi P_\pi^n P_\pi^{rk_0} \\
&\in \psi^* + \lambda_0^{k_0}(\boldsymbol{\Psi} - \psi^*) \\
&= \psi^* + \varepsilon_1(\frac{\lambda_0^{k_0}}{\varepsilon_1}\boldsymbol{\Psi} - \frac{\lambda_0^{k_0}}{\varepsilon_1}\psi^*) \\
&= \psi^* + \varepsilon_1(\frac{\lambda_0^{k_0}}{\varepsilon_1}\boldsymbol{\Psi} + (1 - \frac{\lambda_0^{k_0}}{\varepsilon_1})\psi^* - \psi^*) \\
&\subset \psi^* + \varepsilon_1(\boldsymbol{\Psi} - \psi^*) = \Delta_{\varepsilon_1}\,.
\end{aligned}$$

That means, for each $\psi \in \boldsymbol{\Psi}$ $\psi^{(k)} \in \Delta_{\varepsilon_1} \subset \boldsymbol{\Psi}_s$ for $k \geq rk_0$. Therefore,

$$\begin{aligned}
\boldsymbol{\Psi}^{(rk_0+n)} &= \{\psi \in \boldsymbol{\Psi}_s | \psi^{(k)} \in \boldsymbol{\Psi}_s, \forall k = 1, 2, \ldots, rk_0 + n - 1, \psi^{(rk_0+n)} \in \Delta_{\varepsilon_1}\} \\
&= \{\psi \in \boldsymbol{\Psi}_s | \psi^{(k)} \in \boldsymbol{\Psi}_s, \forall k = 1, 2, \ldots, rk_0 - 1\} \\
&= \{\psi \in \boldsymbol{\Psi}_s | \psi^{(k)} \in \boldsymbol{\Psi}_s, \forall k = 1, 2, \ldots, rk_0 - 1, \psi^{(rk_0)} \in \Delta_{\varepsilon_1}\} \\
&= \boldsymbol{\Psi}^{(rk_0)}\,.
\end{aligned}$$

We thus finish the proof. □

Another algorithm to computer $\boldsymbol{\Psi}_{ss}$ is in the following.

**Algorithm 5.5.4.** (Backward Algorithm)

$$\begin{aligned}
\boldsymbol{\Psi}^{(0)} &= \{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A} \leq \boldsymbol{b}\} = \boldsymbol{\Psi}_s\,, \\
\boldsymbol{\Psi}^{(k)} &= \{\psi \in \boldsymbol{\Psi}^{(k-1)} | \psi P_\pi \in \boldsymbol{\Psi}^{(k-1)}\} \\
&= \{\psi \in \boldsymbol{\Psi}_s | \psi P_\pi^j \in \boldsymbol{\Psi}_s,\ 1 \leq j \leq k\} \qquad \forall k \in \mathbb{N}.
\end{aligned}$$

**Remark 5.5.5.** The difference between the Forward and Backward Algorithm is that the Forward Algorithm starts the candidate of $\boldsymbol{\Psi}_{ss}$ from an invariant safe set that is a subset of $\boldsymbol{\Psi}_{ss}$, and then expands the candidate at each iteration till it attains $\boldsymbol{\Psi}_{ss}$. The Backward Algorithm starts the candidate from $\boldsymbol{\Psi}_{s}$, which is the possibly largest for $\boldsymbol{\Psi}_{ss}$, and then contracts the candidate at each iteration till $\boldsymbol{\Psi}_{ss}$ is reached. An upper bound for the number of iterations to calculate $\boldsymbol{\Psi}_{ss}$ in the Backward Algorithm is given in Section 5.6.3.

### 5.5.2 Special Case as the size of $P_\pi$ is 2×2

In this section we revisit Example 5.3.2. Write

$$P_\pi = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix}, \quad P_\pi^n = \overbrace{P_\pi \cdot P_\pi \cdots P_\pi}^{n} = \begin{bmatrix} p_{11}^n & p_{12}^n \\ p_{21}^n & p_{22}^n \end{bmatrix} \tag{5.25}$$

where $0 \le p, q \le 1$. It is easy to justify the following special cases.

$$p = q = 1 \quad \Rightarrow \quad \boldsymbol{\Psi}_{ss} = \begin{cases} \boldsymbol{\Psi}_s & \text{if } B_1 + B_2 \ge 1 \\ \emptyset & \text{otherwise} \end{cases}, \tag{5.26}$$

$$p + q = 1 \quad \Rightarrow \quad \boldsymbol{\Psi}_{ss} = \begin{cases} \boldsymbol{\Psi} & \text{if } p \le B_1, q \le B_2 \\ \emptyset & \text{otherwise} \end{cases}, \tag{5.27}$$

$$\begin{aligned} & p = q = 0 \\ \Rightarrow \quad & \boldsymbol{\Psi}_{ss} = \begin{cases} \{(x_1, x_2) \ge 0 | x_1 + x_2 = 1, 1 - \underline{B} \le x_1 \le \underline{B}\} & \text{if } \underline{B} \ge \frac{1}{2} \\ \emptyset & \text{otherwise} \end{cases} \end{aligned} \tag{5.28}$$

where $\underline{B} = \min\{B_1, B_2\}$.

From now on, we assume that

$$(p, q) \ne (1, 1), \quad (p, q) \ne (0, 0), \quad \text{and} \quad p + q \ne 1\,. \tag{5.29}$$

Also, (5.12) is assumed to hold so that the invariant safe set is nonempty. The following lemma shows the *monotone* property when the size of $P_\pi$ is 2×2.

**Lemma 5.5.6.** *With the definition in (5.25), and assumption in (5.29) we have for $n \in \mathbb{N}$:*

$$p_{11}^n - p_{21}^n \begin{cases} < 0 & \text{if } p < 1-q \text{ and } n = 1,3,5\cdots \\ > 0 & \text{otherwise} \end{cases},$$

$$p_{11}^{n+1} - p_{11}^n \begin{cases} > 0 & \text{if } p < 1-q \text{ and } n = 1,3,5\cdots \\ < 0 & \text{otherwise} \end{cases},$$

$$p_{21}^{n+1} - p_{21}^n \begin{cases} < 0 & \text{if } p < 1-q \text{ and } n = 1,3,5\cdots \\ > 0 & \text{otherwise} \end{cases}.$$

*Proof.* By applying the mathematical induction it is easy to see that for $n \in \mathbb{N}$

$$p_{11}^n - p_{21}^n = (p+q-1)^n \,. \tag{5.30}$$

Moreover,

$$p_{11}^{n+1} - p_{11}^n = (1-p)(p_{21}^n - p_{11}^n)\,,$$

$$p_{21}^{n+1} - p_{21}^n = (1-q)(p_{11}^n - p_{21}^n)\,,$$

so the results follow. □

Define

$$A_{=B_1}^{(n)} := \{(x_1, x_2) \geq (0,0) \mid x_1 + x_2 = 1,\ x_1 p_{11}^n + x_2 p_{21}^n = B_1\}\,.$$

Also, define $M_{B_1} := \{n \in \mathbb{N} | A_{=B_1}^{(n)} \neq \emptyset\}$. It is apparent that with the assumption in (5.29) we have for $n \in M_{B_1}$

$$\begin{cases} p_{11}^n \leq B_1 \leq p_{21}^n & \text{if } p < 1-q \text{ and } n\text{: odd number} \\ p_{21}^n \leq B_1 \leq p_{11}^n & \text{otherwise} \end{cases}.$$

**Lemma 5.5.7.** *Suppose $(x_1^{(n)}, x_2^{(n)}) \in A_{=B_1}^{(n)}$ for $n \in M_{B_1}$. With the assumption in (5.29) and (5.12) we have*

$$x_2^{(n+1)} - x_2^{(n)} \begin{cases} > 0 & \text{if } p < 1-q \text{ and } n\text{: even number} \\ < 0 & \text{otherwise} \end{cases},$$

$$x_2^{(n+2)} - x_2^{(n)} \begin{cases} > 0 & \textit{if } p < 1-q \textit{ and } n : \textit{odd number} \\ < 0 & \textit{otherwise} \end{cases} .$$

*Proof.* Through some calculation, we obtain

$$\begin{aligned} x_2^{(n+1)} - x_2^{(n)} &= \frac{B_1 - p_{11}^{n+1}}{p_{21}^{n+1} - p_{11}^{n+1}} - \frac{B_1 - p_{11}^{n}}{p_{21}^{n} - p_{11}^{n}} \\ &= \frac{B_1(2-p-q) - (1-q)}{p_{21}^{n+1} - p_{11}^{n+1}} \end{aligned}$$

and

$$\begin{aligned} x_2^{(n+2)} - x_2^{(n)} &= (x_2^{(n+2)} - x_2^{(n+1)}) + (x_2^{(n+1)} - x_2^{(n)}) \\ &= \frac{[B_1(2-p-q) - (1-q)](p+q)}{p_{21}^{n+2} - p_{11}^{n+2}} . \end{aligned}$$

By the assumption in (5.12) and Lemma 5.5.6, we finish the proof. □

**Corollary 5.5.8.** *For $n \in M_{B_1}$ define*

$$A_{\leq B_1}^{(n)} := \{(x_1, x_2) \geq (0,0) \mid x_1 + x_2 = 1,\ x_1 p_{11}^n + x_2 p_{21}^n \leq B_1\},$$

*then with the assumption in (5.29) and (5.12) we have*

$$\begin{cases} A_{\leq B_1}^{(n)} \subset A_{\leq B_1}^{(n+2)} & if \ \ p < 1-q \\ A_{\leq B_1}^{(n)} \subset A_{\leq B_1}^{(n+1)} & otherwise \end{cases} .$$

Similar arguments can be applied to the other upper bound $B_2$. The following theorem is thus ready.

**Theorem 5.5.9.** *Suppose the stochastic matrix $P_\pi$ is defined as in (5.25) and the safety set $\boldsymbol{\Psi}_s$ is defined as in (5.11). Consider the Backward Algorithm to identify the supremal invariant safe set $\boldsymbol{\Psi}_{ss}$:*

$$\begin{aligned} S^{(0)} &= \boldsymbol{\Psi}_s := \{\psi \in \boldsymbol{\Psi} | \psi \leq \boldsymbol{b} = [B_1 \ B_2]\} , \\ S^{(k)} &= \{\psi \in S^{(k-1)} | \psi P_\pi \in S^{(k-1)}\} \\ &= \{\psi \in \boldsymbol{\Psi}_s | \psi P_\pi^j \in \boldsymbol{\Psi}_s,\ 1 \leq j \leq k\} \quad \forall k \in \mathbb{N} . \end{aligned}$$

*If assumptions in (5.29) and (5.12) hold then* $\boldsymbol{\Psi}_{ss} = S^{(1)}$.

*Proof.* If $p \geq 1 - q$ then $\boldsymbol{\Psi}_s$ itself is the supremal invariant safe set by (5.12) as seen in Example 5.3.2. That is: $\boldsymbol{\Psi}_s = \boldsymbol{\Psi}_{ss} = S^{(0)} = S^{(1)}$. If $p < 1 - q$, by Corollary 5.5.8 we just need to consider two iterations of the algorithm. The supremal invariant safe set $\boldsymbol{\Psi}_{ss}$ can thus be written as

$$\boldsymbol{\Psi}_{ss} = \{(x_1, x_2) \geq (0, 0) \mid x_1 + x_2 = 1,\ \underline{x}_2 \leq x_2 \leq \overline{x}_2\}$$

where

$$\underline{x}_2 = \max\left\{\frac{1-p-B_2}{1-q-p}, \frac{p^2+(1-p)(1-q)-B_1}{(1-q-p)^2}, 1-B_1\right\},$$
$$\overline{x}_2 = \min\left\{\frac{B_1-p}{1-q-p}, \frac{B_2-(1-p)(p+q)}{(1-q-p)^2}, B_2\right\}.$$

Since by (5.12)

$$\frac{p^2+(1-p)(1-q)-B_1}{(1-q-p)^2} \leq 1 - B_1$$

and

$$\frac{B_2-(1-p)(p+q)}{(1-q-p)^2} \geq B_2\,,$$

which means the inequalities caused by the second iteration (i.e. $k = 2$) are redundant, so the supremal invariant safe set

$$\begin{aligned}\boldsymbol{\Psi}_{ss} = \{(x_1, x_2)| \quad & x_1 + x_2 = 1, \\ & \max\{\frac{1-p-B_2}{1-q-p}, 1-B_1\} \leq x_2 \leq \min\{\frac{B_1-p}{1-q-p}, B_2\}\}\end{aligned} \tag{5.31}$$

and the proof is complete. □

**Remark 5.5.10.** The conclusion in Theorem 5.5.9 can also be checked with the following argument. Consider a system of $m$ linear inequalities with a

unknown vector $\boldsymbol{x} \in \mathbb{R}^N$ :

$$\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\,, \quad \boldsymbol{x} \geq \boldsymbol{0}\,, \tag{5.32}$$

and an additional inequality $\boldsymbol{d}\boldsymbol{x} \leq d_0$. It is easy to see that if there exists an $\boldsymbol{u} \in \mathbb{R}^m$ satisfying

$$\boldsymbol{u} \geq \boldsymbol{0}\,, \quad \boldsymbol{d} \leq \boldsymbol{u}\boldsymbol{A}\,, \quad \boldsymbol{u}\boldsymbol{b} \leq d_0\,, \tag{5.33}$$

then $\boldsymbol{d}\boldsymbol{x} \leq d_0$ is redundant relative to (5.32).

