PH.D. THESIS

Optimal Risk Sensitive Control of Semi-Markov Decision Processes

by Jay P. Chawla
Advisor: Steven I. Marcus and Mark A. Shayman

Ph.D. 2000-8

ABSTRACT

Title of Dissertation: Optimal Risk Sensitive Control of

Semi-Markov Decision Processes

Jay P. Chawla, Doctor of Philosophy, 2000

Dissertation directed by: Professors Steven I. Marcus and Mark A. Shayman
Department of Electrical and Computer Engineering

In this thesis, we study risk sensitive cost minimization in semi-Markov decision processes. The main thrust of the thesis concerns the minimization of average risk sensitive costs over the infinite horizon. Existing theory is expanded in two directions: the semi-Markov case is considered, and non-irreducible chains are considered. In particular, the analysis of the non-irreducible case is a significant addition to the literature, since many real-world systems do not exhibit irreducibility under all stationary Markov policies. Extension of existing results to the semi-Markov case is significant because it requires the definition of a new dynamic programming equation and a technically challenging adaptation of the Perron-Frobenius eigenvalue from the discrete time case.

In order to determine an optimal policy, new concepts in the classification of Markov chains need to be introduced. This is because in the non-irreducible case, the average risk sensitive cost objective function permits extremely unlikely events to exert a controlling influence on costs. We define equivalence classes of states called ‘strongly communicating classes’ and formulate in terms of them a new characterization of the underlying structure of Markov Decision Problems and Markov chains.

In the risk sensitive case, the expected cost incurred prior to a stopping time with finite expected value can be infinite. For this reason, we introduce an assumption: reachability with finite cost. This is the fundamental assumption required to achieve the major results of this thesis.

We explore existence conditions for an optimal policy, optimality equations, and behavior for large and small risk sensitivity parameter. (Only non-negative risk parameters are discussed in this thesis – i.e. the risk averse and risk neutral cases, not the risk seeking case.) Ramifications for the risk neutral objective function are also analyzed. Furthermore, a simple solution technique we call ‘recursive computation’ to find an optimal policy that is applicable to small state spaces is described through examples.

The countable state space case is explored, and results that hold only for a finite state space are also presented. Other, related objective functions such as sample path cost are analyzed and discussed.

We also explore finite time horizon semi-Markov problems, and present a general technique for solving them. We define a new objective function, the minimization of which is called the ‘deadline problem’. This is a problem in which the probability of reaching the goal state in a set period of time is maximized. We transform the deadline problem objective function into an equivalent finite-horizon risk sensitive objective function.

Optimal Control of
Semi-Markov Decision Processes

by

Jay P. Chawla

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2000

Advisory Committee:

Professors Steven I. Marcus and Mark A. Shayman, Chairmen
Professor Michael C. Fu
Professor Armand M. Makowski
Professor Andre L. Tits

DEDICATION

This thesis is dedicated to my uncle Rajender Chawla, who died 3 days before I defended it. Uncle Raj was a leader throughout his life, and touched the lives of many people. He brought the joy and experience of Indian music to many Americans in Atlanta through his weekly radio broadcasts. He did valuable medical research, most notably in the area of liver function. He was a leader in his community and in our family. He died doing what he loved – he was on travel to present a paper at a conference. I regret that I cannot tell him personally about my accomplishment since his strong encouragement over many years was an important factor in my achieving it.

ACKNOWLEDGEMENTS

I have many people to thank for their encouragement, guidance, and assistance in the completion of this thesis, although I can mention only a few here. Foremost among them are my advisors, Professors Mark Shayman and Steve Marcus. They are both very kind and very smart, and possess a wisdom that has been revealed through their guidance. They are also both excellent educators, and I have had the pleasure of taking classes with each of them. Of course, the greatest act of teaching I have experienced from them is their patient and dedicated guidance over a period of years while I grew slowly in understanding and completed my thesis.

After I completed my Masters degree, I left the university. At the time, I was not sure whether I would come back for my Ph.D.. I approached Dr. Shayman and he agreed to take me on as a student. Later, he was instrumental in getting me a job at MITRE that was related to the research we were doing. I had a very valuable, productive, and enjoyable experience at MITRE, culminating when I left late last year to focus on the final phase of my Ph.D. research. Dr. Shayman is very intuitive and knowledgeable, and he has somehow succeeded at the very challenging job of keeping me focused long enough to write a thesis.

Dr. Marcus is a leader in the university and the academic community at large. He is very knowledgeable in the field in which my thesis is written. He gave me the papers that sparked the research in this thesis, and he provided me with numerous other references, each at the right time and covering the right material to push my research forward and give it context.

Both of my advisors have helped me to channel my often chaotic musings into a coherent body of research, pruning away the rough edges and helping me to understand and greatly improve any ideas I have had. They have shown me tremendous vistas of knowledge, only some of which are reflected in this final written product. I have had the privilege of receiving the greatest and deepest of guidance that a person can receive, and I am very grateful to both of my advisors.

I also would like to thank my fiancee, Terry. She has been like a rock during the often difficult process of completing my thesis, and our love for each other has been like a beacon that has lit the way for me on the journey that now takes me to the completion of my thesis. She has been very patient and tremendously supportive of me, and I am very grateful.

I owe my gratitude also to my parents. Of course the person that I am today is a result of my parents' love and hard work. Also, they have been supportive of me in my research. My father has always encouraged me in my desire to get a Ph.D., and he has given me a firm push on those occasions when I very much needed one. My mother wants what is best for me, and she has provided me with the encouragement and love I have needed along the way.

My fellow students have also been more than just colleagues. They have helped me appreciate and take joy in my work, and understand why I should get a Ph.D. in the first place. Out of all of my fellow students, I especially want to thank Eric Justh and Vahid Reza.

Thanks very much to all of you!

TABLE OF CONTENTS

List of Figures **vii**

Assumptions Used in the Thesis **viii**

1 Introduction and Overview **1**
1.1 Introduction 1
1.2 Motivation for study 2
1.3 Background in the literature 4
1.4 Contributions of the Thesis 6
1.5 Organization of the thesis 9

2 The Semi-Markov Formulation **12**
2.1 Notation for objective functions used in this thesis 18
2.2 Properties of time-invariant Markov chains with countable state space 19

3 Finite Horizon Problems **24**
3.1 The deadline problem 24
3.2 The deadline problem with incremental costs 28
3.3 The discrete time deadline problem without incremental costs . . . 29
3.4 The deadline problem for a general SMDP 31
3.5 Rate of accrual of costs in an SMDP 33
3.6 The cost minimization problem for a finite-horizon SMDP 34

4 Average Costs over the Infinite Horizon **37**
4.1 Average cost objective functions 37
4.2 A dynamic program for the risk sensitive semi-Markov average cost case 39
4.3 Foundational assumptions for existence of optimal policies 50

5 Perron-Frobenius Eigenvalue **53**
5.1 Defining the Perron-Frobenius Eigenvalue and the Round Trip Cost 53
5.2 Round trip cost when the state space is finite 61
5.3 Average cost on the infinite horizon 66

5.4 Performance for a large risk sensitivity parameter 74
5.5 The risk neutral case and its relation to the risk sensitive case . . . 79
5.6 A dynamic program for the case $C^{\theta\to\theta}(\lambda_C^\Pi(\gamma)) < 1$ 82

6 An Optimal Policy Under the Assumption of Reachability with Finite Expected Cost 88
6.1 Reachability with finite expected cost 88
6.2 Borkar's Convex Analytic Approach 99

7 Reachability and Probabilistic Reachability 101

8 The Not Strongly Communicating Case 108
8.1 On the classification of Markov chains 109
8.2 Optimal policies in the not strongly communicating case 111
8.3 Properties of optimal policies for the SMDP in Example 8.2.1 . . . 145
8.3.1 The risk sensitive case 145
8.3.2 The risk neutral case 147

9 Optimal Policies and Optimality Equations for the Finite State Space Case 149
9.1 Optimality equations and a policy optimal from any state 150
9.2 Behavior for a fixed Markov, deterministic, stationary policy (i.e., a reducible Markov chain) 159
9.3 Behavior for large or small risk sensitive parameter 163

10 Some Other Objective Functions 165
10.1 Sample path convergence 165
10.1.1 Ramifications of sample path convergence – optimality . . . 170
10.2 A cost criterion without γ 172

11 Closing Remarks and Suggestions 174
11.1 Summary of major results 174
11.2 On the infinite and its reduction to the finite 175
11.3 Future Research 177
11.4 Speculation on how to determine strongly communicating classes . . 178
11.5 Computational complexity 180

A Index of Notation 181

Bibliography 183

LIST OF FIGURES

2.1 Example of system in which optimal control is dependent on prior costs. 16

5.1 Type I semi-Markov chain – round trip cost is finite at $\lambda = 0$. . . . 57
5.2 Type I semi-Markov chain – round trip cost grows asymptotically. 58
5.3 Type II semi-Markov chain – round trip cost grows discontinuously. ($C^{\theta \to \theta}(\lambda_i) < \infty$) 59
5.4 Type II semi-Markov chain – ($\lambda_i = \lambda_C^{\Pi}(\gamma)$) 60
5.5 Type II semi-Markov chain – ($\lambda_i < \lambda_C^{\Pi}(\gamma)$) 61
5.6 Example of Type II Markov Chain. 62
5.7 A Markov Chain with norm-like costs. 62
5.8 A Markov Chain with norm-like costs and finite transitions into and out of each state. 63
5.9 A simple Markov Chain. 80