We apply the above observation to test the redundant inequalities caused by iterations of the algorithm. Consider the case for $1 - q > p$ only. After one iteration we have

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} x_1 \leq \begin{bmatrix} \overline{x}_1 \\ -\underline{x}_1 \end{bmatrix}\,, \quad x_1 \geq 0 \tag{5.34}$$

where

$$\overline{x}_1 = \min\left\{B_1, \frac{B_2 - q}{1 - q - p}\right\}\,,$$
$$\underline{x}_1 = \max\left\{1 - B_2, \frac{1 - q - B_1}{1 - q - p}\right\}\,.$$

To test if $x_1 p_{11}^m + (1 - x_1) p_{21}^m \leq B_1$ is redundant relative to (5.34), we can check if there exist $u_1 \geq 0$ and $u_2 \geq 0$ such that

$$p_{11}^m - p_{21}^m \leq [u_1 \quad u_2] \begin{bmatrix} 1 \\ -1 \end{bmatrix}\,, \quad [u_1 \quad u_2] \begin{bmatrix} \overline{x}_1 \\ -\underline{x}_1 \end{bmatrix} \leq B_1 - p_{21}^m\,. \tag{5.35}$$

Note that the assumption in (5.12) is the same as

$$\frac{p_{21}^m}{1 + p_{21}^m - p_{11}^m} \leq B_1\,, \qquad \frac{1 - p_{11}^m}{1 + p_{21}^m - p_{11}^m} \leq B_2\,. \tag{5.36}$$

So by (5.36) and Lemma 5.5.6, if $m = 2n$ then we can take

$$\begin{bmatrix} \dfrac{B_1 - p_{21}^{2n}}{B_1} & 0 \end{bmatrix}$$

as a solution of $[u_1\ u_2]$ for (5.35). If $m = 2n+1$, by using (5.30) we obtain a solution

$$\begin{bmatrix} 0 & \dfrac{(p_{21}^{2n+1} - B_1)(1-q-p)}{1-q-B_1} \end{bmatrix}.$$

Similarly argument can be applied to show for $m \geq 2$ the redundancy of $x_1 p_{12}^m + (1-x_1)p_{22}^m \leq B_2$. So we conclude that the supremal invariant safety set is

$$\begin{aligned} \boldsymbol{\Psi}_{ss} = \{(x_1, x_2) | \quad & x_1 + x_2 = 1, \\ & \max\{\frac{1-q-B_1}{1-q-p}, 1-B_2\} \leq x_1 \leq \min\{\frac{B_2-q}{1-q-p}, B_1\}\}, \end{aligned}$$

which is equivalent to (5.31).

## 5.6 Remark on the Assumption

The purpose of the condition *irreducible and aperiodic* in Assumption 5.5.1 is to make sure the *unique* existence of $P_\pi$'s invariant distribution. If this condition fails, we are still able to study $\psi P_\pi^k$ for $k \in \mathbb{N}$. In particular, $\lim_{k\to\infty} \psi P_\pi^k$, if exists, might depend on $\psi$. In the following we first study the spectral representation of a matrix product to relate $\lim_{k\to\infty} P_\pi^k$ to $P_\pi$'s eigenvalues. Then we analyze $\lim_{k\to\infty} P_\pi^k$ when $P_\pi$ is not both irreducible and aperiodic.

### 5.6.1 Spectral Representation of A Matrix Function

In this section we develop the spectral representation of a matrix function based on the following lemma from linear algebra.

**Lemma 5.6.1.** *Let $\mathbb{C}$ be the set of complex numbers and $A, B \in \mathbb{C}^{n\times n}$. Suppose*

$\{v_1, \ldots, v_n\}$ *is any basis of* $\mathbb{C}^n$. *Then* $Av_i = Bv_i$ *for* $i = 1, 2, \cdots, n$ *implies* $A = B$.

The feature of our presentation is that no determination of eigenvectors is necessary.

**Theorem 5.6.2.** *Suppose the minimal polynomial of a matrix* $A \in \mathbb{R}^{N\times N}$ *is*

$$f(x) = \prod_{i=0}^{M} (x - \lambda_i)^{d_i} .$$

*Define*

$$f_j(x) = \prod_{i=0, i\neq j}^{M} \frac{(x-\lambda_i)^{d_i}}{(\lambda_j - \lambda_i)^{d_i}} , \quad f_j(A) = \prod_{i=0, i\neq j}^{M} \frac{(A-\lambda_i I)^{d_i}}{(\lambda_j - \lambda_i)^{d_i}} ,$$

*and the identity matrix* $I \in \mathbb{R}^{N\times N}$. *Denote* $q^{(i)}(x)$ *the* $i^{th}$ *derivative of* $q(x)$ *and suppose we can write*

$$q(x) = \sum_{n=0}^{\infty} \frac{q^{(n)}(0)}{n!} x^n .$$

*Then we have*

$$q(A) = \sum_{i=0}^{M} \sum_{j=0}^{d_i - 1} R_{ij} (A - \lambda_i I)^j \frac{q^{(j)}(\lambda_i)}{j!} \tag{5.37}$$

*where*

$$R_{ij} = f_i(A) \sum_{k=0}^{d_i - 1 - j} c_{i,k} (A - \lambda_i I)^k$$

*and for* $k = 0, 1, \ldots, d_i - 1$

$$c_{i,0} = 1 , \quad c_{i,k} = -\sum_{l=1}^{k} \sum_{0\leq k_s \leq d_s, \sum k_s = l} \prod_{s=0, s\neq i}^{M} \frac{\begin{pmatrix} d_s \\ k_s \end{pmatrix}}{(\lambda_i - \lambda_s)^{k_s}} c_{i,k-l} .$$

*Proof.* This is a slightly generalized version of the result in [60]. The proof is recapitulated in Appendix C.2. □

By the *Gerschgorin Disk Theorem* the absolute value of each eigenvalue in a stochastic matrix is at most 1. In (5.37), it is clear that only those $R_{ij}$ with $|\lambda_i| = 1$ need to be considered when we study $\lim_{n\to\infty} q(A)$ where $q(A) = A^n$. If $P_\pi$ is aperiodic, then $\lambda_0 = 1$ is the only eigenvalue with its absolute value equal to 1. Applying the idea similar to that in equality (C.1) in Appendix C.2 we may relate the invariant distribution $\psi^*$ with $P_\pi$'s eigenvalues. That is,

$$\lim_{k\to\infty} P_\pi^k \quad \longrightarrow \quad \prod_{i=1}^{M} \frac{(P_\pi - \lambda_i I)^{d_i}}{(1-\lambda_i)^{d_i}} \,.$$

This convergent property is obtained by only assuming that $P_\pi$ is aperiodic, so there may not exist a $\psi_s^* \in \boldsymbol{\Psi}$ such that $\lim_{k\to\infty} P_\pi^k = \mathbf{1}\psi_s^*$. That means $\lim_{k\to\infty} \psi P_\pi^k$ might depends on $\psi$. Hence the condition $\psi^* \boldsymbol{A} \leq \boldsymbol{b} - \varepsilon \mathbf{1}_{N_b}^T$ in Assumption 5.5.1 needs to be replaced with $\lim_{k\to\infty} \psi P_\pi^k \boldsymbol{A} \leq \boldsymbol{b} - \varepsilon \mathbf{1}_{N_b}^T$ where $\psi$ is the initial state probability distribution of the system.

### 5.6.2 Periodicity

In this section we consider a simple example that the transition matrix $P_\pi$ is periodic. Suppose $P_\pi$ is expressed as

$$P_\pi := \begin{bmatrix} \mathbf{0}_{2\times 2} & A_{2\times 3} \\ B_{3\times 2} & \mathbf{0}_{3\times 3} \end{bmatrix}$$

where each row of $A$ and $B$ is a positive probability vector. So this is a irreducible Markov chain with period 2, and

$$P^{2k} = \begin{bmatrix} (AB)^k_{2\times 2} & \mathbf{0}_{2\times 3} \\ \mathbf{0}_{3\times 2} & (BA)^k_{3\times 3} \end{bmatrix} \quad \longrightarrow \quad \begin{bmatrix} \mathbf{1}_{2\times 1} u^T & \mathbf{0}_{2\times 3} \\ \mathbf{0}_{3\times 2} & \mathbf{1}_{3\times 1} v^T \end{bmatrix} \qquad \text{as} \quad k \to \infty \,,$$

$$P^{2k+1} = \begin{bmatrix} \mathbf{0}_{2\times 2} & A(BA)^k \\ B(AB)^k & \mathbf{0}_{3\times 3} \end{bmatrix} \quad \longrightarrow \quad \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{1}_{2\times 1} v^T \\ \mathbf{1}_{3\times 1} u^T & \mathbf{0}_{3\times 3} \end{bmatrix} \qquad \text{as} \quad k \to \infty \,,$$

where $u = [\, u(1) \;\; u(2) \,]^T$, $v = [\, v(1) \;\; v(2) \;\; v(3) \,]^T$ are both positive probability vectors. Suppose $\psi = [\, \psi(1) \;\; \psi(2) \;\; \psi(3) \;\; \psi(4) \;\; \psi(5) \,]$, then

$$\psi P^{2k} \to [\, (\psi(1) + \psi(2))u^T \;\; (\psi(3) + \psi(4) + \psi(5))v^T \,] := \psi^*_{ev}$$

and

$$\psi P^{2k+1} \to [\, (\psi(3) + \psi(4) + \psi(3))u^T \;\; (\psi(1) + \psi(2))v^T \,] := \psi^*_{od}\,.$$

Again the limiting state probability distribution depends on the initial state probability distribution $\psi$. So we need to add in Assumption 5.5.1 that the initial distribution $\psi$ should satisfy

$$\psi^*_{od}\boldsymbol{A} \le \boldsymbol{b} - \varepsilon \mathbf{1}^T \quad \text{and} \quad \psi^*_{ev}\boldsymbol{A} \le \boldsymbol{b} - \varepsilon \mathbf{1}^T\,. \tag{5.38}$$

### 5.6.3 Scrambling Condition

We study in this section the weaker hypothesis which implies that $\lim_{k\to\infty} \psi P^k_\pi$ exists and does not depend on $\psi$. Specifically, consider the condition that there exists an integer $n \ge 1$ and a scalar $\alpha > 0$ such that

$$\sum_{j \in \boldsymbol{X}} \min\{(P^n_\pi)_{i_1 j}, (P^n_\pi)_{i_2 j}\} \ge \alpha \qquad \forall i_1, i_2 \in \boldsymbol{X}\,.$$

This condition is called *scrambling condition*. It is well known (see [25]) that for an aperiodic transition matrix, if the scrambling condition is satisfied, then there exists an integer $n \ge 1$ and a scalar $\rho > 0$ such that there exists an invariant distribution $\psi^* \in \boldsymbol{\Psi}$ for which

$$\sum_{j \in \boldsymbol{X}} |(P^m_\pi)_{ij} - \psi^*(j)| \le 2(1-\rho)^{\lfloor m/n \rfloor} \tag{5.39}$$

for all $i \in \boldsymbol{X}$ and $m \geq 1$, where $\lfloor x \rfloor$ denotes the largest integer not exceeding $x$. Under the assumption it is expected that the Backward Algorithm to calculate $\boldsymbol{\Psi}_{ss}$:

$$\begin{aligned}
\boldsymbol{\Psi}^{(0)} &= \boldsymbol{\Psi}_s = \{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A} \leq \boldsymbol{b}\}, \\
\boldsymbol{\Psi}^{(k)} &= \{\psi \in \boldsymbol{\Psi}^{(k-1)} | \psi P_\pi \in \boldsymbol{\Psi}^{(k-1)}\} \qquad \forall k = 1, 2, \dots,
\end{aligned}$$

will again terminate in finite steps. To seek an upper bound $k^*$ for the number of steps, we obtain from (5.39) that for each $i, j \in \boldsymbol{X}$,

$$(P_\pi^k \boldsymbol{A})_{ij} \leq \psi^* A_{\cdot j} + (1-\rho)^{\lfloor \frac{k}{n} \rfloor} (\overline{A}_{\cdot j} - \underline{A}_{\cdot j})$$

where $\overline{A}_{\cdot j}, \underline{A}_{\cdot j}$ is the maximal, minimal value in the $j$th column of $\boldsymbol{A}$, respectively. If Assumption 5.5.1-(b) is satisfied, then there exists an $\varepsilon > 0$ such that

$$\psi^* A_{\cdot j} + \varepsilon \leq \boldsymbol{b}(j) \qquad \forall j \in \boldsymbol{X}.$$

So we can identify a upper bound

$$k^* = n \times \left\lceil \max_{j \in \boldsymbol{X}} \frac{\log \frac{\varepsilon}{\overline{A}_{\cdot j} - \underline{A}_{\cdot j}}}{\log(1-\rho)} \right\rceil$$

where $\lceil x \rceil$ is the smallest integer exceeding $x$.

# Chapter 6

# Safety Control with Partial Observations

## 6.1 Introduction and Notation

This chapter is a continuation of our study on the safety control of stochastic discrete-time event systems (DESs). In Chapter 5 we studied the case when the stochastic DESs are modelled by complete observed Markov decision processes (COMDPs). We explore in this chapter the case when the stochastic DESs are modelled by partially observed Markov decision processes (POMDPs). It is mentioned in Section 1.4.4 that the conventional approach to deal with a POMDP model is to transform it into an equivalent COMDP model by considering the system's information states. Therefore we study the safety control of a system modelled by a POMDP based on these information states and define a safety specification as a convex set $\boldsymbol{\Psi}_s$ in which the information states of the system must lie. Our primary interests are the set $\boldsymbol{\Psi}_{in} \subset \boldsymbol{\Psi}_s$ and the admissible policy $\pi \in \Pi$ such that if the system starts with any initial probability distribution in $\boldsymbol{\Psi}_{in}$, then under $\pi$ the future information states will be in $\boldsymbol{\Psi}_s$ with probability 1 at each time step. Here $\pi$ is called an *a.s.* (*almost surely*) *safe policy* and $\boldsymbol{\Psi}_{in}$ is called an *a.s. safe set under* $\pi$.

We propose in Section 6.2 an algorithm to characterize a set $\boldsymbol{\Psi}_L$ which is an a.s. safe set under our conditions. The property of $\boldsymbol{\Psi}_L$ is that any other a.s. safe set under some a.s. safe policy is a subset of $\boldsymbol{\Psi}_L$. If under some policy the

a.s. safe set $\boldsymbol{\Psi}_{in}$ equals the safety specification $\boldsymbol{\Psi}_s$, then we called this policy induces an *a.s. safety enforcing controller*. Since the class of deterministic polices plays a crucial role in the optimal control of the POMDP model, we study in Section 6.3 the necessary and sufficient condition for a deterministic policy to induce an a.s. safety enforcing controller for a given $\boldsymbol{\Psi}_s$. We present under some hypothesis a linear programming type of characterization for that condition.

Finally a numerical simulation is implemented for a machine replacement problem. We comment on the simulation result and discuss the possibly supplemental role of the safety control to the optimal control of the POMDP model.

## 6.2 Almost Surely Safe Set

Let the Borel measurable map $\pi : \boldsymbol{\Psi} \to \boldsymbol{U}$ be a deterministic policy. As mentioned in Section 6.1, a set $\boldsymbol{\Psi}_{in} \subset \boldsymbol{\Psi}_s$ is called an a.s. safe set under $\pi$ if for all $\psi \in \boldsymbol{\Psi}_{in}$ and $k \in \mathbb{N}_0$

$$\mathbb{P}_{\psi}^{\pi}(\psi_k \in \boldsymbol{\Psi}_s) = 1\,. \tag{6.1}$$

In (6.1) $\pi$ is called an a.s. safe policy. Define

$$T(\psi, u) := \bigcup_{y \in \boldsymbol{Y}} T(\psi, y, u) = \bigcup_{y \in \boldsymbol{Y}} \frac{\psi \cdot Q(y, u)}{\psi \cdot Q(y, u) \cdot \mathbf{1}}\,. \tag{6.2}$$

and consider the following algorithm:

**Algorithm 6.2.1.** For $k = 0, 1, \cdots$

$$\boldsymbol{\Psi}^{(0)} := \{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A} \leq \boldsymbol{b}\} = \boldsymbol{\Psi}_s \,, \tag{6.3}$$

$$\begin{aligned} \mathcal{F}_u^{(k)} &:= \{\psi \in \boldsymbol{\Psi}^{(k-1)} | T(\psi, u) \subset \boldsymbol{\Psi}^{(k-1)}\} \\ &= \bigcap_{y \in \boldsymbol{Y}} \left[\boldsymbol{\Psi}^{(k-1)} \cap (T(\boldsymbol{\Psi}^{(k-1)}, y, u))^{(-1)}\right] \,, \end{aligned} \tag{6.4}$$

$$\boldsymbol{\Psi}^{(k)} := \bigcup_{u \in \boldsymbol{U}} \mathcal{F}_u^{(k)} \,, \tag{6.5}$$

$$\boldsymbol{\Psi}^{(\infty)} := \bigcap_{k \geq 0} \boldsymbol{\Psi}^{(k)} \,. \tag{6.6}$$

Some properties regarding Algorithm 6.2.1 are in the following.

**Lemma 6.2.2.** *For $k = 0, 1, \cdots$, the sets $\boldsymbol{\Psi}^{(k)}$ defined in Algorithm 6.2.1 are closed.*

*Proof.* We use induction to prove this lemma. $\boldsymbol{\Psi}^{(0)}$ is closed by (6.3). Suppose $\boldsymbol{\Psi}^{(k-1)}$ is closed. If $\{\psi_n\} \subset \boldsymbol{\Psi}^{(k)}$ is a sequence converging to some $\overline{\psi}$, we have $\overline{\psi} \in \overline{\boldsymbol{\Psi}^{(k)}}$, the closure of $\boldsymbol{\Psi}^{(k)}$, and $\overline{\boldsymbol{\Psi}^{(k)}} \subset \boldsymbol{\Psi}^{(k-1)}$ by (6.4) and (6.5). Also, by (6.5) there exists a sequence $\{u_n\} \subset \boldsymbol{U}$ such that $T(\psi_n, u_n) \subset \boldsymbol{\Psi}^{(k-1)}$. If $\overline{u}$ is a limit point of $\{u_n\}$, then the continuity of $(u, \psi) \to T(\psi, y, u)$ along with the assumption that $\boldsymbol{\Psi}^{(k-1)}$ is closed imply $T(\overline{\psi}, \overline{u}) \subset \boldsymbol{\Psi}^{(k-1)}$. Thus, by (6.4) and (6.5) $\overline{\psi} \in \mathcal{F}_u^{(k)} \subset \boldsymbol{\Psi}^{(k)}$. That means $\boldsymbol{\Psi}^{(k)}$ is also closed and the proof is completed. □

**Lemma 6.2.3.** *Suppose $\boldsymbol{\Psi}^{(\infty)} \neq \emptyset$. Let*

$$\mathcal{F}_u^{(\infty)} := \{\psi \in \boldsymbol{\Psi}^{(\infty)} | T(\psi, u) \subset \boldsymbol{\Psi}^{(\infty)}\} \,. \tag{6.7}$$

*Then*

$$\boldsymbol{\Psi}^{(\infty)} = \bigcup_{u \in \boldsymbol{U}} \mathcal{F}_u^{(\infty)} \,. \tag{6.8}$$

*Proof.* Suppose not, then there exists a $\psi' \in \boldsymbol{\Psi}^{(\infty)}$ such that $T(\psi', u)$ does not belong to $\boldsymbol{\Psi}^{(\infty)}$ for all $u \in \boldsymbol{U}$. That means for each $u \in \boldsymbol{U}$ there exists $y_n \in \boldsymbol{Y}$ such that $T(\psi', y, u) \notin \boldsymbol{\Psi}^{(\infty)}$. By the continuity of $u \to T(\psi, y, u)$ for each $u \in \boldsymbol{U}$ there exists an open neighborhood $V_u$ such that $u \in V_u$ and $T(\psi', y_u, u') \notin \boldsymbol{\Psi}^{(\infty)}$ for all $u' \in \overline{V_u}$. Let $V_{u_1}, \cdots, V_{u_l}$ be a finite subcover of the cover $\{V_u\}_{u \in \boldsymbol{U}}$ for $\boldsymbol{U}$. Then the set

$$\mathcal{T}(\psi') := \bigcup_{i=1}^{l} \bigcup_{u \in \overline{V_{u_i}}} T(\psi', y_{u_i}, u) \tag{6.9}$$

is compact and $\mathcal{T}(\psi') \cap \boldsymbol{\Psi}^{(\infty)} = \emptyset$. Since $S^{(k)}$ is apparently bounded and, by Lemma 6.2.2, is closed, we obtain that $\{S^{(k)}\}_{k=1}^{\infty}$ is a sequence of compact sets and thus $\mathcal{T}(\psi') \cap \boldsymbol{\Psi}^{(k')} = \emptyset$ for some finite $k'$. That means $\psi' \notin \mathcal{F}_u^{(k'+1)}$ for all $u \in \boldsymbol{U}$ so $\psi' \notin \boldsymbol{\Psi}^{(k'+1)}$. But $\psi' \in \boldsymbol{\Psi}^{(\infty)} \subset \boldsymbol{\Psi}^{(k'+1)}$ so the contradiction results. □

Let $2^{\boldsymbol{U}}$ represent the collection of all nonempty subset of $\boldsymbol{U}$. We conclude this section with the following theorem. Some technical background can be seen in Appendix D.

**Theorem 6.2.4.** *Suppose $\boldsymbol{\Psi}^{(\infty)} \neq \emptyset$. Let $\mathcal{G} : \boldsymbol{\Psi}^{(\infty)} \to 2^{\boldsymbol{U}}$ be the set-valued map defined by*

$$\mathcal{G}(\psi) := \{u \in \boldsymbol{U} | \psi \in \mathcal{F}_u^{(\infty)}\}. \tag{6.10}$$

*Let $\pi : \boldsymbol{\Psi}^{(\infty)} \to \boldsymbol{U}$ be any measurable selection from $\mathcal{G}$. Then, $\boldsymbol{\Psi}^{(\infty)}$ is an a.s. safe set under $\pi$. Conversely, if $\boldsymbol{\Psi}_{in}$ is an a.s. safe set under some deterministic policy, then $\boldsymbol{\Psi}_{in} \subset \boldsymbol{\Psi}^{(\infty)}$.*

*Proof.* The first assertion follows from Lemma 6.2.3. The second assertion is evident, since from the structure of Algorithm 6.2.1 every safe set is contained in $\boldsymbol{\Psi}^{(k)}$ for each $k \in \mathbb{N}_0$. □

## 6.3 Almost Surely Safety Enforcing Controller

If in (6.1) $\boldsymbol{\Psi}_{in}$ equals $\boldsymbol{\Psi}_s$, then the policy $\pi$ is called to induce an a.s. safety enforcing controller. Here we present a necessary and sufficient condition for a given deterministic policy $\pi$ to induce an a.s. safety enforcing controller. Suppose the safety specification is $\boldsymbol{\Psi}_s = \{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A} \leq \boldsymbol{b}\}$ as defined in (5.1). Define

$$\begin{aligned} B_y(i) := \max_{\psi} \quad & \frac{\psi \cdot Q(y, \pi(\psi)) \cdot \boldsymbol{A}_{\cdot i}}{\psi \cdot Q(y, \pi(\psi)) \cdot \mathbf{1}} \\ \text{subject to} \quad & \psi \in \boldsymbol{\Psi}\,, \\ & \psi \boldsymbol{A} \leq \boldsymbol{b}\,. \end{aligned} \tag{6.11}$$

It is easy to see $B_y(i) \leq \boldsymbol{b}(i)$ for all $y \in \boldsymbol{Y}$ and $i \in \boldsymbol{X}$ if and only if $\pi$ induces an a.s. safety enforcing controller corresponding to $\boldsymbol{\Psi}_s$. (6.11) is a nonlinear programming due to the fractional objective function and the dependence on $\pi$.