8.1 An MDP for which no single policy is optimal from every initial state. 144
8.2 An MDP illustrating why Theorem 8.2.1 does not work for a risk neutral objective function. 148

ASSUMPTIONS USED IN THIS THESIS

Assumption 2.0.1 (finite action space) Page 14

Assumption 2.0.2 (compact action space) Page 14

Assumption 3.1.1 (probabilistic reachability) Page 25

Assumption 3.4.1 (transition time nonzero with positive probability) Page 32

Assumption 4.2.1 (uniform upper bound on a measure of transition time distribution) Page 41

Assumption 4.2.2 (uniform upper bound on costs) Page 41

Assumption 4.2.3 (norm like costs) Page 45

Assumption 4.3.1 (uniform upper bound on a measure of transition time distribution) Page 51

Assumption 4.3.2 (norm like costs) Page 51

Assumption 4.3.3 (finite branching assumption) Page 52

Assumption 5.2.1 (nondegenerate cost structure) Page 64

Assumption 5.2.2 (transition times nonzero with positive probability) Page 64

Assumption 6.1.1 (reachability from any state to any other with finite expected cost) Page 88

Assumption 7.0.1 (reachability w.p.1 or w.p.0, no in between) Page 102

Assumption 7.0.2 (the system can be driven from any state to SR) Page 102

Assumption 8.2.1 (If a subset is reachable, it is reachable with finite expected cost) Page 112

Assumption 8.2.2 (every state can reach a finite set of equivalence classes) Page 114

Assumption 10.1.1 (one of 4 growth conditions) Page 169

Assumption 10.1.2 (one of 4 growth conditions) Page 169

Assumption 10.1.3 (one of 4 growth conditions) Page 169

Assumption 10.1.4 (one of 4 growth conditions) Page 169

Chapter 1

Introduction and Overview

1.1 Introduction

For the criterion of average or discounted risk neutral costs on the infinite horizon, policy or value iteration can be used to find optimal policies for semi-Markov decision processes (see, e.g., Ronald Howard's books [22] and [21]). However, when the time between transitions varies over a continuous time interval and is not exponentially distributed, and either the time horizon is finite or the cost function to be optimized is not a linear sum of costs, the standard framework for solving MDPs is no longer applicable. (There are trivial exceptions, such as when the time between transitions is restricted to the positive integers, a case covered by Howard and Matheson [23] for a risk sensitive objective function.) In this thesis, we extend treatment of semi-Markov decision processes to the risk sensitive cost criterion, both on the finite and the infinite horizon.

The main contribution of this thesis is in laying the theoretical groundwork for a study of optimal average cost policies on the infinite horizon when the standard irreducibility assumption is removed. We focus on the risk sensitive objective function because it has interesting and useful properties, including robustness under parameter uncertainty. There has been considerable research in the area of risk sensitive control recently (see Section 1.3), and in this thesis we aim to push progress forward in terms of the complexity and scope of problems to which the risk sensitive optimality criterion can be applied. Future researchers may build on these results to determine efficient and convergent solution techniques for these optimization problems.

1.2 Motivation for study

In this thesis, we focus on optimizing the objective function of average risk sensitive costs on the infinite horizon. Aside from the mathematical interest of the problem, there are many practical reasons to pursue this avenue of study. The most direct reason is that sometimes one needs to avoid costly realizations and is willing to sacrifice somewhat in terms of average (risk neutral) performance. In this regard, a risk sensitive criterion objective has an advantage over a minimax objective since it balances risk with average performance. (Note: In [13], another objective function is proposed that balances the same tradeoff: mixed risk neutral/minimax control.)

One of the most natural applications of risk sensitive control is in maximizing financial return. This is because financial returns are inherently multiplicative, rather than additive – if one earns 5% in a year, one's portfolio value is multiplied by 1.05. In [6], risk sensitive portfolio managment is studied. A finite state space and discrete time formulation is used to model a number of factors including macroeconomic conditions, and portfolio performance is optimized with respect to the risk sensitive average cost objective function on the infinite horizon. Even for a finite time frame, a controller for the average cost objective function will perform well since performance converges yielding nearly optimal behavior on a finite time horizon. Another common objective function used in portfolio analysis is mean-variance control in which the mean gain plus minus a factor times the variance of the gain is maximized. This is similar to risk sensitive control, in fact it constitutes the first two terms of the Taylor series expansion of the exponential. However, it leads to undesirable behavior including history-dependence of the optimal policy, as we will later illustrate in an example.

In general, aside from the advantage of minimizing risk (i.e., reducing the probability of a very costly realization), a risk sensitive controller outperforms a risk neutral controller when system parameters are not known with certainty or they are not constant (see e.g., [39], [15], [8], [3], [12], and [13]). This is due to the connection between risk sensitive and robust control first pointed out by Glover and Doyle in [17]. (For references to subsequent literature, see [15].)

In [8], a variational representation is used to connect risk-sensitive and robust control. It is shown that for a stochastic differential game, strategies that are nearly maximizing for the robust problem can be used to define nearly minimizing controls for the risk-sensitive problem with small risk parameter. In [15], some robustness properties of a risk sensitive controller are stated and proved, including a stochastic small gain theorem.

In [3], a framework is created for solving robust control problems using a risk sensitive controller. Specifically, for the case in which there are uncertainties in system parameters, a risk sensitive control problem is formulated and solved using an information state. The optimal controller for this problem performs robustly. In [39], a risk sensitive criterion is used to perform decision theoretic diagnosis with application to communication network failures. The reason a risk sensitive criterion is used is because network parameters are changing and can only be estimated. The robustness of a risk sensitive controller allows it to perform well under those conditions.

Note: For more information about risk sensitive control of partially observed MDPs (POMDPs) including large and small risk limit results, see [16]. See the seminal paper [41] for the risk neutral POMDP case. We assume full state observations throughout this thesis.

1.3 Background in the literature

There is a rich literature in risk sensitive control stretching back over half a century. In this background section we focus on results in the literature that are of direct relevance to our work.

In [35], Puterman covers a wealth of issues involving MDPs, SMDPs, dynamic programming, existence of optimal policies, policy iteration, value iteration, and linear programming. The average cost case is covered in depth, including the semi-Markov case. However, Puterman only covers the risk neutral case. Risk sensitive objective functions are not discussed. However, for the risk neutral case, the results in the literature are well explained. The Average Cost Optimality Inequality (ACOI) is described, based on Sennott's work as described in [37]. Another excellent overview text is [19], which covers much of the same ground as Puterman, including a detailed discussion of the linear programming approach to solving the risk neutral average cost control problem. Bertsekas has written two canonical volumes, [4] and [5], on all aspects of risk neutral optimal control. His texts are an excellent source for a first time reading of the material because they are very intuitively written. In addition, they are broad in scope and cover all of the relevant material.

In [37], Sennott explores optimal policies for Markov decision problems, also for the risk neutral case only. She shows that in the countable state space case for bounded costs, that the ACOI holds. ([37], P. 135)

The risk sensitive objective function was first addressed by Ronald Howard, ([21], [22], and (with Matheson) [23]) who covered the discrete time, finite horizon case. The discounted costs case on the infinite horizon was covered by Chung and Sobel in [10]. Unfortunately, in the discounted costs case the optimal policy is nonstationary in general, although as time gets large it converges to a stationary policy. Recent work in the area has been done by Coraluppi, [14] who discussed tradeoffs between various objective functions and further explored the discounted costs case. Patek [34] recently considered the risk sensitive stochastic shortest path problem for a finite state space in discrete time. He showed the existence of an optimal stationary policy and proved the convergence of value and policy iteration.

The risk sensitive objective function is 'harder' to analyze than the risk neutral objective function in the infinite horizon case. The reason for this is because the risk neutral objective function can take advantage of ergodicity in a direct way: if a stationary Markov policy induces an ergodic distribution, then the average risk neutral cost on the infinite horizon is given by the cost function weighted by the ergodic distribution. In addition, in the limit as the discount factor approaches 1, the discounted risk neutral cost approaches the average risk neutral cost. This 'vanishing discount' approach is described in, e.g. [4] and [35].

In [18], Hernandez and Marcus extend the risk sensitive results by applying a method similar to the vanishing discount approach to the Isaacs equation of an ergodic cost stochastic dynamic game. (Note: Fleming and Hernandez used the Isaacs equation in this way earlier for the finite state case.) In [9], Cavazos-Cadena and Fernandez-Gaucherand extend the risk sensitive results in the same way but without resorting to a limiting argument. In both [18] and [9] the result is that if costs are bounded over the entire (countable) state space and (simultaneous Doeblin condition) every policy returns the system to a specific recurrent state within an expected time that is uniformly bounded starting from any state, then an optimal policy exists and an optimality equality holds for that policy. In [9] it is additionally pointed out that unique problems arise in the risk sensitive case. In particular, the expected cost to escape a state need not even be finite! A simple example is used to show that because of this potentially infinite transition cost unless the risk sensitivity parameter is sufficiently small, the average cost for a given stationary policy is not necessarily the same starting from every initial state despite the strong recurrence (Doeblin) condition.

In [31], Di Masi and Stettner extend the results in [18] by retaining the bounded costs assumption and replacing the Doeblin condition with a very strong assumption on the transition probabilities. (Essentially that the difference in transition probabilities from any two states is uniformly bounded.)

$$P[C|x_1, a_1] - P[C|x_2, a_2] \leq \delta.$$

In addition to the results in [18], Di Masi and Stettner also show that the limit as the risk sensitivity parameter goes to zero from above of the risk sensitive cost is equal to the risk neutral cost.