Suppose the action space $\boldsymbol{U}$ is finite and the set $\{\psi \in \boldsymbol{\Psi} | \pi(\psi) = u\}$ can be expressed by a polyhedral convex set $\{\psi \in \boldsymbol{\Psi} | \psi \boldsymbol{A}_u^* \leq \boldsymbol{b}_u^*\}$ for each $u \in \boldsymbol{U}$, where $\boldsymbol{A}_u^* \in \mathbb{R}^{N_x \times N_{b_u^*}}$ and $\boldsymbol{b}_u^*$ is a row vector with size $N_{b_u^*}$, then we can apply the following transformation to convert the original nonlinear programming problem to a linear programming problem. Specifically, let $\hat{\psi}$ be a nonnegative row vector with the same size as that of $\psi$ and satisfy

$$\hat{\psi} = t\psi\,,$$

$$\hat{\psi} \cdot Q(y, \pi(\psi)) \cdot \mathbf{1} = 1\,,$$

for some nonnegative $t$. Define

$$\begin{aligned} B_{yu}(i) := \max_{\tilde{\psi}} \quad & \tilde{\psi} \cdot c_{yu}(i) \\ \text{subject to} \quad & K\tilde{\psi}^T \leq \mathbf{0}_{N_x \times 1}\,, \\ & \Gamma_{yu}\tilde{\psi}^T = [1 \quad 0]^T\,, \\ & \tilde{\psi}^T \geq \mathbf{0}_{N_x \times 1}\,, \end{aligned} \tag{6.12}$$

where

$$\begin{aligned} \tilde{\psi} &= [\hat{\psi} \quad t]\,, \\ c_{yu}(i) &= [(Q(y,u) \cdot \boldsymbol{A}_{\cdot i})^T \quad 0]^T\,, \\ K &= \begin{bmatrix} \boldsymbol{A}^T & -\boldsymbol{b}^T \\ \boldsymbol{A}_u^{*T} & -\boldsymbol{b}_u^{*T} \end{bmatrix}\,, \\ \Gamma_{yu} &= \begin{bmatrix} (Q(y,u) \cdot \mathbf{1})^T & 0 \\ \mathbf{1}_{1 \times N_x} & -1 \end{bmatrix}\,. \end{aligned}$$

We conclude that the policy $\pi$ induces an a.s. safety enforcing controller if and only if in (6.12) $B_{yu}(i) \leq \boldsymbol{b}(i)$ for each $i \in \boldsymbol{X}$, $y \in \boldsymbol{Y}$ and $u \in \boldsymbol{U}$.

## 6.4 A Machine Replacement Example

In this section we study a machine replacement example in detail. The setting is in the following.

- $\boldsymbol{X} = \boldsymbol{Y} = \{0, 1\}$, where '0' denotes that the system is down and '1' good.
- $\boldsymbol{U} = \{0, 1\}$, where '0' means to continue the use of the machine and '1' to replace the machine.

- The state transition probability of the system $P(u)$ is specified in the following:

$$P(0) = \begin{bmatrix} 1-\eta & \eta \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad P(1) = \begin{bmatrix} 1 & 0 \\ \alpha & 1-\alpha \end{bmatrix}$$

  where $\eta \in (0,1)$ is the failure rate of the system and $\alpha \in (0.5,1]$ is the replacement degree when the replacement action is performed.

- The correct observation rate is $q \in (0.5,1]$ and the transition law of the partially observed system is expressed by

$$Q(0,0) = \begin{bmatrix} 1-\eta & \eta \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} q & 0 \\ 0 & 1-q \end{bmatrix},$$

$$Q(1,0) = \begin{bmatrix} 1-\eta & \eta \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1-q & 0 \\ 0 & q \end{bmatrix},$$

$$Q(0,1) = \begin{bmatrix} 1 & 0 \\ \alpha & 1-\alpha \end{bmatrix} \cdot \begin{bmatrix} q & 0 \\ 0 & 1-q \end{bmatrix},$$

$$Q(1,1) = \begin{bmatrix} 1 & 0 \\ \alpha & 1-\alpha \end{bmatrix} \cdot \begin{bmatrix} 1-q & 0 \\ 0 & q \end{bmatrix}.$$

The information state space $\boldsymbol{\Psi}$ is parameterized in the form $\psi = [\,1-p \quad p\,]$ where $p \in [\,0,1\,]$ is the probability that the system is down. Thus the corresponding $V(p,y,u)$ and $T(p,y,u)$ for each each $y \in \boldsymbol{Y}$ and $u \in \boldsymbol{U}$ can be obtained by

$$V(p,y,u) = \psi \cdot Q(y,u) \cdot \mathbf{1}$$

and

$$T(p,y,u) = \begin{cases} \frac{\psi \cdot Q(p,y,u) \cdot [0 \;\; 1]^T}{V(p,y,u)} & \text{if } V(p,y,u) \neq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Specifically,

$$
\begin{aligned}
V(p,0,0) &= -p(1-\eta)(2q-1) + (1-q)\eta + q(1-\eta)\,; \\
V(p,1,0) &= 1 - V(p,0,0)\,; \\
V(p,0,1) &= q - p(1-\alpha)(2q-1)\,; \\
V(p,1,1) &= 1 - V(p,0,1)\,; \\
T(p,0,0) &= \frac{p(1-\eta)(1-q) + \eta(1-q)}{(1-q)\eta + q(1-\eta) - p(1-\eta)(2q-1)}\,; \\
T(p,1,0) &= \frac{pq(1-\eta) + q\eta}{p(1-\eta)(2q-1) + q\eta + (1-q)(1-\eta)}\,; \\
T(p,0,1) &= \frac{p(1-q)(1-\alpha)}{q - p(1-\alpha)(2q-1)}\,; \\
T(p,1,1) &= \frac{pq(1-\alpha)}{(1-q) + p(1-\alpha)(2q-1)}\,.
\end{aligned}
$$

From above expressions, we can show that

1. $T(\cdot, y, u)$ is nondecreasing on $[0,1]$ for every $y, u \in \{0,1\}$.

2. $T(\cdot, 0, u)$ is convex and $T(\cdot, 1, u)$ is concave both on $[0,1]$ for every $u \in \{0,1\}$.

Furthermore, for all $p \in [0,1]$,

$$
\begin{aligned}
T(p,1,0) - T(p,0,0) &= \frac{[p + (1-p)\eta](1-\eta)(1-p)(2q-1)}{V(p,1,0)V(p,0,0)} \geq 0\,, \\
T(p,1,0) - T(p,1,1) &= \frac{(1-q)q\{p\alpha(1-\eta) + \eta[1-p(1-\alpha)]\}}{V(p,1,0)V(p,1,1)} \geq 0\,, \\
T(p,1,1) - T(p,0,1) &= \frac{p(1-\alpha)(2q-1)(1-p(1-\alpha))}{V(p,1,1)V(p,0,0)} \geq 0\,,
\end{aligned}
$$

$$T(p,1,0)-p=-\frac{(p(1-\eta)(2q-1)+q\eta)(p-1)}{V(p,1,0)}\geq 0\,,$$
$$T(p,0,1)-p=\frac{(1-\alpha)(2q-1)p(p-\frac{\frac{q}{1-\alpha}+q-1}{q+q-1})}{V(p,0,1)}\leq 0\,,$$
$$T(p,1,1)-p=-\frac{(1-\alpha)(2q-1)p(p-\frac{q-\frac{1-q}{1-\alpha}}{q-(1-q)})}{V(p,1,1)}$$
$$\Rightarrow\begin{cases}\geq 0 & \text{if} \quad q\geq\frac{1}{2-\alpha}, \quad \text{and} \quad p\leq\frac{q(1-\alpha)-(1-q)}{(1-\alpha)(2q-1)}\\ \leq 0 & \text{otherwise}\end{cases}.$$

Suppose the safety specification is $\boldsymbol{\Psi}_s=\{[1-p\ p\,]|0\leq p\leq B\}$. Define the criterion $p^*$ for the policy $\pi_{p^*}$ in the following:

$$\pi_{p^*}(p)=\begin{cases}0 & \text{if } 0\leq p\leq p^*\\ 1 & \text{otherwise}\end{cases}.$$

Also, define $T^{-1}(B,y,u)$ such that

$$T(T^{-1}(B,y,u),y,u)=B\,.$$

Based on above analysis, we plot in Figure 6.1 the relation between the priori probability $p$ and the posteriori probability $T(p,y,u)$ for various $y$ and $u$ and classify all the possibilities into following cases.

*Case 1:*

$$q\leq\frac{1}{2-\alpha}$$

or

$$q\geq\frac{1}{2-\alpha}\quad\text{and}\quad\frac{q(1-\alpha)-(1-q)}{(1-\alpha)(2q-1)}\leq B\,.$$

In this case $\boldsymbol{\Psi}_s$ is the largest safe set and there exists a class of deterministic safe policies $\pi_{p^*}$ corresponding to it. Since for each $p\in[0\ 1]$ we

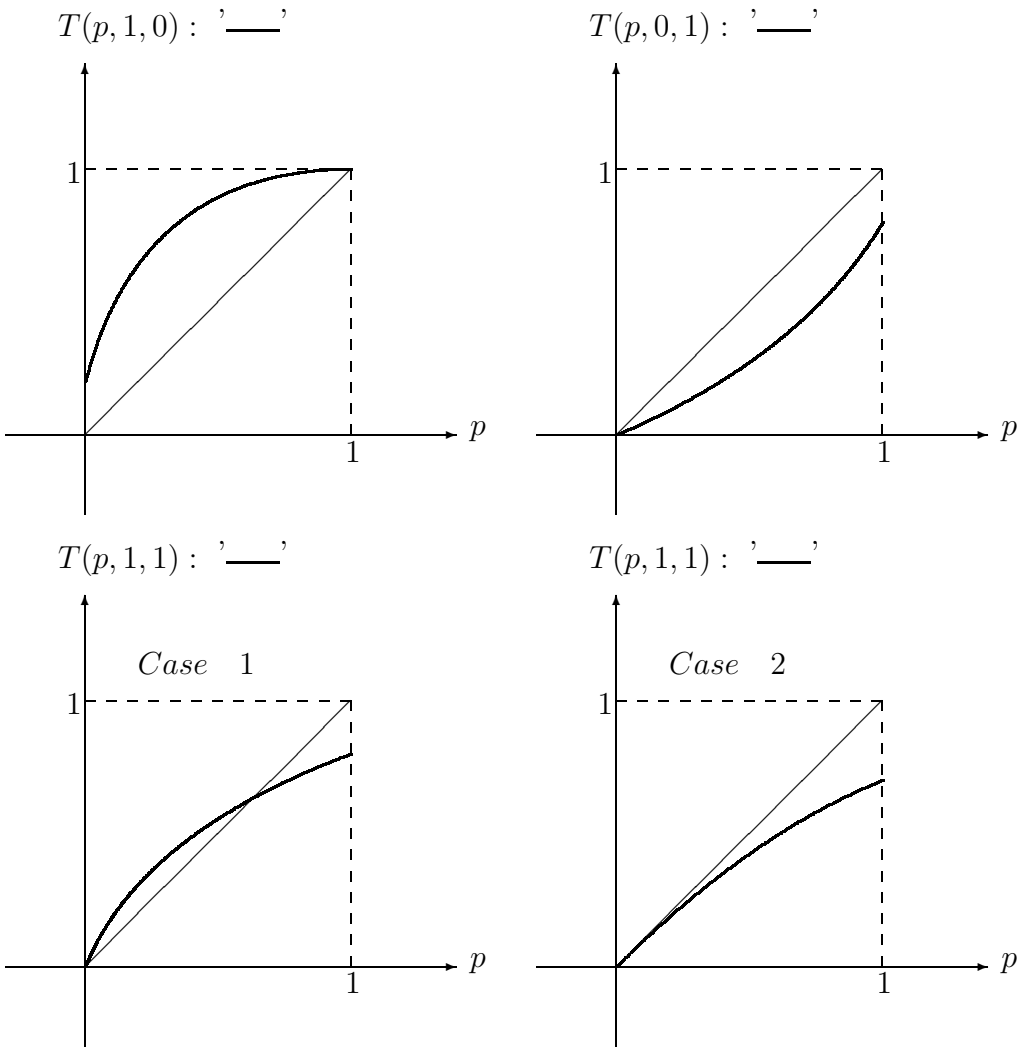


Figure 6.1: the analytical relation for the priori probability $p$ and posteriori probability $T(p, y, u)$ for various $y$ and $u$

have $\max_{y\in \boldsymbol{Y},u\in \boldsymbol{U}} T(p,y,u) = T(p,1,0)$. Those policies can be characterized by asking $p^* \leq T^{-1}(B,1,0)$ where

$$T^{-1}(t,1,0) = \begin{cases} \frac{t(q\eta+(1-q)(1-\eta))-q\eta}{(1-\eta)(q-t(2q-1))} & \text{if } t \in [0,1] \\ 0 & \text{otherwise} \end{cases}.$$

*Case 2:*

$$q \geq \frac{1}{2-\alpha} \quad \text{and} \quad 0 \leq B \leq \frac{q(1-\alpha)-(1-q)}{(1-\alpha)(2q-1)}.$$

In this case $[0, T^{-1}(B,1,1)]$ is the largest safe set under the policies $\pi_{p^*}$ where $p^* \leq T^{-1}(T^{-1}(B,1,1),1,0)$ and

$$T^{-1}(t,1,1) = \begin{cases} \frac{t(1-q)}{q(1-\alpha)-t(1-\alpha)(2q-1)} & \text{if } t \in [0,1] \\ 0 & \text{otherwise} \end{cases}.$$

Note that both $T^{-1}(t,1,0)$ and $T^{-1}(t,1,1)$ are increasing functions in $t$. In order to obtain $T^{-1}(t,1,0) \geq 0$, we need

$$t \geq \frac{q\eta}{q\eta+(1-q)(1-\eta)}.$$

So in *Case 2* we need the following condition

$$\frac{B(1-q)}{(1-\alpha)(q-B(2q-1))} \geq \frac{q\eta}{q\eta+(1-q)(1-\eta)}$$

to get a nonnegative $p^*$.

In Figure 6.2 we present an example of an simulated relation between the deterministic policies and the incurred average costs. The used parameters are: system's failure rate $\eta = 0.2$; correct observation rate $q = 0.8$; replacement degree $\alpha = 0.8$; initial probability $p$ that the system is down: 0; replacement cost $R = 180$; the cost $C$ that the system is down but no replacement is

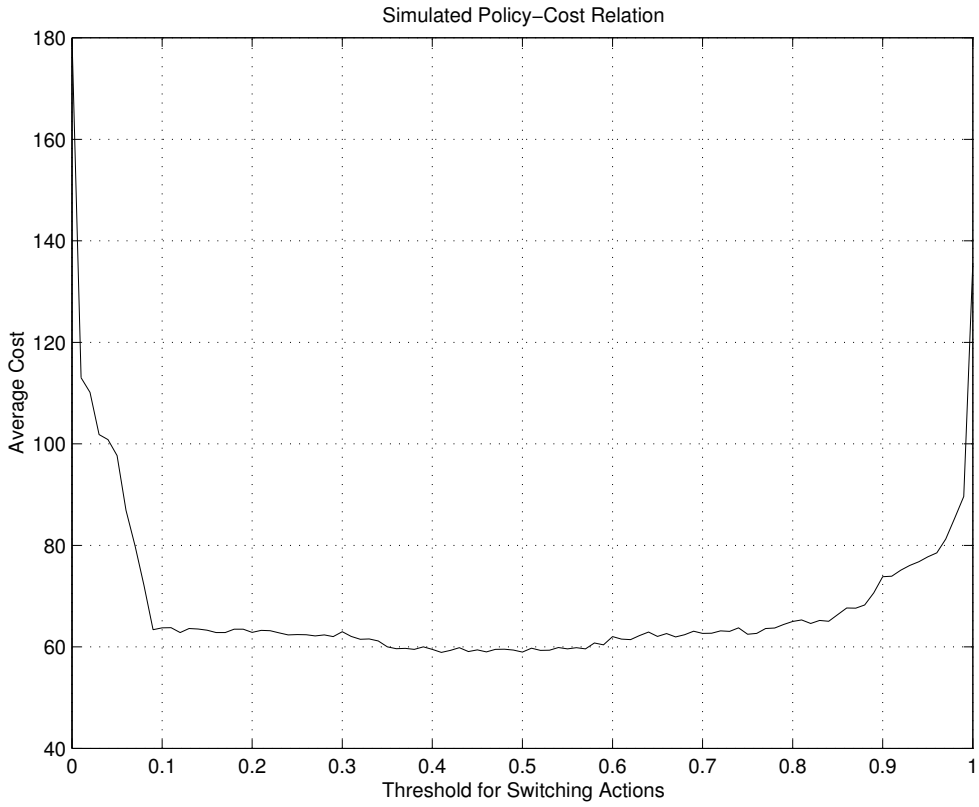


Figure 6.2: the simulated relation between the stationary policies and their incurred average costs. (The used parameters are $\eta = 0.2$, $q = 0.8$, $\alpha = 0.8$, $p = 0$, $R = 180$, and $C = 150$. The $MatLab^{©}$ random number generator is used and the time horizon is calculated up to 30,000 steps.)

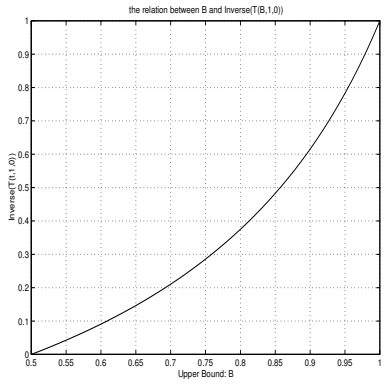


Figure 6.3: the relation between the upper bound $B$ and $T^{-1}(B, 1, 0)$ with the same parameters as those used for Figure 6.2

performed: 150. The $MatLab^{©}$ random number generator is used and time horizon is calculated up to 30,000 steps. As the simulation shows, the optimal policy suggests the threshold $p^* \approx 0.41$, which incurs the minimal average cost around 59. As for the safety control, since $q = 0.8 \leq \frac{1}{2-\alpha} = 0.833$, these parameters belong to *Case 1*. The relation between $T^{-1}(p, 1, 0)$ and $p$ is plotted in Figure 6.3 with the same parameters. In order to make the optimal policy safe with probability 1, the safety specification, i.e. the upper bound $B$, can not be less than

$$T(p^*, 1, 0) = \frac{p^* q(1-\eta) + q\eta}{p^*(1-\eta)(2q-1) + q\eta + (1-q)(1-\eta)} \approx 0.8173\,.$$

It is noted that as the threshold is set between 0.09 and 0.82, the incurred costs differ not much. It means that if we would like to pay around 64, not much more than the minimal cost 59, then we can set the threshold at around 0.09 and control the system so that the probability of the system being down never exceed 0.6, if the starting probability of the system being down is also less than 0.6.

# Chapter 7

# Conclusions and Future Work

In this dissertation we have dealt with several topics regarding the discrete-time Markov decision process. Some of the topics have been studied for decades, for example, ergodic control and zero-sum stochastic game, but the study of others is just beginning, for example, safety control. In this chapter we give our conclusion and point out possible directions for future research.

## 7.1 Single and Dual Controllers Ergodic Control

In Chapter 2 we proposed an assumption to prove the existence of Bellman's average cost optimality equation, which functionally characterizes the optimal value and policy for the controller. An example was analyzed in detail to compare our assumption with those proposed before. We concluded that our approach is able to handle more cases and is thus less restrictive.

An interesting future topic is the relationship between our assumption and those classical assumptions. Even though our assumption seems to handle more problems, it is not clear whether there exists an implication between them.

We considered in Chapter 3 the two-controller zero-sum partially observable game problems. The feature of the class of problems we studied is

that there exists a state that is completely observable. We show the existence of the game's value and the optimal policies, which are characterized by Shapley's average cost equation, for both controllers. Actually, the completely observable state serves as the ergodic information state that is visited in every time step with positive probability. That is the main condition contributing to the problem's tractability. It would be interesting to know whether similar results can be obtained when the complete observation condition is released. Note that for the optimal control in the POMDP model, this release is possible due to the concavity of the $\beta$-discounted value function. In the POMG model, however, the lack of the convex (or concave) property in the $\beta$-discounted value function makes the release very difficult, if not impossible.

## 7.2 Adaptive and Safety Control

In Chapter 4 the adaptive control for the POMDP model was explored. We followed the methodology of [31] to treat this topic and provide justifiable ergodic assumption, which implies the condition necessary in the methodology. In particular, our approach could be regarded as an alternative condition to that in [17]. With further assumptions on the structure and properties of the parameter space, we were able to show that the proposed adaptive control policy is self-optimizing in an appropriate sense.

In Chapter 5 we studied the safety control of a stochastic system modelled by the discrete-time completely observable finite MDP. The goal of safety control was specified as a convex set, called the safety specification, in which the system's state probability distribution vector must lie at each time step.

By appropriate choices of policy and initial state probability distribution, a controlled system is able to achieve this goal. We first obtained a necessary and sufficient condition for a safe transition matrix such that the safety is enforced; that is, with this transition matrix, the state probability distribution of the system is always safe (inside the safety specification) if the initial state probability distribution of the system is safe. Next we searched a safe policy that guarantees that the long-term state probability distribution is safe. In particular, we introduced a cost function as in the MDP model and identified the optimal safe policy in the sense of a long-run average among all safe policies. The problem was formulated as a linear programming and its feasibility was analyzed. Finally, given a safe set, we provided algorithms to characterize the largest set such that a system starting from an initial state probability distribution inside that set would have its future state probability distributions remain safe. The algorithms were shown to terminate in finite steps under mild conditions. In particular, we proved that the algorithms terminate in one iteration when the system has only two states.

However, even though we could provide a finite upper bound on the number of iterations of the algorithm, practically speaking, the algorithm still generates too many inequalities to characterize the supremal invariant safe set. One possible direction for future studies is to identify a policy that makes the safety specification itself an invariant safe set, that is, to identify a policy that induces a safety enforcing controller. If the policy is not unique, we can introduce a cost function as in the MDP model and identify the optimal policy in the sense of a long-run average among those that induce safety enforcing controllers. The expected challenge is that the solution of our new problem

involves solving a nonlinear programming and the existence of an exact solution is not clear. Another possible approach for alleviating the *dimension curse* of the algorithm is to conjecture a possible redundancy of inequalities created in running the algorithm. From Section 5.5.2 we showed that as the size of the transition matrix is $2\times2$, only 1 iteration is enough to characterize the supremal invariant safe set, while theoretically the upper bound we provided for the number of iterations of the algorithm could be much greater (depending on each element in the transition matrix and the $\varepsilon$ value). This discovery leads naturally to the question about the relation between the number of redundant inequalities and the size of the transition matrix. If the relation can be found, the number of inequalities required to characterize the supremal invariant safe set is expected to be significantly smaller than that indicated by the upper bound of iterations of the algorithm.

Chapter 6 is a continuation of our study on the safety control of stochastic discrete-time event systems. We explored in this chapter the case when the stochastic systems are modelled by the partially observed Markov decision processes. The safety specification is defined by a convex set in which the information states of the system must lie. We focus on a type of safety called a.s. (almost surely) safety. Our primary interests are a subset of the safety specification and an admissible policy such that if the system starts with any initial state probability distribution in the subset, then under this policy the following information states will be in the safety specification with probability 1 at each time step. We first proposed an algorithm to characterize a set which is an a.s. safe set under some conditions. The property of the set is that any other a.s. safe set under some a.s. safe policy is its subset. Next we studied

the necessary and sufficient conditions for a deterministic policy to induce an a.s. safety enforcing controller for a given safety specification; that is, under this policy the whole safety specification is an a.s. safe set. We presented under our hypothesis a linear programming type of characterization for that condition. Finally, a numerical simulation was implemented for a machine replacement problem. We commented on the simulation result and discussed the possibly supplemental role of the safety control to the optimal control of the POMDP model.

A possible direction of future work concerning Chapter 6 is the study of another safety concept called *safety in the mean*. In that concept the safety specification is defined as a convex set in which the expected value of the information states must lie in at each time step. It would be interesting to study the safe policy and the safe set in this new sense.

# Appendix A

# Proof of Lemma 2.2.1

*Proof.* Let $u_*^{k-1}$ denote the decisions with respect to the starting state $\psi_*$, the optimal strategy $\pi_\beta$, and the observations $y^{k-1}$. Define $\boldsymbol{Y}^{k*} \subset \boldsymbol{Y}^k$ such that $V(\psi^*, y^k, u_*^{k-1}) \neq 0 \ \forall \ y^k \in \boldsymbol{Y}^{k*}$ and define $\boldsymbol{Y}_*^k \subset \boldsymbol{Y}^k$ such that $V(\psi_*, y^k, u_*^{k-1}) \neq 0 \ \forall \ y^k \in \boldsymbol{Y}_*^k$. Since

$$h_\beta(\psi_*) \geq \beta^k \sum_{y^k \in \boldsymbol{Y}_*^k} V(\psi_*, y^k, u_*^{k-1}) h_\beta(T(\psi_*, y^k, u_*^{k-1}))$$

and

$$\begin{aligned} h_\beta(\psi^*) \leq \|c\|_\infty + \beta\|c\|_\infty + \cdots + \beta^{k-1}\|c\|_\infty \\ + \beta^k \sum_{y^k \in \boldsymbol{Y}^{k*}} V(\psi^*, y^k, u_*^{k-1}) h_\beta(T(\psi^*, y^k, u_*^{k-1}))\,, \end{aligned}$$

under the assumption, there exists $\boldsymbol{Y}_c^k \subset \boldsymbol{Y}_*^k \cap \boldsymbol{Y}^{k*}$ such that if we define

$$\tilde{\psi}(y^k) := \frac{T(\psi_*, y^k, u_*^k) - \varepsilon \cdot T(\psi^*, y^k, u_*^k)}{1 - \varepsilon}\,,$$

then $\tilde{\psi}(y^k) \in \boldsymbol{\Psi} \ \forall \ y^k \in \boldsymbol{Y}_c^k$.