In [14], Coraluppi points out that the discounted risk sensitive cost, as the risk sensitivity parameter goes to ∞, approaches the discounted maximum cost. This was already known to be true in the finite horizon case.

Balaji, Borkar, and Meyn have made significant contributions to the area recently. In [2], Balaji and Meyn studied ergodicity for an irreducible Markov chain with risk sensitive costs. This extends earlier ergodicity work (see [32]) in the risk neutral costs area. The most important result in [2] shows that if there is a Lyapunov function that satisfies a growth condition, then the average risk sensitive cost over the infinite horizon exists and is independent of the initial state. In [7], Borkar and Meyn use the results in [2] to prove the existence of an optimal policy. Their result is quite broad and assumes only three things: the costs are norm-like, the (countable) state space is irreducible under all Markov stationary policies, and there exists a policy that induces a finite average risk sensitive cost.

In [20] and [26], results are presented that show the existence of a sample path optimal (risk neutral average costs) policy. The conditions are different in the two references, and will be discussed in section 10.1.

1.4 Contributions of the Thesis

In this thesis, we cover the risk sensitive case in depth. Our results extend the results in [7] and [2] by covering the semi-Markov case and removing the irreducibility assumption. In particular, the removal of the standard assumptions that all policies are unichain and that the entire state space is irreducible under any policy is a major contribution to the literature. Although the risk neutral objective function has been studied without the irreducibility assumption, we are not aware of any literature that studies the risk sensitive objective function without a strong irreducibility assumption. In particular, as Cavazos-Cadena and Fernandez-Gaucherand point out in ([9], P. 4): "it is well known that a 'communicating' condition is necessary in order to have the optimal average cost be independent of the initial state, and that a strong recurrence condition is required for the existence of a bounded solution to the average cost optimality equation." We remove the 'communicating' condition in this thesis and we also remove the strong recurrence condition, allowing the existence of a sequence of policies that approach null recurrence, and allowing policies to be null recurrent or not recurrent at all. We are not aware of other work in which these strong assumptions have been removed in studying the risk sensitive average costs objective function on the infinite horizon.

As an aside, we want to point out that the classification and relevance of the 'communicating' aspects of MDPs are different when applied to the study of risk sensitive versus risk neutral objective functions. We explore that difference in this thesis.

In this thesis, we prove two verification theorems: one for the case of bounded costs and one for the case of norm like costs. These results are an extension of the verification theorem result in [18] because they involve semi-Markov decision problems. Furthermore, we provide a verification theorem in which the bounded costs assumption of [18] is replaced by a norm-like costs assumption. This verification theorem is used in the same way the verification theorem of [18] is used, to complete a proof of the existence of an optimal policy.

We find conditions under which there exists an optimal policy, both for the strongly communicating case and the not strongly communicating case, based on two fundamental assumptions: that if a set of states is reachable w.p.1, then it is reachable with finite costs, and that costs are norm-like (I.e., for any given bound, there are only a finite number of states with transition costs below the bound.) We believe that these are natural assumptions to make. The irreducibility assumption is unnatural because not every policy will hit every state infinitely many times w.p.1. The norm-like assumption is natural because the farther the system gets from its 'core' set of states, the more costly it should be. The assumption that a reachable set of states can be reached with finite costs is natural both because its converse is unnatural and because of the fact, pointed out by Cavazos-Cadena and Fernandez-Gaucherand in [9] that without that assumption the long-term average cost depends on the initial state. (In our non-irreducible framework, the corresponding ramification is that the long-term average cost *within an irreducible subclass induced by a policy* depends on the initial state.)

In order to prove our general results, we had to examine the behavior of semi-Markov, as opposed to discrete time, processes. The main work in this area is done in the proof of the verification theorems. We also had to classify communicating properties of controlled Markov chains in ways previously not relevant to optimization problems. This work culminates in the strong optimality resuls near the end of the thesis.

We also cover, as do Di Masi and Stettner (in [31]) the behavior of the risk sensitive cost as the risk sensitivity parameter goes to zero. However, our results are much broader, not requiring bounded costs. Furthermore, we eliminate the irreducibility assumption and describe the limiting behavior, something that has not been done before for the risk sensitive average cost objective function. (Although it has been done for the risk neutral case. See e.g., [35].) We also cover the case where the risk sensitivity parameter goes to ∞. In that case (in discrete time only – we do not cover the semi-Markov case), the average risk sensitive cost approaches the average maximum cost, when it is defined.

We present broad conditions under which a policy's sample path (risk neutral average) costs equal its expected costs w.p.1. This result can be used to extend the results in [20] and [26].

1.5 Organization of the thesis

In Chapter 2, we present the mathematical definition of the optimization problem this thesis addresses: the semi-Markov decision problem formulation. The properties of the state space, action space, and transition time and cost structure are defined. In addition, some unique features of risk sensitive and risk neutral objective functions are discussed. Also, some basic notation used throughout the thesis and some basic properties of time-invariant Markov chains are introduced.

In Chapter 3, we describe the deadline problem. This is a new problem that apparently has not been described in the literature. The deadline problem is to find the policy that will take the system to the *goal* state within a predefined time limit with the highest probability. In other words, the speed with which the system reaches the goal state or its 'closeness' if it does not hit the goal state are irrelevant. All that matters is reaching the goal state within the time limit. The deadline problem, it is shown, can be reduced to an equivalent risk sensitive control problem and solved using standard methods.

In Section 3.5, the rate of accrual of costs in an SMDP is further defined. Then in Section 3.6 a simple algorithm is defined to solve any finite horizon SMDP. It is an extension of the familiar dynamic programming technique to solve a finite horizon MDP. Chapter 3, in addition to defining and solving the deadline problem, addresses the problem of finite horizon SMDPs in general.

In Chapter 4, the objective function this thesis addresses is defined: the average cost risk sensitive objective function on the infinite horizon. The continuous time dynamic program (4.2) is introduced as well. Then the two verification theorems are presented and proved. The first verification theorem, Theorem 4.2.1, covers the case of bounded costs over the countable state space. It shows that if the dynamic program (4.2) has a solution, then there exists an optimal policy defined by the dynamic program, and furthermore this optimal policy is stationary, Markov, and deterministic. The second verification theorem, Theorem 4.2.2, covers the case of norm-like costs. Chapter 4 concludes with a list of assumptions that will be used in future chapters to prove the main results of this thesis.

In Chapter 5, we define the Perron-Frobenius eigenvalue $\lambda_C^\Pi(\gamma)$, which is shown to be equal to the long term average cost of a stationary, Markov policy within one of its strongly communicating classes. Furthermore, the round trip cost $C^{\theta\to\theta}(\lambda)$ is defined. These core concepts are used to explore behavior for large and small values of the risk sensitivity parameter and to discover a recursive equality (5.13) that must hold within a recurrence class induced by a policy with finite Perron-Frobenius eigenvalue.

In Chapter 6, the fundamental Assumption 6.1.1 is stated. This assumption is that the system can be driven from any state to any other state w.p.1, and furthermore it can be driven with finite expected risk sensitive cost. This assumption places the risk sensitive average costs control problem on a par with the risk neutral average costs control problem because it eliminates the problem of infinite costs to get between states. (Recall that this problem was pointed out by Cavazos-Cadena and Fernandez-Gaucherand in [9].) The second theorem in this chapter shows that even if the dynamic program (4.2) fails to hold because the round trip cost at the Perron-Frobenius eigenvalue is less than one, then an optimality inequality (6.10) still holds.

In Chapter 7, reachability, probabilistic reachability, and equivalence classes of states that can reach each other (strongly communicating classes) are defined. Several lemmas used in later chapters are proved based on these definitions. In Chapter 7 it is demonstrated that the state space is composed of several strongly communicating classes of states that are self-reachable plus a set of transient states. Furthermore, some of these strongly communicating classes can reach others, creating a relationship that defines a partial ordering on the strongly communicating classes themselves.

In Chapter 8, we delve into the heart of this thesis. First an example is used to illustrate the maximum cost nature of the risk sensitive average cost objective function. (Therefore, the optimal control will execute minimaxing over the set of reachable strongly communicating classes.) Then, Section 8.1 discusses how our classification of Markov chains differs from that used by Puterman in [35] and why: he applied the classification to the solution of risk neutral average costs and we apply it to the solution of risk sensitive average costs. Then in Section 8.2, a series of lemmas culminates in Theorem 8.2.1, which shows that starting from any initial state there is an optimal policy in the not strongly communicating case. Section 8.3 is devoted to showing why Theorem 8.2.1 does not hold independent of the initial state, and further showing that if the optimal policy is independent of the initial state in the risk sensitive case, it also is in the risk neutral case.

In Chapter 9, the finite state space assumption is utilized. Theorem 9.1.1, a powerful result, starts the chapter. Theorem 9.1.1 generalizes Theorem 8.2.1 by showing that there is a policy that is optimal starting from any state. Theorem 9.1.2, also a powerful result, shows the optimality equations for the not strongly communicating case. Lemma 9.1.1 is another verification 'theorem' like those in Chapter 4. This lemma holds in the more general not strongly communicating case, though.

In the last two sections, Sections 9.2 and 9.3, we delve into more detail as to why the optimality equations and the limit of risk sensitive costs as $\gamma \downarrow 0$ look the way they do. We classify all of the realizations starting from a given state under a policy and find the probability that the realization falls in each class.

In Chapter 10, two topics are discussed: sample path convergence and the elimination of γ from the risk sensitive average cost objective function. In particular, the discussion in Subsection 10.1.1 is valuable in understanding the thesis as a whole.

Finally, in Chapter 11, some closing remarks and suggestions for future research are made.

Chapter 2

The Semi-Markov Formulation

Let (S, A, P, Z) be a semi-Markov control model. Put simply, a semi-Markov model (also called a semi-Markov decision process or SMDP) consists of a state space, S, an action space, A, a set of transition probabilities, P, that specify the probability of transitioning to a given state from a given other state under a given action, and lastly, Z. Z is what makes the semi-Markov model different from a Markov model. In a discrete time Markov model, transition times and transition costs are fixed. In a semi-Markov model, both times and costs are random, and they are described by a joint probability distribution dependent on the state and the action.

The state space, S, may be either finite or countably infinite and endowed with the discrete topology, and the action or control space A is a Borel space. The state evolves in continuous time and is piecewise constant. Those times when the state changes are called *decision times*, and a control action must be selected at each decision time. For every x in S, $\alpha(x) \subset A$ is the set of admissible actions when the system is in state x. The set of admissible pairs is denoted $K = \{(x, a) : a \in \alpha(x), x \in S\}$. (Clearly, $K \subset S \times A$.) The state process is continuous from the right, and immediately after each state change, a new action must be selected from those admissible actions for the new state, which is completely observed. The state occupied and action taken at the k^{th} decision epoch are denoted x_k, a_k respectively.

The time elapsed between the k^{th} and $k+1^{th}$ decision epochs, i.e., the 'transition time', is denoted t_k.

The manner of choosing an action at the k^{th} decision epoch is a mapping or decision rule $d_k : x \rightarrow \alpha(x);\ x \in S$. (Note that $d_k : S \rightarrow A$.) A decision rule may depend on the history $h_k = (x_0, a_0, ..., x_{k-1}, a_{k-1}, x_k)$ of the process up to the k^{th} decision epoch or it may depend only on x_k. Such decision rules are called history-dependent or Markovian, respectively. A decision rule may be randomized, specifying a probability density q_{d_k} on the set of actions. I.e., the probability that action $a \in \alpha(x_k)$ is chosen at the k^{th} decision epoch is $q_{d_k}(h_k)(a) \geq 0$, with $\sum_{a\in\alpha(x_k)} q_{d_k}(h_k)(a) = 1$ for all possible values of the history. Following [35], we denote the set of all decision rules at decision epoch k by D_k. There are 4 classes of decision rules: history dependent and randomized (HR), history dependent and deterministic (HD), Markov and randomized (MR), and Markov and deterministic (MD). We denote the class of decision rule by a superscript.

A policy Π is a sequence of decision rules, $\Pi = (d_1, d_2, ...)$ Let Π^L denote the set of all policies of class L; $L \in \{\text{HR,HD,MR,MD}\}$. Thus, $\Pi^L = D_0^L \times D_1^L \times ....$ We call a policy stationary if $d_k = d\ \forall k$. Also, we can see that

$$\Pi^{HR} \supset \Pi^{HD} \cup \Pi^{MR}; \text{ and } \Pi^{MD} \subset \Pi^{HD} \cap \Pi^{MR}.$$

As soon as an action a is selected, the next state y is determined from the transition law P, which is a stochastic kernel on S given (x, a). Z is a stochastic kernel on $\Re^+ \times \Re^+$ given (x, a, y). Z determines the transition time (i.e., the time between decision epochs) $t(x, a, y)$ and the transition cost $c(x, a, y)$ given the state and action selected, and the state to which the system transitions. So transition time and cost are not independent in general. Furthermore, we require that transition times be positive and transition costs be non-negative.

Throughout the remainder of the thesis, we will assume there is no dependence of the cost and time of a transition on the state transitioned to; i.e., $t(x, a, y) = t(x, a)$ and $c(x, a, y) = c(x, a)$. This assumption is made without loss of generality, because it can be imposed by adding states to the state space, while maintaining the finiteness or countability of the state space.

When describing the performance of a policy starting from a given initial state, we will use the notation $E_x^{\Pi}[\cdot]$ to denote the expected value of a random variable under policy Π starting from state x; and the notation $P_x^{\Pi}[\cdot]$ to denote the probability of an event under policy Π starting from state x.

The following theorem (2.0.1) is taken from [35] (p. 536):

Theorem 2.0.1 *Let* $\Pi = (d_1, d_2, ...) \in \Pi^{HR}$. *Then, for each* $x_0 \in S$, *there exists a policy* $\Pi^{'} = (d_1^{'}, d_2^{'}, ...) \in \Pi^{MR}$ *satisfying*

$$P^{\Pi^{'}}[x_k = j, a_k = a, t_k = \tau | x_0] = P^{\Pi}[x_k = j, a_k = a, t_k = \tau | x_0]$$

for $k = 1, 2, 3, ...$.

Assumption 2.0.1 (finite action space) $\alpha(x)$ *is finite* $\forall x \in S$.

Assumption 2.0.2 (compact action space) $\alpha(x)$ *is compact* $\forall x \in S$; *and* $P(y|x,a)$, $Z(t,c|x,a)$ *are continuous in* a.

Theorem 2.0.2 *Let* $L : (y, c, t) \rightarrow \Re$ *be a measurable function and assume either Assumption 2.0.2 or Assumption 2.0.1. Then,*

$$\inf_{d_k \in D^{MD}} E_{x_k}^{d_k}[L(x_{k+1}, c_k(x_k, a_k), t_k(x_k, a_k))] = \inf_{d_k \in D^{MR}} E_{x_k}^{d_k}[L(x_{k+1}, c_k(x_k, a_k), t_k(x_k, a_k))]. \tag{2.1}$$

Furthermore, the infimum is achieved.

Proof:

Since $D^{MD} \subset D^{MR}$, left hand side $\geq$ right hand side.

Now choose $d_k \in D^{MR}$. Suppose that $x_k = s$. Under d_k, there is a probability density function on a_k given by $q_{d_k}(\cdot)$. Denote $\bar{L}_{x_k}(a) = E[L(x_{k+1}, c_k(x_k, a_k), t_k(x_k, a_k))|a_k =$

$a]$. Then, under either Assumption 2.0.2 or Assumption 2.0.1, $\exists a^{'} \in \alpha(x_k)$ such that $\bar{L}_x(a^{'}) = \inf_{a\in\alpha(x)} \bar{L}(a)$. Therefore,

$$E^{d_k}_{x_k}[L(x_{k+1}, c_k(x_k, a_k), t_k(x_k, a_k))] = \int_{a\in\alpha(x_k)} \bar{L}_{x_k}(a) \cdot q_{d_k}(a) da$$

$$\geq \bar{L}_{x_k}(a^{'}) = E^{d^{'}_k}_{x_k}[L(x_{k+1}, c_k(x_k, a_k), t_k(x_k, a_k))].$$

where $d^{'}_k \in D^{MD}$ is the decision rule of taking action $a^{'}$ in state s. □

From Theorems 2.0.1 and 2.0.2, it can be seen that for the purpose of optimizing a risk neutral cost criterion we can confine our investigation to Markov, deterministic policies. However, while Theorem 2.0.2 is still applicable in the risk sensitive case (i.e., Theorem 2.0.2 applies to both multiplicative and additive dynamic programs), Theorem 2.0.1 is no longer relevant since the nature of the objective function (product of costs) brings dependency on the joint distribution of the state, rather than just its distribution at a given decision epoch. Later, we will see that under certain assumptions the optimal policy for infinite horizon risk sensitive average cost problems is a stationary, Markov, deterministic policy.

Clearly, if the time horizon is finite, the horizon effect will bring about a time dependence in the optimal policy. An interesting question to ask for both finite horizon and infinite horizon problems is "when does an optimal policy have to depend on accrued costs?"

In [23] it is pointed out that there exists an optimal control that is independent of past costs for (total, average, or discounted) risk neutral and risk sensitive objective functions (both finite and infinite horizon) in discrete time, i.e., if $t(x, a) \equiv 1 \; \forall x, a$ and $c(x, a)$ is deterministic. This makes intuitive sense, since if the objective function is risk neutral, the objective is to minimize the expected value of the sum of future costs regardless of past costs, which are merely additive. And if the objective function is risk sensitive, future costs are a multiplier to accrued costs; this multiplier should be minimized regardless of what costs have already been accrued. For all of the objective functions we study in this thesis, the use of past costs to determine actions will not improve performance.

However, there are objective functions in which past costs do affect future actions, i.e., in which an optimal policy must be dependent on past costs. One such objective function is the square of the total cost, as the following example demonstrates:

Example 2.0.1 (Cost dependence of optimal control)

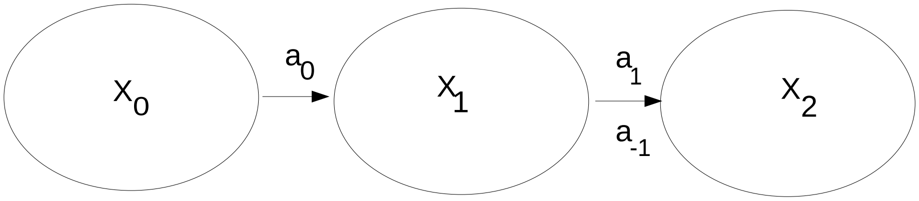

Figure 2.1: Example of system in which optimal control is dependent on prior costs.