Since $h_\beta(\cdot)$ is concave in $\psi$, we can write

$$\begin{aligned}
\Delta h \le N_0\|c\|_\infty &+ \beta^k \sum_{y^k \in \boldsymbol{Y}_c^k} V(\psi^*, y^k, u_*^{k-1}) h_\beta(T(\psi^*, y^k, u_*^{k-1})) \\
&- \beta^k \sum_{y^k \in \boldsymbol{Y}_c^k} V(\psi_*, y^k, u_*^{k-1})[(1-\varepsilon)h_\beta(\tilde{\psi}(y^k)) + \varepsilon \cdot h_\beta(T(\psi^*, y^k, u_*^{k-1}))] \\
&+ \beta^k \sum_{y^k \in \boldsymbol{Y}^{k*} \setminus \boldsymbol{Y}_c^k} V(\psi^*, y^k, u_*^{k-1}) h_\beta(T(\psi^*, y^k, u_*^{k-1})) \\
&- \beta^k \sum_{y^k \in \boldsymbol{Y}_*^k \setminus \boldsymbol{Y}_c^k} V(\psi_*, y^k, u_*^{k-1}) h_\beta(T(\psi_*, y^k, u_*^{k-1}))
\end{aligned}$$

$$\begin{aligned}
= N_0\|c\|_\infty &+ \beta^k \sum_{y^k \in \boldsymbol{Y}_c^k} (V(\psi^*, y^k, u_*^{k-1}) - \varepsilon \cdot V(\psi_*, y^k, u_*^{k-1})) h_\beta(T(\psi^*, y^k, u_*^{k-1})) \\
&- \beta^k \sum_{y^k \in \boldsymbol{Y}_c^k} V(\psi_*, y^k, u_*^{k-1})(1-\varepsilon) h_\beta(\tilde{\psi}(y^k)) \\
&+ \beta^k \sum_{y^k \in \boldsymbol{Y}^{k*} \setminus \boldsymbol{Y}_c^k} V(\psi^*, y^k, u_*^{k-1}) h_\beta(T(\psi^*, y^k, u_*^{k-1})) \\
&- \beta^k \sum_{y^k \in \boldsymbol{Y}_*^k \setminus \boldsymbol{Y}_c^k} V(\psi_*, y^k, u_*^{k-1}) h_\beta(T(\psi_*, y^k, u_*^{k-1}))
\end{aligned}$$

$$\begin{aligned}
= N_0\|c\|_\infty &+ \beta^k \sum_{y^k \in \boldsymbol{Y}_c^k} (V(\psi^*, y^k, u_*^{k-1}) - \varepsilon \cdot V(\psi_*, y^k, u_*^{k-1})) \\
&\qquad \times \{h_\beta(T(\psi^*, y^k, u_*^{k-1})) - h_\beta(\psi_*)\} \\
&- \beta^k \sum_{y^k \in \boldsymbol{Y}_c^k} V(\psi_*, y^k, u_*^{k-1})(1-\varepsilon)\{h_\beta(\tilde{\psi}(y^k)) - h_\beta(\psi_*)\} \\
&+ \beta^k \sum_{y^k \in \boldsymbol{Y}^{k*} \setminus \boldsymbol{Y}_c^k} V(\psi^*, y^k, u_*^{k-1})\{h_\beta(T(\psi^*, y^k, u_*^{k-1})) - h_\beta(\psi_*)\} \\
&- \beta^k \sum_{y^k \in \boldsymbol{Y}_*^k \setminus \boldsymbol{Y}_c^k} V(\psi_*, y^k, u_*^{k-1})\{h_\beta(T(\psi_*, y^k, u_*^{k-1})) - h_\beta(\psi_*)\}\,.
\end{aligned}$$

Let $\beta$=1, drop the minus terms and replace the appropriate terms in $\{\cdot\}$ with $\Delta h$, then we have

$$\begin{aligned}\Delta h &\le N_0\|c\|_\infty + \sum_{y^k \in \boldsymbol{Y}_c^k} (V(\psi^*, y^k, u_*^{k-1}) - \varepsilon \cdot V(\psi_*, y^k, u_*^{k-1}))\Delta h \\ &\quad + \sum_{y^k \in \boldsymbol{Y}^{k*} \setminus \boldsymbol{Y}_c^k} V(\psi^*, y^k, u_*^{k-1})\Delta h \\ &\le N_0\|c\|_\infty + (1-\varepsilon^2)\Delta h\,.\end{aligned}$$

Thus,

$$\overline{h}_\beta(\psi) \le \Delta h \le \frac{N_0\|c\|_\infty}{\varepsilon^2}\,,$$

and the proof is completed. □

# Appendix B

# Proofs Regarding Chapter 4

## B.1 Proof of Lemma 4.2.1

*Proof.* By the assumption on $B_m$ we have

$$\begin{aligned}
&(B_m)_{i_1 s} \geq \varepsilon \cdot (B_m)_{i_2 s}\,,\\
\Rightarrow &(B_m)_{i_1 s}(B_{m+1}^{m+r})_{sj} \geq \varepsilon \cdot (B_m)_{i_2 s}(B_{m+1}^{m+r})_{sj}\,,\\
\Rightarrow &\sum_s (B_m)_{i_1 s}(B_{m+1}^{m+r})_{sj} \geq \varepsilon \cdot \sum_s (B_m)_{i_2 s}(B_{m+1}^{m+r})_{sj}\,,\\
\Rightarrow &(B_m^{m+r})_{i_1 j} \geq \varepsilon \cdot (B_m^{m+r})_{i_2 j}\,,\\
\Rightarrow &\sum_j (B_m^{m+r})_{i_1 j} \geq \varepsilon \cdot \sum_j (B_m^{m+r})_{i_2 j}\,.
\end{aligned}$$

Since $B_m$ is row-allowable $\forall m \in \mathbb{N}$, so will be their product and the result follows. □

## B.2 Proof of Lemma 4.2.3

*Proof.* As $k$=1, the property holds by assumption. If it holds as $k = n$ then as $k = n$+1

$$\frac{(B^{n+1})_{i_1 .}\mathbf{1}}{(B^{n+1})_{i_2 .}\mathbf{1}} = \frac{\sum_j (B_1)_{i_1 j}(B_2^{n+1})_{j.}\mathbf{1}}{\sum_j (B_1)_{i_2 j}(B_2^{n+1})_{j.}\mathbf{1}} = \frac{(B_1)_{i_1 .}\mathbf{1}\sum_j \frac{(B_1)_{i_1 j}}{(B_1)_{i_1 .}\mathbf{1}}(B_2^{n+1})_{j.}\mathbf{1}}{(B_1)_{i_2 .}\mathbf{1}\sum_j \frac{(B_1)_{i_2 j}}{(B_1)_{i_2 .}\mathbf{1}}(B_2^{n+1})_{j.}\mathbf{1}}$$

$$\geq \varepsilon \cdot \frac{\sum_j \frac{(B_1)_{i_1 j}}{(B_1)_{i_1 .}\mathbf{1}}(B_2^{n+1})_{j.}\mathbf{1}}{\sum_j \frac{(B_1)_{i_2 j}}{(B_1)_{i_2 .}\mathbf{1}}(B_2^{n+1})_{j.}\mathbf{1}} \geq \varepsilon \cdot \frac{\min_j (B_2^{n+1})_{j.}\mathbf{1}}{\max_j (B_2^{n+1})_{j.}\mathbf{1}} \geq \varepsilon^{n+1}\,,$$

thus the result follows. □

## B.3 Proof of Lemma 4.2.4

*Proof.* Since

$$\frac{(B_{m+r}^{m+n})_{i_1 j}}{(B_{m+r}^{m+n})_{i_1 .}\mathbf{1}} - \frac{(B_{m+r}^{m+n})_{i_2 j}}{(B_{m+r}^{m+n})_{i_2 .}\mathbf{1}}$$

$$= \sum_s \left\{ \frac{(B_{m+r})_{i_1 s} \cdot (B_{m+r+1}^{m+n})_{s.}\mathbf{1}}{(B_{m+r}^{m+n})_{i_1 .}\mathbf{1}} - \frac{(B_{m+r})_{i_2 s} \cdot (B_{m+r+1}^{m+n})_{s.}\mathbf{1}}{(B_{m+r}^{m+n})_{i_2 .}\mathbf{1}} \right\} \frac{(B_{m+r+1}^{m+n})_{sj}}{(B_{m+r+1}^{m+n})_{s.}\mathbf{1}} .$$

Define

$$A_{i_1 s}^{rn} := \frac{(B_{m+r})_{i_1 s} \cdot (B_{m+r+1}^{m+n})_{s.}\mathbf{1}}{(B_{m+r}^{m+n})_{i_1 .}\mathbf{1}} , \quad A_{i_2 s}^{rn} := \frac{(B_{m+r})_{i_2 s} \cdot (B_{m+r+1}^{m+n})_{s.}\mathbf{1}}{(B_{m+r}^{m+n})_{i_2 .}\mathbf{1}} .$$

Since for $a = i_1$ or $i_2$ we have

$$(B_{m+r}^{m+n})_{a.}\mathbf{1} = \sum_s (B_{m+r})_{as} \cdot (B_{m+r+1}^{m+n})_{s.}\mathbf{1} ,$$

that is

$$\sum_s \left\{ A_{i_1 s}^{rn} - A_{i_2 s}^{rn} \right\} = 1 - 1 = 0 := \left( \sum_{s^+} + \sum_{s^-} \right) \left( A_{i_1 s}^{rn} - A_{i_2 s}^{rn} \right) ,$$

where

$$A_{i_1 s}^{rn} - A_{i_2 s}^{rn} \geq 0 \quad \forall s \in s^+ \quad \text{and} \quad A_{i_1 s}^{rn} - A_{i_2 s}^{rn} < 0 \quad \forall s \in s^- .$$

On the other hand, we have $A_{i_2 s}^{rn} \geq \varepsilon^2 A_{i_1 s}^{rn}$ from Corollary 4.2.2. So

$$0 \leq \sum_{s^+} \left\{ A_{i_1 s}^{rn} - A_{i_2 s}^{rn} \right\} \leq \sum_{s^+} (1 - \varepsilon^2) A_{i_1 s}^{rn} \leq (1 - \varepsilon^2) .$$

Finally

$$\frac{(B_{m+r}^{m+n})_{i_1 j}}{(B_{m+r}^{m+n})_{i_1 .}\mathbf{1}} - \frac{(B_{m+r}^{m+n})_{i_2 j}}{(B_{m+r}^{m+n})_{i_2 .}\mathbf{1}}$$

$$= \sum_{s^+} \left( A_{i_1 s}^{rn} - A_{i_2 s}^{rn} \right) \frac{(B_{m+r+1}^{m+n})_{sj}}{(B_{m+r+1}^{m+n})_{s.}} + \left( - \sum_{s^-} \left( A_{i_1 s}^{rn} - A_{i_2 s}^{rn} \right) \right) \left( - \frac{(B_{m+r+1}^{m+n})_{sj}}{(B_{m+r+1}^{m+n})_{s.}\mathbf{1}} \right)$$

$$\leq (1 - \varepsilon^2) \max_s \frac{(B_{m+r+1}^{m+n})_{sj}}{(B_{m+r+1}^{m+n})_{s.}\mathbf{1}} + (1 - \varepsilon^2) \left( - \min_s \frac{(B_{m+r+1}^{m+n})_{sj}}{(B_{m+r+1}^{m+n})_{s.}\mathbf{1}} \right) .$$

By induction on $r$ we have

$$\frac{(B_m^{m+n})_{i_1 j}}{(B_m^{m+n})_{i_1 \cdot}\mathbf{1}} - \frac{(B_m^{m+n})_{i_2 j}}{(B_m^{m+n})_{i_2 \cdot}\mathbf{1}} \leq (1-\varepsilon^2)^n \left\{ \max_s \frac{(B_{m+n})_{sj}}{(B_{m+n})_{s\cdot}\mathbf{1}} - \min_s \frac{(B_{m+n})_{sj}}{(B_{m+n})_{s\cdot}\mathbf{1}} \right\}.$$

Since the above argument holds for each $i_1, i_2 \in \{1, 2, \ldots, N_x\}$ and the difference in the big bracket is less than 1, we obtain the result. □