Suppose that, as shown in figure 2.1, we have a discrete-time, finite horizon problem in which transitions are deterministic and independent of the action taken, there are two time steps, and the system starts in state x_0. Suppose that at time 0, only one control is admissible: a_0, and at time 1, two controls are admissible: a_1 or a_{-1}. Suppose the costs are as follows:

$$c(x_0, a_0) = \begin{cases} 1 \text{ with probability } \frac{1}{2} \\ 100 \text{ with probability } \frac{1}{2} \end{cases}$$

$$c(x_1, a_1) = 7$$

$$c(x_1, a_{-1}) = \begin{cases} 1 \text{ with probability } \frac{1}{2} \\ 11 \text{ with probability } \frac{1}{2} \end{cases}$$

Suppose further that the objective function is $J = E[(c_0 + c_1)^2]$, i.e., the objective function is the square of the total cost. If the cost incurred at time 0 is observed, then our choice of control at time 1 will depend on the observed cost. It can be seen that if cost 100 is observed, then we choose control a_{-1}, whereas if cost 1 is observed, we choose control a_1. Therefore, the optimal policy depends on the prior costs.

Now let's look at an objective function that is useful in financial applications: a weighted sum of the mean + the variance. (As pointed out in the introduction, this kind of objective function is used, e.g., in financial applications, although with the objective of maximizing benefits rather than minimizing costs. However, this example could be suitably modified to address profit maximization.)

Suppose we are trying to minimize $E[(c_0 + c_1) + \gamma(c_0 + c_1)^2]$. Then clearly if γ is large enough we would again choose a_{-1} when cost 100 is observed and a_1 when cost 1 is observed. For γ small enough, we would choose a_{-1} no matter what cost is observed at time 0.

This example brings to mind an interesting point. Observation of accrued costs in the problems we study is irrelevant to optimizing performance. However, there are problems and objective functions (such as mean-variance as shown above) in which cost observation is essential to maximizing performance.

In the following, we will restrict our attention to completely observed risk sensitive and risk neutral objective functions. We also introduce the completely observed *deadline problem*, in which the objective is to reach the *goal* state within a time deadline, which also has the property that past costs (actually the past probabilities of not reaching the goal state – the deadline problem does not deal with 'costs' per se) do not affect the optimal policy. From the fact that the deadline problem has this same 'nice' property, we might guess that it can be transformed into either a risk neutral or risk sensitive problem. We will show that it is in fact equivalent to a risk sensitive problem in which the 'costs' are a function of past probabilities of not reaching the goal state.

It is also worth mentioning that the minimax optimal controller obtained by using as the objective function the maximum possible (additive) cost under a policy, also has an optimal policy independent of past costs. Again, a relationship to the risk sensitive problem might be inferred from this fact, and in fact the minimax objective function is the limit of the risk sensitive objective function as the risk sensitivity parameter (defined later) tends to ∞. Also, each of the objective functions mentioned admits a dynamic programming formulation, which if taken for the infinite horizon average cost problem depends only on the current state.

2.1 Notation for objective functions used in this thesis

We assume without loss of generality that there is a cost to be minimized, rather than a reward to be maximized. Furthermore, the capital letter J is used to represent the objective function to be minimized. A superscript of Π indicates that policy Π is used to select actions. A subscript of x_0 indicates that the system begins at time zero just having transitioned to state x_0, i.e., at a decision epoch in state x_0. (These two rules of notation hold for expected values as well as objective functions.) Because we will always be assuming the system starts at a decision epoch, we will always have a state in the subscript of any objective function or expectation operator. An objective function with risk-sensitive costs is denoted J, and one with risk neutral costs is denoted $\mathcal{J}$. Continuous time is assumed to be used, but a superscript of ', as in $J^{'}$ denotes discrete time. An objective function is by default average cost, but a bar above the J (as in $\bar{J}$ or $\bar{\mathcal{J}}$) denotes a finite horizon objective function. An infinite horizon average cost objective function that takes the ratio of cost to time in the same number of transitions is denoted with a tilde above the J (as in $\tilde{J}$ or $\tilde{\mathcal{J}}$). A value function V denotes the infimum over all possible policies of the value of the objective function to be minimized. Because all decisions are made at decision epochs, a value function denotes this infimum taken at a decision epoch. For the deadline problem, we will denote the value function $V(R, s)$ where R is the time remaining until the deadline and s is the state to which the system has just transitioned.

To summarize the objective function notation,

$$J_{x_0}^{\Pi} = \text{risk sensitive objective function}$$

$$\mathcal{J}_{x_0}^{\Pi} = \text{ risk neutral objective function}$$

$$J_{x_0}^{'\Pi} \text{ or } \mathcal{J}_{x_0}^{'\Pi} = \text{ discrete time}$$

$$\bar{J}_{x_0}^{\Pi} \text{ or } \bar{\mathcal{J}}_{x_0}^{\Pi} = \text{finite time horizon}$$

$$\tilde{J}_{x_0}^{\Pi} \text{ or } \tilde{\mathcal{J}}_{x_0}^{\Pi} = \text{infinite horizon ratio of total cost to total time}$$

$$J_d(R, s) = \text{ objective function for the deadline problem}$$

$$\text{at decision epoch in state } s \text{ with time } R \text{ remaining}$$

where the superscript of Π means that policy Π is used and the subscript of x_0 means that the initial state is x_0. (For a continuous time objective function, the system is assumed to begin at a decision epoch.)

The deadline problem will be defined in Chapter 3.

2.2 Properties of time-invariant Markov chains with countable state space

The notation τ_G for $G \subset S$ and τ_y for $y \in S$ will be used throughout this thesis. Define $\tau_G \doteq \min(k \in \{1, 2, 3, ...\} | x_k \in G)$ and define $\tau_y \doteq \tau_{\{y\}}$. Recall that the initial state is denoted x_0; we denote $\sigma_G \doteq \min(k \in \{0, 1, 2, 3, ...\} | x_k \in G)$. (Note: it can easily be shown that τ_G is a stopping time.)

If a stationary policy $\Pi \in \Pi^{MR}$ is applied, then the embedded chain becomes a time-invariant Markov chain. Following Chapter 4 of [32], for $x, y \in S$ we define the relationship $x \overset{\Pi}{\to} y$ to be true if $P_x^{\Pi}(\tau_y < \infty) > 0$ and we define $x \overset{\Pi}{\leftrightarrow} y$ to be true if $x \overset{\Pi}{\to} y$ and $y \overset{\Pi}{\to} x$. If $x \overset{\Pi}{\leftrightarrow} x$, then the state x is called *probabilistically self-reachable* under stationary, Markov randomized policy Π. We denote the set of all probabilistically self-reachable states under Π as $PSR^{\Pi} = \{x \in S | x \overset{\Pi}{\leftrightarrow} x\}$.

Property 2.2.1 $\overset{\Pi}{\leftrightarrow}$ *is an equivalence relation on* PSR^{Π}.

If $x \in PSR^{\Pi}$, define the communicating class containing x as $\Omega_{\Pi}(x) = \{y \in PSR^{\Pi} | x \overset{\Pi}{\leftrightarrow} y\} = \{y \in S | x \overset{\Pi}{\leftrightarrow} y\}$. ($\Omega_{\Pi}(x)$ is the equivalence class containing x induced by $\overset{\Pi}{\leftrightarrow}$.) If $z \in \Omega_{\Pi}(x)$, $w \in \Omega_{\Pi}(y)$, and $z \overset{\Pi}{\to} w$, then we denote $\Omega_{\Pi}(x) \overset{\Pi}{\to} \Omega_{\Pi}(y)$. For $x \in PSR^{\Pi}$, if $x \overset{\Pi}{\to} y$ implies that $y \in \Omega_{\Pi}(x)$, then we say that $\Omega_{\Pi}(x)$ is *absorbing.* Denote the set of all states that are contained in any absorbing communicating class as Ω_{Π}^{a}. Ω_{Π}^{a} is absorbing, but not necessarily communicating.

The following Lemma is a formal restatement of an argument contained in ([32], P. 84):

Lemma 2.2.1 *For any stationary, Markov policy* Π, *any state* $x \in S$, *and any finite set* $G \subset S$,

$$\limsup_{k \to \infty} P_x^{\Pi}[x_k \in G^c \cup \Omega_{\Pi}^{a}] = 1.$$

In words, Lemma 2.2.1 says that no matter what the initial state, the system will eventually either go to infinity or enter an absorbing communicating class. (Or possibly both.)

Define $M_{\Pi} = \{x \in S | P_x^{\Pi}[\tau_x < \infty] = 1\}$. Clearly, $M_{\Pi} \subset \Omega_{\Pi}^{a}$, and M_{Π} is itself absorbing.

Lemma 2.2.2 *For any stationary, Markov policy* Π*, any state* $x \in S$*, and any finite set* $G \subset S$*,*

$$\lim_{k\to\infty} P_x^{\Pi}[x_k \in G^c \cup M_\Pi] = 1.$$

Proof:

By Lemma 2.2.1, all that we are required to show is that

$$\lim_{k\to\infty} P_x^{\Pi}[x_k \in G \cap (\Omega_\Pi^a - M_\Pi)] = 0.$$

Since G is a finite set, so is $G \cap (\Omega_\Pi^a - M_\Pi)$. For each $x \in G \cap (\Omega_\Pi^a - M_\Pi)$, we have that $P_x^{\Pi}[\tau_x < \infty] < 1$. Therefore, we are guaranteed that the state will eventually leave $G \cap (\Omega_\Pi^a - M_\Pi)$ and never return.

□

Define $\eta_G^N = \sum_{k=0}^{N}[I(x_k \in G)]$ and let $\eta_G = \lim_{N\to\infty} \eta_G^N$. The following Lemma will be useful in establishing a verification theorem for a risk-sensitive average cost-optimal policy when costs are unbounded.

Lemma 2.2.3 *Let* Π *be a Markov, stationary policy,* $G \subset S$*,* $|G| < \infty$*, and* $x \in S$*. If* $P_x^{\Pi}[\eta_G = \infty] = 1$*, then* $\exists B \subset M_\Pi \cap G$ *such that* $P_x^{\Pi}[\tau_B < \infty] = 1$*.*

Proof:

The result follows immediately from Lemma 2.2.2.

Lemma 2.2.3 says that if the system hits a finite set an infinite number of times w.p.1 ($E_x^{\Pi}[\eta_G] = \infty$ is NOT sufficient since it does not guarantee that $\eta_G = \infty$ w.p.1.) then it enters an absorbing communicating class that is positive recurrent under Π w.p.1. (*Positive recurrent* is defined below.)

Definition 2.2.1 *A Markov chain is called irreducible if* $\Omega_\Pi(x) = S \ \forall x \in S$*.*

For any state $x \in \Omega_\Pi^a$, define $d(x) = g.c.d.\{n \geq 1 | P_x^{\Pi}[x_n = x] > 0\}$.

Definition 2.2.2 ([32]) *An irreducible chain is called aperiodic if* $d(x) = 1\ \forall x \in S$.

The following theorem is taken from [11]:

Theorem 2.2.1 *If a time-invariant semi-Markov process (i.e., an SMDP for a fixed stationary policy* $\Pi \in \Pi^{MR}$*) is irreducible and the states are periodic with period* δ*, then the cumulative distribution* $F_x(\cdot)$ *of the transition time* $t(x,a)$ *is a step function with jumps in the set* $\{\delta_x, \delta_x + \delta, \delta_x + 2\delta, ...\}$*. for some* $\delta_x \geq 0$*. Moreover, if* $P(x_1|x_0, \Pi(x_0)), P(x_2|x_1, \Pi(x_1)), ..., P(x_n|x_{n-1}, \Pi(x_{n-1}) > 0$ *and* $x_0 = x_n$*, then* $\delta_{x_0} + \delta_{x_1} + ... + \delta_{x_{n-1}}$ *is equal to an integer multiple of* δ*.*

Because of the necessary condition in this theorem, the controlled semi-Markov process is periodic only under unusual circumstances (e.g., if transition times are all the same w.p.1, i.e., the discrete time case.) When considering the convergence of policy and value iteration (which is beyond the scope of this thesis), we are concerned with periodicity of the embedded Markov chain because many results require the embedded Markov chain to be aperiodic. Furthermore, the aperiodicity transformation ([32], P. 371), which is used to transform a periodic Markov chain to an aperiodic Markov chain, only works in the case of a risk neutral objective function. However, for our purposes it turns out that certain technical assumptions can be used instead of the assumption of aperiodicity in order for our existence and uniqueness results to hold.

Definition 2.2.3 ([32], P. 500) *A subset* $C \subset S$ *is called a positive recurrent subclass induced by stationary, Markov policy* Π *if* $\forall x \in C$, $\Omega_\Pi(x) = C$ *and* $\forall A \subset C$, $A \neq \emptyset$, $\lim_{n\to\infty} P_x^\Pi[x_n \in A] > 0$.

Definition 2.2.4 ([32], P. 500) *A subset* $C \subset S$ *is called a null recurrent subclass induced by stationary, Markov policy* Π *if* $\forall x, y \in C$, $\Omega_\Pi(x) = \Omega_\Pi(y) = C$ *and* $\lim_{n\to\infty} P_x^\Pi[x_n = y] = 0$.

Note: It can be easily shown that M_Π is the set of all absorbing communicating classes that are also positive recurrent.

If S is a positive (null) recurrent subclass induced by stationary, Markov policy Π, then the induced Markov chain is called positive (null) recurrent.

Note: The limit $\lim_{n\to\infty} P_x^\Pi[x_n \in C] > 0$ is guaranteed to exist ([32], P. 230) if $\Omega_\Pi(x) = C$, C is absorbing, and Π is stationary and Markov. Furthermore ([32], P. 500), an absorbing, communicating class C induced by Markov, stationary policy Π must be either positive recurrent or null recurrent.

The following definition of a positive recurrent subclass can be shown to be equivalent to Definition 2.2.3:

Definition 2.2.5 *Given a policy Π, a set $C \subset S$ is called a 'positive recurrent subclass' induced by Π if*

$$E_x^\Pi[\tau_y] < \infty; \forall x, y \in C.$$

and

$$P_x^\Pi[\tau_z < \infty] = 0; \forall x \in C, z \notin C.$$

Chapter 3

Finite Horizon Problems

3.1 The deadline problem

Before we consider the more general cases (risk neutral and risk sensitive objective functions), let us consider the case of a Semi-Markov Decision Problem (SMDP) with deterministic but nonuniform time between transitions in which the objective is to reach the goal state within a given time budget.

For convenience, we track the SMDP in discrete time, i.e., at transition times, with the state at 'time k' being the state after the k^{th} transition. The actual (continuous) time after the k^{th} transition is equal to the sum of the first k state occupancy (or transition) times.

The problem is as follows: There is a finite set of states $\{s_0, s_1, s_2, ..., s_n\}$ and a finite set of m_i control actions $\{a_1^i, a_2^i, ..., a_{m_i}^i\}$ possible in each state, s_i. (So the set of admissible pairs is $K = \{(s_i, a_k^i) | 0 \leq i \leq n; 1 \leq k \leq m_i\}$.) There is a unique *goal* state s_0. At each discrete time k, the controller selects an action $a(k)$, where $a(k)$ is selected from the set $\alpha(s(k))$, the set of actions that can be taken when the system is in state $s(k)$. The next state is selected according to the transition law P, i.e. the probability of transitioning to state y from state x under action a is given by $P(y|x, a)$. Denote by $r(x, a)$ the set $\{y | P(y|x, a) > 0\}$. The time elapsed between the k^{th} and $k+1^{th}$ transitions, denoted t_k, depends on the current state and the action selected, i.e., $t_k = g(s(k), a(k))$ where $g(\cdot, \cdot)$ is a deterministic function of its arguments. (So the transition kernel Z is degenerate with no costs and with deterministic transition times.) The objective is to reach the goal state s_0 while keeping the total time spent below a budgeted (or deadline) time, B.

Once the goal state is reached, the process terminates. We also make the assumption

Assumption 3.1.1 (A1) *For any pair of states* (s_i, s_j)*, there exists a policy that takes the system from* s_i *to* s_j *with nonzero probability in a finite number of transitions.*

Because the state space if finite, **(A1)** is equivalent to the assumption that you can drive the system from s_i to s_j w.p.1 given an infinite amount of time.

The problem starts at time 0 in an arbitrarily selected non-goal state. At transition time k, if time X has elapsed prior to time k, we say that the *cost budget remaining,* denoted $R(k)$, is $B - X$. Thus, $R(0) = B$. Let us define the value function, $V^{\Pi}(R, s) =$ the probability of not reaching the goal state within budget by following policy Π given that the system just transitioned to state s with time R remaining.

Define the optimal value function $V(R, s) = \inf_{\Pi \in \Pi^{HR}} V^{\Pi}(R, s)$. By Theorem A, we have that $V(R, s) = \inf_{\Pi \in \Pi^{MR}} V^{\Pi}(R, s)$. A policy Π^* is said to be *optimal* if $V^{\Pi^*}(R, s) \leq V_{\Pi}(R, s)$; $\forall \Pi, R, s$; i.e., if $V^{\Pi^*}(R, s) = V(R, s)$.

Let us now examine some properties of the optimal value function:

We see that

$$V(R, s_0) = \begin{cases} 0 \text{ if } R \geq 0 \\ 1 \text{ if } R < 0 \end{cases} \tag{3.1}$$

And for $s \neq s_0$,

$$V(R,s) = 1;\text{ if R } \leq 0 \tag{3.2}$$

Lemma 3.1.1

$$V(R,s) = \inf_{a \in \alpha(s)} \sum_{x \in r(s,a)} P(x|s,a)V(R - g(s,a), x);\ \textit{if } R > 0 \tag{3.3}$$

Proof:

$V(R,s) = \inf_{\Pi \in \Pi^{MR}} V^{\Pi}(R,s)$. Because we are considering only Markov policies, we can decompose a given Π into $\Pi = \Pi(R) \cup \{\Pi(r)|r < R\}$. And $\Pi(R) = d_R(s)$ for some randomized decision rule d_R. (I.e., a policy $\Pi \in \Pi^{MR}$ can be decomposed into its decision rule at each time. That decision rule is just a randomized mapping from state to action.) We have by definition of V,

$$V(R,s) = \inf_{d_R(s) \in D^{MR}, \{\Pi(r)|r<R\}} V^{d_R(s) \cup \{\Pi(r)|r<R\}}(R,s)$$

$$= \inf_{d_R(s) \in D^{MR}} \inf_{\{\Pi(r)|r<R\}} V^{d_R(s) \cup \{\Pi(r)|r<R\}}(R,s)$$

$$= \inf_{d_R(s) \in D^{MR}} \sum_{a \in \alpha(s)} q_{d_R}(a) \cdot \{P(s_0|s,a) \cdot I[g(s,a) > R] +$$

$$\sum_{x \neq s_0} P(x|s,a) \inf_{\{\Pi(r)|r<R\}} V^{\{\Pi(r)|r<R\}}(R - g(s,a), x)\}$$

$$= \inf_{d_R(s) \in D^{MR}} \sum_{a \in \alpha(s)} q_{d_R}(a) \cdot \{P(s_0|s,a) \cdot I[g(s,a) > R] + \sum_{x \neq s_0} P(x|s,a)V(R - g(s,a), x)\}$$

$$= \inf_{d_R(s) \in D^{MR}} \sum_{a \in \alpha(s)} q_{d_R}(a) \cdot \{P(s_0|s,a) \cdot V(R - g(s,a), s_0) + \sum_{x \neq s_0} P(x|s,a)V(R - g(s,a), x)\}$$

$$= \inf_{d_R(s) \in D^{MR}} \sum_{a \in \alpha(s)} q_{d_R}(a) \cdot \sum_{x \in r(s,a)} P(x|s,a)V(R - g(s,a), x)\}.$$