## B.4 Proof of Lemma 4.2.5

*Proof.* Suppose $\psi_1, \psi_2 \in \boldsymbol{\Psi}$ and $B_k$ is row-allowable for $k \in \mathbb{N}$. Denote $\psi_{is} = \psi_i(s)$, $i = 1, 2$. We have

$$\left\| \frac{\psi_1 B^n}{\psi_1 B^n \mathbf{1}} - \frac{\psi_2 B^n}{\psi_2 B^n \mathbf{1}} \right\|_1 = \sum_{i=1}^{N_x} \left| \sum_{s=1}^{N_x} \frac{\psi_{1s} B_{si}^n}{\psi_1 B^n \mathbf{1}} - \sum_{s=1}^{N_x} \frac{\psi_{2s} B_{si}^n}{\psi_2 B^n \mathbf{1}} \right|$$

$$:= \sum_{i=1}^{N_x} \left| \sum_{s=1}^{N_x} (\hat{\psi}_{1s} - \hat{\psi}_{2s}) \hat{B}_{si}^n \right| = \left\| (\hat{\psi}_1 - \hat{\psi}_2) \hat{B}^n \right\|_1 \leq \tau_1(\hat{B}^n) \left\| \hat{\psi}_1 - \hat{\psi}_2 \right\|_1$$

where

$$\hat{\psi}_{1s} := \frac{\psi_{1s} B_{s\cdot}^n \mathbf{1}}{\psi_1 B^n \mathbf{1}}, \quad \hat{\psi}_{2s} := \frac{\psi_{2s} B_{s\cdot}^n \mathbf{1}}{\psi_2 B^n \mathbf{1}}, \quad \hat{B}_{si}^n := \frac{B_{si}^n}{B_{s\cdot}^n \mathbf{1}}.$$

Since

$$\left\| \hat{\psi}_1 - \hat{\psi}_2 \right\|_1 \leq \left\{ \left\| \hat{\psi}_1 - \hat{\psi}_0 \right\|_1 + \left\| \hat{\psi}_2 - \hat{\psi}_0 \right\|_1 \right\},$$

if we let $\hat{\psi}_0 = [\hat{\psi}_{01}, \hat{\psi}_{02}, \ldots, \hat{\psi}_{0N_x}]$ with $s^{th}$ component $\frac{\psi_{2s} B_{s\cdot}^n \mathbf{1}}{\psi_1 B^n \mathbf{1}}$, then

$$\left\| \hat{\psi}_2 - \hat{\psi}_0 \right\|_1 = \sum_{s=1}^{N_x} \left| \frac{\psi_{2s} B_{s\cdot}^n \mathbf{1}}{\psi_2 B^n \mathbf{1}} - \frac{\psi_{2s} B_{s\cdot}^n \mathbf{1}}{\psi_1 B^n \mathbf{1}} \right| = \sum_{s=1}^{N_x} \left| \frac{(\psi_1 - \psi_2) B^n \mathbf{1} \psi_{2s} B_{s\cdot}^n \mathbf{1}}{\psi_1 B^n \mathbf{1} \psi_2 B^n \mathbf{1}} \right|$$

$$= \left| \frac{(\psi_1 - \psi_2) B^n \mathbf{1}}{\psi_1 B^n \mathbf{1}} \right| \leq \sum_{s=1}^{N_x} \left| \frac{\psi_{1s} B_{s\cdot}^n \mathbf{1}}{\psi_1 B^n \mathbf{1}} - \frac{\psi_{2s} B_{s\cdot}^n \mathbf{1}}{\psi_1 B^n \mathbf{1}} \right| = \left\| \hat{\psi}_1 - \hat{\psi}_0 \right\|_1.$$

That is,

$$\left\| \hat{\psi}_1 - \hat{\psi}_2 \right\|_1 \leq 2 \sum_{s=1}^{N_x} \left| \frac{(\psi_{1s} - \psi_{2s}) B_{s\cdot}^n \mathbf{1}}{\psi_1 B^n \mathbf{1}} \right| \leq 2 \sum_{s=1}^{N_x} |\psi_{1s} - \psi_{2s}| \cdot \max_s \left\{ \frac{B_{s\cdot}^n \mathbf{1}}{\psi_1 B^n \mathbf{1}} \right\}.$$

Since $\tau_1(B) \leq 1$ for a stochastic matrix $B$, if $(B_k)_{i_1\cdot}\mathbf{1} \geq \varepsilon \cdot (B_k)_{i_2\cdot}\mathbf{1}$ then from Lemma 4.2.3

$$\left\|\hat{\psi}_1 - \hat{\psi}_2\right\|_1 \leq \frac{2}{\varepsilon^n} \left\|\psi_1 - \psi_2\right\|_1 .$$

Thus part (1) is proved. If $(B_k)_{i_1 j} \geq \varepsilon \cdot (B_k)_{i_2 j}$ then from Lemma 4.2.4

$$\begin{aligned}
\tau_1(\hat{B}^n) &= \frac{1}{2} \max_{i_1, i_2} \sum_{j=1}^{N_x} \left| \hat{B}^n_{i_1 j} - \hat{B}^n_{i_2 j} \right| \\
&= \frac{1}{2} \max_{i_1, i_2} \sum_{j=1}^{N_x} \left| \frac{B^n_{i_1 j}}{B^n_{i_1\cdot}\mathbf{1}} - \frac{B^n_{i_2 j}}{B^n_{i_2\cdot}\mathbf{1}} \right| \\
&\leq \frac{N_x}{2} (1 - \varepsilon^2)^{n-1} .
\end{aligned}$$

Also, by Corollary 4.2.2

$$\left\|\hat{\psi}_1 - \hat{\psi}_2\right\|_1 \leq \frac{2}{\varepsilon} \left\|\psi_1 - \psi_2\right\|_1 ,$$

therefore part (2) is proved. □

# Appendix C

# Proofs Regarding Chapter 5

## C.1 Proof of Lemma 5.4.1

*Proof.* Let $\beta$ be any other vector satisfying $\beta \geq \mathbf{0}$ and $W\beta \leq Z$. Define $\mathbf{d} := \beta - \beta^*$, then

$$\|r_\beta\|^2 = \|r_{\beta^*}\|^2 + \|R\mathbf{d}\|^2 + 2r_{\beta^*}^T(R\mathbf{d})\,.$$

Since

$$(r_{\beta^*}^T R + v^T W)\beta \geq 0\,,$$

$$(r_{\beta^*}^T R + v^T W)\beta^* = 0\,,$$

$$v^T W\beta^* = v^T Z\,.$$

Therefore

$$r_{\beta^*}^T R\mathbf{d} + v^T(W\beta - Z) \geq 0\,.$$

Furthermore, because $v \geq \mathbf{0}$, we have

$$r_{\beta^*}^T R\mathbf{d} \geq v^T(Z - W\beta) \geq 0$$

and it follows that

$$\|r_\beta\|^2 \geq \|r_{\beta^*}\|^2\,.$$

□

**Remark C.1.1.** If $\beta^*$ is a regular point, we can apply the *Karush-Kuhn-Tucker* conditions to the convex programming. That is, write $f(\beta) = \|R\beta - S\|^2$. A feasible $\beta^*$ is a minimizer if and only if there exists an $u^T := [u_1^T \quad u_2^T] \geq \mathbf{0}^T$ such that

$$\nabla f(\beta^*) + \tilde{W}^T u = \mathbf{0}$$

and

$$u_l(\tilde{W}_l.\beta^* - \tilde{Z}) = 0\,, l = 1, 2\,.$$

where $\tilde{W}^T = [W^T \quad -I_m]$, $\tilde{Z}^T = [Z^T \quad \mathbf{0}^T]$, and $\nabla f$ is the gradient of $f$. So we have

$$2R^T r_{\beta^*} + W^T u_1 = I_m u_2 = u_2 \geq \mathbf{0}\,,$$

$$u_1^T(W\beta^* - Z) = 0\,,$$

and

$$u_2^T \beta^* = 0\,.$$

Take $v = u_1/2$ then the result follows.

## C.2 Proof of Theorem 5.6.2

*Proof.* Since $\lambda_m$ is the eigenvalue of $A$ with geometric multiplicity $d_m$. Denote its corresponding eigenvectors with $V_{m,n}$, $n = 0, \ldots, d_m - 1$ where

$$AV_{m,0} = \lambda_m V_{m,0}\,, \quad AV_{m,n} = \lambda_m V_{m,n} + V_{m,n-1} \quad \forall n = 1, \ldots, d_m - 1\,.$$

It follows that for $k \in \mathbb{N}_0$

$$(A - \lambda_m I)^k V_{m,n} = \begin{cases} V_{m,n-k} & k \leq n \\ \mathbf{0} & k > n \end{cases} \tag{C.1}$$

where $\mathbf{0}$ is of the same size as $V_{m,n-k}$. It is not difficult to see that for $m = 0, \ldots, M$

$$q(A)V_{m,n} = \sum_{j=0}^{n} \frac{q^{(j)}(\lambda_m)}{j!} V_{m,n-j} \quad n = 0, \ldots, d_m - 1\,. \tag{C.2}$$

On the other hand, from the definitions of $R_{ij}$, $f_m(A)$, and equality (C.1) we have

$$\begin{aligned}
&\sum_{i=0}^{M} \sum_{j=0}^{d_i-1} R_{ij}(A-\lambda_i I)^j \frac{q^{(j)}(\lambda_i)}{j!} V_{m,n} \\
=&\sum_{j=0}^{d_m-1} R_{mj}(A-\lambda_m I)^j \frac{q^{(j)}(\lambda_m)}{j!} V_{m,n} \\
=&\sum_{j=0}^{n} R_{mj} \frac{q^{(j)}(\lambda_m)}{j!} V_{m,n-j} \\
=&\sum_{j=0}^{n} f_m(A) \sum_{l=0}^{d_m-1-j} c_{m,l}(A-\lambda_m I)^l \frac{q^{(j)}(\lambda_m)}{j!} V_{m,n-j} \\
=&\sum_{j=0}^{n} f_m(A) \sum_{l=0}^{n-j} c_{m,l} V_{m,n-j-l} \frac{q^{(j)}(\lambda_m)}{j!} \\
=&\sum_{j=0}^{n} \sum_{l=0}^{n-j} \sum_{s=0}^{n-j-l} c_{m,l}\alpha_{m,s} V_{m,n-j-l-s} \frac{q^{(j)}(\lambda_m)}{j!}
\end{aligned} \tag{C.3}$$

where the equality in (C.3) follows from the fact that

$$f_m(A) = \prod_{i=0, i\neq m}^{M} \frac{(A-\lambda_i I)^{d_i}}{(\lambda_m - \lambda_i)^{d_i}} = \sum_{j=0}^{M-1} \alpha_{m,j}(A-\lambda_m I)^j$$

and

$$\alpha_{m,j} = \frac{f_m^{(j)}(x)}{j!}|_{x=\lambda_m} = \sum_{0\leq k_s \leq d_s, \sum k_s = j} \prod_{s=0, s\neq m}^{M} \frac{\begin{pmatrix} d_s \\ k_s \end{pmatrix}}{(\lambda_m - \lambda_s)^{k_s}}\,. \tag{C.4}$$

Note that $\alpha_{m,0} = 1$ for $m = 0, 1, \ldots, M$, also

$$\begin{aligned}
&\sum_{l=0}^{n-j} \sum_{s=0}^{n-j-l} c_{m,l}\alpha_{m,s} V_{m,n-j-l-s} = \sum_{l=0}^{n-j} \sum_{s=0}^{l} \alpha_{m,s} c_{m,l-s} V_{m,n-j-l} \\
=& c_{m,0} V_{m,n-j} + \sum_{l=1}^{n-j} \sum_{s=0}^{l} \alpha_s c_{m,l-s} V_{m,n-j-l}\,.
\end{aligned}$$

If for $m = 0, \ldots, M$, we set $c_{m,0} = 1$, and

$$\sum_{s=0}^{l} \alpha_{m,s} c_{m,l-s} = 0, \quad l = 1, \ldots, n-j\,.$$

That is

$$c_{m,l} = \begin{cases} 1 & l = 0 \\ -\sum_{s=1}^{l} \alpha_s c_{m,l-s} & l = 1, \ldots, n-j \end{cases}, \tag{C.5}$$

then following (C.3) we have for $m = 0, \ldots, M$

$$\begin{aligned}
&\sum_{i=0}^{M} \sum_{j=0}^{d_i - 1} R_{ij} (A - \lambda_i I)^j \frac{q^{(j)}(\lambda_i)}{j!} V_{m,n} \\
=& \sum_{j=0}^{n} \frac{q^{(j)}(\lambda_m)}{j!} V_{m,n-j} \quad n = 0, \ldots, d_m - 1\,.
\end{aligned} \tag{C.6}$$

Thus by (C.2), (C.6), Lemma 5.6.1 as well as (C.4), (C.5) we conclude the proof. □

# Appendix D

# Set-Valued Function and Measurable Selectors

For the convenience of reference, we quote definitions and some properties about multifunction and measurable selectors based on [38, Appendix D] and [27, Appendix], where more detailed information not mentioned here is available.