And the Lemma follows by Theorem 2.0.2. □

Lemma 3.1.2 *There exists an optimal policy* Π^* *Furthermore,* Π^* *is Markov and deterministic.*

Proof:

Because $\alpha(s)$ is compact and $P(x|s,a), g(s,a)$ are continuous in a; the infimum in (3.3) is achieved. Since it is achieved, and since there are a finite number of possible realizations since the budgeted time $B < \infty$, the policy of choosing the infimum in (3.3) is the optimal decision rule, i.e., its value function is the optimal value function. Clearly, this decision rule is Markov and deterministic. □

Lemma 3.1.3 $V(R,s)$ *is a piecewise constant, nonincreasing function of* R*, with* $V(R,s) = 1$ *for* $R < 0$ *and* $\lim_{R\to\infty} V(R,s) = 0$*. Furthermore,* $V(R,s)$ *is continuous from the right in* R*.*

Proof:

For $R \leq 0$, the required conditions hold by (3.2). Also, the required conditions hold for $V(\cdot, s_0)$ by (3.1). Let us proceed with a proof by induction on R. Let $\epsilon = \min_{x\in S, a\in\alpha(x)} g(x,a)$. Assume the required conditions hold $\forall s\ \forall R < n\epsilon$ with $0 \leq n \in \Re$. We will show that the required conditions hold $\forall s\ \forall R < (n+1)\epsilon$. For simplicity, if a function is piecewise constant, nonincreasing and continuous from the right, we call it a PCNICR function.

Let us examine the value of $V(R,\bar{s})$ for some state $\bar{s}$. By the DP (3.3), we see that over the interval $[n\epsilon, (n+1)\epsilon]$, $V(\cdot,\bar{s})$ is, for each R, the minimum over the set of admissible actions of $\sum_{x\in r(\bar{s},a)} P(x|\bar{s},a)\cdot V(R - g(\bar{s},a), x)$. This term is the weighted average of PCNICR functions, so it is PCNICR. Furthermore, the minimum of a finite set of PCNICR functions is PCNICR. Therefore, we see that $V(\cdot,\bar{s})$ is PCNICR over the interval $[n\epsilon, (n+1)\epsilon]$, and the induction is established. □

3.2 The deadline problem with incremental costs

Instead of viewing the value function in Chapter 3 as the probability of not reaching the goal state by the deadline time, it could be viewed as the expected value of the cost, where the MDP terminates when the goal state is reached, and the only cost ever incurred is a cost of 1 for not being in the goal state when the deadline time is reached. Let us call this deadline penalty cost DLP and allow it to take on values other than 1. Furthermore, in this section, we add an incremental cost equal to ρ times the time elapsed prior to reaching the goal state in excess of the budgeted time.

Define $\zeta(s)$ as the optimal value function for the stochastic shortest path problem of reaching the goal state in the shortest time. $\zeta(\cdot)$ can be found using standard methods. It can be seen that $\zeta(s_0) = 0$.

Lemma 3.2.1 *The value function iteration (for $R > 0$) for the deadline problem with incremental costs is*

$$V(R, s) = \min_{a \in \alpha(s)} \sum_{x \in r(s,a)} P(x|s, a) \cdot V(R - g(s, a), x); R > 0 \tag{3.4}$$

with the boundary conditions

$$V(R, s) = DLP + \rho \cdot (\zeta(s) - R); R < 0$$

$$V(0, s_0) = 0$$

$$V(0, s) = DLP + \rho \cdot \zeta(s); s \neq s_0$$

Proof:

The boundary conditions are the only part that needs to be proved, since the dynamic program was shown to be true in Lemma 3.1.1. The boundary conditions are seen to hold since the penalty DLP is assessed if the goal state has not been reached before $R = 0$. And the incremental penalty is ρ times the additional time required to reach the goal state. □

Notice that we obtain the value function iteration of the original problem without incremental costs if we set $DLP = 1$ and eliminate the incremental cost term.

Lemma 3.2.2 *If incremental costs are included, then $V(R, s)$ is piecewise linear, nonincreasing, and continuous from the right (PLNCR) in R, with $V(R, s) = DLP + \rho R$ for $R < 0$ and $\lim_{R \to \infty} V(R, s) = 0$.*

Proof:

Given the boundary conditions (the value of $V(R, s)$ for $R \leq 0$ from Lemma 3.2.1, we see that the lemma is satisfied for $R \leq 0$.

As in Lemma 3.1.3, we proceed by induction:

Let $\epsilon = \min_{x \in S, a \in \alpha(x), y \in r(s,a)} g(x, a)$ Assume that the conditions of the lemma are satisfied $\forall s \ \forall R \leq n\epsilon$ where $0 \leq n \in \Re$. We will show that the conditions of the lemma are then satisfied $\forall s \ \forall R \leq (n + 1)\epsilon$.

Let us examine the value of $V(R, \bar{s})$ for some state $\bar{s}$. By the DP (equation 3.4), we see that over the interval $[n\epsilon, (n+1)\epsilon]$, $V(\cdot, \bar{s})$ is, for each R, the minimum over the set of admissible actions of $\sum_{x \in r(s,a)} P(x|s, a) \cdot V(R - g(s, a), x)$. This term is the weighted average of PLNCR functions, so it is PLNCR. Furthermore, the minimum of a finite set of PLNCR functions is PLNCR. Therefore, we see that $V(\cdot, \bar{s})$ is PLNCR over the interval $[n\epsilon, (n + 1)\epsilon]$, and the induction is established. □

3.3 The discrete time deadline problem without incremental costs

Suppose we ignore the continuous time between transitions and focus on the objective of reaching the goal state within B transitions, i.e., suppose we set $g(\cdot, \cdot) = 1$. The definition of $V(R, s)$ is unchanged: $V_\Pi(R, s) =$ the probability of not reaching the goal state within R transitions by following policy Π.

In order to examine this system more easily, let us transform it into an equivalent finite horizon, risk-sensitive MDP. The objective function for a risk-sensitive MDP (with risk sensitivity parameter $\gamma = 1$ and time horizon R) is

$$\bar{J}_d'^{\Pi}(R, x_0) = E_{x_0}^{\Pi}[e^{\sum_{i=0}^{R-1} c(x_i, a_i)}]$$

where the d subscript is for 'deadline problem'.

Since no costs are defined for the deadline problem, and in fact the objective function of the deadline problem is the probability of not reaching the goal state within the time budget, we set the costs for the equivalent risk sensitive problem to be the log of the probability of not reaching the goal state in a single transition. Since such a framework doesn't make sense once the goal state is reached, we eliminate the goal state and set the boundary cost for each state to 1; i.e. $V(0, s) = 1\ \forall s$.

We must transform the transition probabilities in order to eliminate the goal state. (Note: Assumption A1 guarantees that the following procedure is well-defined.)

For each (s, a), set $r(s, a) = r(s, a) - \{s_0\}$ and set $P(x|s, a) = \frac{P(x|s,a)}{1-P(s_0|s,a)}$. And set the transition cost to $c(s, a) = \log(1 - P(s_0|s, a))$.

It can be seen that the deadline problem (3.3) is equivalent to the following optimization problem in the transformed system:

$$V(R, s) = \min_{a \in \alpha(s)} [e^{c(s,a)} \sum_{x \in r(s,a)} P(x|s, a) V(R - 1, x)]$$

which is the standard D.P. equation for a risk sensitive control problem. Standard value iteration for a completely observed risk sensitive MDP can then be used to find the optimal policy $\Pi(R)$.

The following facts are standard for a finite state, discrete time risk sensitive control problem ([23]):

1. There exists a number $\beta < \infty$ such that, $\forall\ R_1, R_2 > \beta$, $\Pi(R_1) = \Pi(R_2)$. Furthermore, they each equal the stationary policy that optimizes the infinite horizon risk-sensitive average cost

$$J_{x_0}^{\prime\Pi} = \lim_{R\to\infty} \frac{1}{R} log[\bar{J}_d^{\Pi}(x_0, R)],$$

denoted Π_{AC}.

2. $\Pi(R)$ is a contraction mapping from $V(R)$ to $V(R-1)$, and the largest eigenvalue of Π_{AC} is less than or equal to the largest eigenvalue of any other policy.

3. Let W_{AC}^{MAX} denote the eigenvector corresponding to the largest eigenvalue of Π_{AC}, and let w_{AC}^{MAX} be the eigenvalue. Then there exists a constant γ such that $lim_{R\to\infty} \frac{V(R)}{[w_{AC}^{MAX}]^R} = \gamma W_{AC}^{MAX}$.