Suppose $V$ and $W$ are both nonempty Borel spaces and denote $2^W$ the collection of all nonempty subsets of $W$ . A mapping $D$ which associates with each $v \in V$ a nonempty subset $D(v)$ of $W$ is called a *set-valued function* (or *correspondence*, or *multifunction*) from $V$ to $W$. Namely, A set-valued function $D$ from $V$ to $W$ is a map $D : V \to 2^W$. $D$ is said to be compact-valued (measurable-valued) if for each $v \in V$, $D(v)$ is compact (measurable) subset of $W$. A *selector* (or *selection*) of $D$ is a function $d : V \to W$ such that $d(v) \in D(v)$ for each $v \in V$. In the text, $D(x)$ is the set of admissible action(s) when $x \in \boldsymbol{X}$ represents a system state, and any $d \in D$ is seen as a Markov stationary policy.

Let $(V, \mathcal{V})$ be a (Borel) measurable space and $W$ be a Polish space. A set-valued map $D : V \to 2^W$ is said to be *measurable* if for every *closed* set $w \subset W$ the set $\{v \in V | D(v) \cap w \neq \emptyset\}$ belongs to $\mathcal{V}$. A set-valued map $D : V \to 2^W$ is said to be *lower measurable* if for every *open* set $w \subset W$ the set $\{v \in V | D(v) \cap w \neq \emptyset\}$ belongs to $\mathcal{V}$.

# Bibliography


[1] E. Altman. *Constrained Markov decision processes.* Chapman and Hall/CRC, 1999.

[2] A. Arapostathis, V. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled markov processes with average cost criterion: a survey. *SIAM Journal on Control and Optimization*, 31:282–344, 1993.

[3] A. Arapostathis, R. Kumar, and S. Tangirala. Controlled markov chains and safety criteria. *Proceedings of the 40th IEEE Conference on Decision and Control*, 2:1675–1680, 2001.

[4] A. Arapostathis, R. Kumar, and S. Tangirala. Safety control of completely observed markov chains. *Proceedings of 2000 International Workshop on Discrete Event Systems*, August, 2000.

[5] A. Arapostathis and S. I. Marcus. Analysis of an identification algorithm arising in the adaptive estimation of markov chains. *Mathematics of Control, Signal and System*, 3:1–29, 1990.

[6] K. J. Astrom. Optimal control of markov processes with incomplete state information ii: the convexity of the loss function. *Journal of Mathematical Analysis and Applications*, 26:403–406, 1969.

[7] R. Atar and O. Zeitouni. Exponential stability for nonlinear filtering. *Probability et Statistiques*, 33:697–725, 1997.

[8] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[9] R. Bellman. *Adaptive Control Processes: A Guidede Tour*. Princeton University Press, Princeton, NJ, 1961.

[10] D. P. Bertsekas. *Dynamic programming and stochastic control*. Academic Press, 1976.

[11] D. P. Bertsekas and S. E. Shreve. *Stochastic optimal control: the discrete time case*. Academic Press, 1978.

[12] D. Blackwell. Discrete dynamic programming. *Annals of Mathematical Statistics*, 33:719–726, 1962.

[13] K. C. Border. *Fixed point theorems with applications to economics and game theory*. Cambridge University Press, 1985.

[14] V. S. Borkar. N-person non-cooperative stochastic games with partial information. *Game theory and economic applications (New Delhi, 1990), Lecture Notes in Econom. and Math. Systems, 389, Springer, Berlin,*, pages 98–113, 1992.

[15] V. S. Borkar. Ergodic control of partially observed markov chains. *System and Control Letters*, 99:185–189, 1998.

[16] R. Cavazos-Cadena and O. Hernandez-Lerma. Recursive adaptive control of markov decision processes with the average reward criterion. *Applied Mathematics and Optimization*, 23:193–207, 1991.

[17] D.-M. Chuang. *Risk-sensitive control of discrete-time partially observed Markov decision processes*. Ph.D. Dissertation, The University of Texas at Austin, 1999.

[18] D.-M. Chuang. Risk-sensitive control of partially observed controlled markov chains. 1999.

[19] D.-M. Chuang and A. Arapostathis. Some new results on the ergodic control of partially observed markov chains. *Proceedings of the 38th IEEE conference on decision and control*, pages 1908–1909, 1999.

[20] C. Dermen. On sequential decisions and markov chains. *Management Science*, 9:16–24, 1962.

[21] T. E. Duncan, B. Pasik-Duncan, and L. Stettner. Adaptive control of a partially observed discrete time markov process. *Applied Mathematics and Optimization*, 37:269–293, 1998.

[22] E. B. Dynkin and A. A. Yushkevich. *Controlled Markov processes*. Springer-Verlag, 1995.

[23] K. Fan. Fixed-point and minimax theorems in locally convex topological linear spaces. *Proceedings of the National Academy of Sciences of the U.S.A.*, 38:121–126, 1952.

[24] A. Federgruen. On n-person stochastic games with denumerable state space. *Advances in Applied Probability*, 10:452–471, 1978.

[25] A. Federgruen, A. Hordijk, and H. C. Tijms. Recurrent conditions in denumberable state markov decision processes. *Dynamic programming and its application*, pages 3–22, 1978. P. L. Puterman ed. Academic Press, New York.

[26] A. Federgruen and H. C. Tijms. The optimality equation in average cost denumberable state semi-markov decision problems, recurrent conditions and algorithms. *Journal of Applied Probability*, 15:356–373, 1978.

[27] E. Fernández-Gaucherand. *Controlled Markov processes on the infinite planning horizon: optimal & adaptive control.* Ph.D. Disertation, The University of Texas at Austin, 1991.

[28] E. Fernández-Gaucherand, A. Arapostathis, and S. I. Marcus. On the adaptive control of partially observable markov decision processes. *Proceedings of the 27th IEEE Conference Decision and Control*, pages 1204–1210, 1988.

[29] E. Fernández-Gaucherand, A. Arapostathis, and S. I. Marcus. On partially observable markov decision processes with an average cost criterion. *Proceedings of the 28th IEEE Conference Decision and Control*, pages 1267–1272, 1989.

[30] E. Fernández-Gaucherand, A. Arapostathis, and S. I. Marcus. Convex stochastic control problems. *Proceedings of the 31th IEEE Conference Decision and Control*, pages 2179–2180, 1992.

[31] E. Fernández-Gaucherand, A. Arapostathis, and S. I. Marcus. A methodology for the adaptive control of markov chains under partial state information. *Proceedings of the 31th IEEE Conference Decision and Control*, pages 2750–2752, 1992.

[32] E. Fernández-Gaucherand, A. Arapostathis, and S. I. Marcus. Analysis of an adaptive control scheme for a partially observed controlled markov chain. *IEEE transactions on Automatic Control*, 38:987–993, 1993.

[33] J. Filar and K. Vrieze. *Competitive Markov decision processes*. Springer-Verlag, 1997.

[34] M. K. Ghosh and A. Bagchi. Stochastic games with average payoff criterion. *Applied Mathematics and Optimization*, 38:283–301, 1998.

[35] I. Glicksberg. A further generalization of the kakutani fixed point theorem with application to nash equilibrium points. *Proceedings of the American Mathematical Society*, 3:170–174, 1952.

[36] T. L. Graves and T. L. Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM Journal on Control and Optimization*, 35, 1997.

[37] J. Gunnarsson. *Symbolic methods and tools for discrete envent dynamic systems*. Ph.D thesis, Linkopin University, Linkping, Sweden, January, 1997.

[38] O. Hernández-Lerma. *Adaptive Markov control processes*. Springer Verlag, 1989.

[39] O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov control processes*. Springer Verlag, 1996.

[40] L. E. Holloway, B. H. Krogh, and A. Giua. A survey of petri net methods for controlled discrete event systems. *Journal of Discrete Event Dynamical Systems: Theory and Applications*, 7(2):151–190, 1997.

[41] A. Hordijk and L. C. M. Kallenberg. Constrained undiscounted stochastic dynamic programming. *Mathematics of Operations Research*, 9(2):276–289, 1984.

[42] R. Howard. *Dynamic programming and Markov decision processes*. MIT Press, Cambridge, MA, 1960.

[43] A. Neyman J. F. Mertens. Stochastic games. *International Journal of Game Thoery*, 10:53–66, 1981.

[44] A. F. Karr. *Probability*. Springer-Verlag, 1993.

[45] P. R. Kumar. Optimal adaptive control of linear quadratic gaussian systems. *SIAM Journal on Control and Optimization*, 21:163–178, 1983.

[46] P. R. Kumar. A survey of some results in stochastic adaptive control. *SIAM Journal on Control and Optimization*, 23:329–380, 1985.

[47] P. R. Kumar and T. H. Shiau. Existence of value and randomized strategies in zero-sum discrete-time stochastic dynamic games.

[48] P. R. Kumar and P. Varaiya. *Stochastic Systems: estimation, identification and adaptive control*. Prentice-Hall, 1986.

[49] R. Kumar, V. K. Garg, and S. I. Marcus. Predicates and predicate transformers for supervisory control of discrete event systems. *IEEE Transactions on Automatic Control*, 38:232–247, 1993.

[50] K. Kuratowski and C. Ryll-Nardzewski. A general theorem on selectors. *Bull. Acad. Polon. Sci.*, 13:379–403, 1965.

[51] H. J. Kushner and C. G. Yin. *Stochastic approximation algorithm and applications*. Springer-Verlag, New York, 1997.

[52] A. Manne. Linear programming of sequential decisions. *Management Science*, 6:259–267, 1960.

[53] S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag, 1993.

[54] A. S. Nowak. Stationary equilibria for nonzero-sum average payoff ergodic stochastic games with general state space. *Advances in dynamic games and applications (Geneva, 1992)*. Annals of International Society of Dynamic Games, 1, Birkhauser Boston, Boston, MA.

[55] A. S. Nowak. On zero-sum stochastic games with general state space. i. *Probability and Mathematical Statistics*, 4:13–32, 1984.

[56] A. S. Nowak. On zero-sum stochastic games with general state space. ii. *Probability and Mathematical statistics*, 4:143–152, 1984.

[57] A. S. Nowak. Zero-sum average payoff stochastic games with general state space. *Games and Economic Behavior*, 7:221–232, 1994.

[58] A. S. Nowak. Optimal strategies in a class of zero-sum ergodic stochastic games. *Mathematical methods of operations research*, 50:399–419, 1999.

[59] K. R. Parthasarathy. *Probability measures on Metric Spaces*. Academic Press, New York, 1967.

[60] N.-F. Peng. Spectral representations of the transition probability matrices for continuous time finite markov chains. *Journal of Applied Probability*, 33:28–33, 1996.

[61] L. K. Platzman. Optimal infinite-horizon undiscounted control of finite probabilistic systems. *SIAM Journal on Control and Optimization*, 18:362–380, 1980.

[62] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

[63] P. J. Ramadge and W. M. Wonham. Supervisory control of a class of discrete event processes. *SIAM Journal of Control and Optimization*, 25(1):206–230, 1987.

[64] R. Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[65] S. M. Ross. Arbitrary state markovian decision processes. *Annals of Mathematical Statistics*, 39:2118–2122, 1968.

[66] H. L. Royden. *Real Analysis*. Maxwell Macmillan, 3 edition, 1989.

[67] W. Rudin. *Real and complex analysis*. McGraw-Hill, 3 edition, 1987.

[68] W. J. Runggaldier and L. Stettner. *Approximations of discrete time partially observed control problems*. Applied Mathematics Monographs 6, Pisa, Italy, 1994.

[69] S. Sankatha, B. Watson, and P. Srivastava. *Fixed point theory and best approximation: the KKM-map principle*. Kluwer Academic Publishers, 1997.

[70] E. Seneta. *Non-negative matrices and Markov chains*. Springer Veralg, 1981.

[71] L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the U.S.A.*, 39:1095–1100, 1953.

[72] H. M. Wagner. On the optimality of pure strategies. *Management Science*, 6:268–269, 1960.

[73] A. Wald. *Sequential Analysis*. John Wiley, New York, 1947.

[74] D. J. White. Dynamic programming of markov chains and the method of successive approximation. *Journal of Mathematical Analysis and Application*, 6:373–376, 1963.

# Vita

Shun-Pin Hsu was born in Kaohsiung, Taiwan, on June 14, 1971, the third son of Mr. Ping-Jiun Hsu and Shu-Huei Hsieh. He obtained his Bachelor of Science degree in Control Engineering in June 1993, and his Master of Science degree in Mathematical Statistics in June 1995, both from the National Chiao-Tung University, Hsin-Chu, Taiwan. In October 1995 he was commissioned a second lieutenant in the Taiwan Army and stationed in Nan-Tou, a county in central Taiwan. He was discharged from military service in June 1997, and in August he entered the master program in the Electrical and Computer Engineering Department at the University of Texas at Austin. He obtained his second Master of Science degree in May 1999 and began a doctoral program in August 1999 in the same department.

Permanent address: 50 Jing-Chuan St. Kaohsiung 804, Taiwan

This dissertation was typeset with LaTeX[†] by author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.