3.4 The deadline problem for a general SMDP

Now suppose that transition times are arbitrarily distributed, i.e., instead of $\Delta t = g(s, a)$, we have a general density function $f_{s,a}(\tau)$ on Δt such that $\Delta t > 0$ w.p.1 $\forall\{s, a\}$ and $E[\Delta t] < \infty$ $\forall\{s, a\}$.

Lemma 3.4.1 *Equations 3.1 and 3.2 still hold, but equation 3.3 is replaced by*

$$V(R, s) = \min_{a\in\alpha(s)} \sum_{x\in r(s,a)} P(x|s, a) \int_{\tau=0}^{R} V(R-\tau, x) f_{s,a}(\tau) d\tau; \text{ if } R > 0 \qquad (3.5)$$

Proof:

The boundary conditions hold trivially. (3.5) can be shown to be true by arguments similar to those used in the proof of Lemma 3.1.1. □

This dynamic program is very difficult to solve directly, so we will construct the optimal value function to the deadline problem as the limit of a sequence of value functions of *truncated* deadline problems. Recall that $V(R, s)$ is defined as the probability of not reaching the goal state within the time remaining under the optimal policy.

The truncated deadline problem has an additional constraint: the goal state must be reached not only within the deadline time, but also within k transitions. If k is given, the problem is called the *k-truncated* deadline problem. Define $V^k(R,s)$ as the probability of not reaching the goal state within k transitions in the time remaining under the optimal policy to the k-truncated deadline problem.

We can see that equations 3.1 and 3.2 hold $\forall k$, and a recursion to determine $V^k(R,s)$, in terms of $V^{k-1}(R,s)$ is

$$V^k(R,s) = \min_{a\in\alpha(s)} \sum_{x\in r(s,a)} P(x|s,a) \int_{\tau=0}^{R} V^{k-1}(R-\tau,x) f_{s,a}(\tau)d\tau; \text{ if } R>0 \tag{3.6}$$

Clearly,

$$V^0(R,s) = \begin{cases} 0 \text{ if } s = s_0 \\ 1 \text{ if } s \neq s_0 \end{cases},$$

so $V^k(R,s)$ can be solved by convolving known functions, adding, and taking a minimum over admissible actions.

In order for this recursion to converge to the optimal value function, we need to assure that only a finite number of transitions take place in a finite time interval.

Assumption 3.4.1 *([35])*

There exist $\epsilon > 0$ and $\delta > 0$ such that

$$P[\Delta t \leq \delta] \leq 1-\epsilon$$

$\forall x \in S$ and $a \in \alpha(x)$.

Lemma 3.4.2 *Under Assumption 3.4.1, for any $R \in \Re^+$, $\lim_{k\to\infty} V^k(R,s) = V(R,s)$.*

Proof:

Let $R < \infty$ be given. Under policy Π starting from initial state s, let $T(k)$ denote the time elapsed in the first k transitions. Out of the first k transitions, let the number that have transition times exceeding δ be denoted by $B(k)$. Denote $G = \lceil \frac{R}{\delta} \rceil$; i.e., G is the least integer that is no less than $\frac{R}{\delta}$. We have

$$P[T(k) > R] \leq P[B(k) > G]$$

and by Assumption 3.4.1,

$$P[B(k) \leq G] \leq \sum_{i=0}^{G} (1-\epsilon)^{k-i}(\epsilon)^i.$$

Clearly, $\lim_{k\to\infty} P[B(k) \leq G] = 0$, so we have that $\lim_{k\to\infty} P[T(k) > R] = 1$. The optimal value function can be bounded by

$$V^k(R,s) - P[T(k) < R] \leq V(R,s) \leq V^k(R,s)$$

And $\lim_{k\to\infty} P[T(k) < R] = 0$. □

We have now an algorithm for approximating the value function of the deadline problem to arbitrary precision.

This algorithm can be applied to the deadline problem with deterministic transition times to obtain the exact value function. It is clear that $V(R,s) = V^k(R,s)$ for $k > \frac{R}{\epsilon}$ where $\epsilon = \min_{x,a} g(x,a)$.

3.5 Rate of accrual of costs in an SMDP

In order to study a finite or infinite horizon control problem with costs, one must know more than the joint density function on total cost and time to complete a transition. One must know the rate of accrual of costs. This is because in a finite horizon control problem, the time limit may be reached at a time other than a transition time. And in an infinite horizon control problem, the limit must be reached uniformly, not just at transition times.

$c(x,a)$, the total transition cost between decision epochs, and $t(x,a)$, the transition time betweem decision epochs, have a joint density function given by Z. Let $G_{x,a}(\tau)$ denote the rate of accrual of costs, and in addition, let $h(x,a)$ denote a one time cost that takes place at the time of transition. We have

$$c(x,a) = [h(x,a) + \int_0^{t(x,a)} G_{x,a}(\tau)d\tau] \text{ w.p.1.}$$

We assume that $G_{x,a}(\tau) \geq 0$ w.p.1 $\forall \tau \geq 0$ and that $h(x,a) \geq 0$ w.p.1. Note that $h(x,a)$ does not depend on the transition time or the accrued costs up to the transition time.

Given that the total cost of transition is generated in this way, we can see that

$$E[f(c(x,a))|t(x,a) = A_1] \geq E[f(c(x,a))|t(x,a) = A_2]; \text{ if } A_1 \geq A_2. \tag{3.7}$$

for any monotone increasing function $f(\cdot)$. This fact will be useful later on in proving Lemma 4.2.2 and Theorem 4.2.2 by bounding the costs.

For simplicity, we will combine $G(\cdot)$ and $h(\cdot,\cdot)$ into $g(\cdot) \doteq G(\cdot)+$ a delta function of magnitude $h(\cdot,\cdot)$ at each transition time. Thus, the total (risk neutral) cost up to time T is given by $\int_0^T g(t)dt$.

It should be noted that a joint probability density function on $c(x,a)$ and $t(x,a)$ is general enough to model rates of accrual of costs other than the one we adopt here.

3.6 The cost minimization problem for a finite-horizon SMDP

A finite horizon cost minimization problem can take one of two forms:

$$\bar{J}^{\Pi}_{x_0}(T) = \frac{1}{\gamma T} \ln E^{\Pi}_{x_0}[e^{\gamma \int_{t=0}^{T} g(t)dt}] \tag{3.8}$$

or

$$\bar{\mathcal{J}}^{\Pi}_{x_0}(T) = E^{\Pi}_{x_0}[\int_{t=0}^{T} g(t)dt],$$

where the time horizon is T.

We can extend the same procedure to solve either of these problems that we used to solve the deadline problem for a general SMDP.

Define

$$\bar{J}^{*}_{x_0}(T) = \inf_{\Pi\in\Pi^{HR}} \bar{J}^{\Pi}_{x_0}(T),$$

and similarly for $\bar{\mathcal{J}}^{*}_{x_0}(T)$. Existence and dynamic programming results similar to Lemma 3.1.2 and Lemma 3.1.1 can be shown in both the risk sensitive and risk neutral cases.

Recall the definition of the 'k-truncated' deadline problem. We define the *k-truncated* finite horizon, risk sensitive, cost minimization problem in terms of its objective function as follows:

$$\bar{J}^{*k}_{x_0}(T) = E^{\Pi}_{x_0}[e^{\gamma\int_{t=0}^{T} g(t)h_k(t)dt}],$$

where $nt(t)$ = the number of transitions that have taken place up to time t,

$$h_k(t) = \begin{cases} 0 \text{ if } nt(t) \geq k \\ 1 \text{ if } nt(t) < k \end{cases},$$

and we have gotten rid of the log and the normalization in (3.8) for simplicity (and without affecting the optimal policy).

It is easily seen that $\bar{J}^{*0}_{x_0}(T) = 0$. Furthermore, the following recursion can be shown to hold:

$$\bar{J}^{*k}_{s}(T) = \min_{a\in\alpha(s)}\{P[t(x,a) > T]\cdot E[e^{\gamma\int_{t=0}^{T} g(t)dt}|t(x,a) > T]+$$

$$\sum_{x\in r(s,a)} P(x|s,a)\int_{\tau=0}^{T} E[e^{\gamma c(s,a)}|t(x,a) = \tau]\bar{J}^{*k-1}_{s}(T-\tau)f_{s,a}(\tau)d\tau\}.$$

Similarly, for the risk neutral case it can be shown that

$$\bar{\mathcal{J}}_s^{*k}(T) = \min_{a \in \alpha(s)} \{P[t(x,a) > T] \cdot E[\int_{t=0}^{T} g(t)dt | t(x,a) > T] +$$

$$\sum_{x \in r(s,a)} P(x|s,a) \int_{\tau=0}^{T} E[c(s,a)|t(x,a) = \tau] \bar{\mathcal{J}}_s^{*k-1}(T-\tau) f_{s,a}(\tau) d\tau\}$$

and $\bar{\mathcal{J}}_{x_0}^{*0}(T) = 0$.

Chapter 4

Average Costs over the Infinite Horizon

4.1 Average cost objective functions

Risk neutral control problems have been well explored in both the finite horizon and infinite horizon cases. For the infinite horizon, the semi-Markov risk neutral cost case can be solved through value and policy iteration for both average and discounted costs. This case was first studied in [22]; it is examined in more detail in [5]; and [35] gives a thorough treatment with references. In the average cost case, the objective function is given by:

$$\mathcal{J}_{x_0}^{\Pi} = \limsup_{T\to\infty} \frac{1}{T} E_{x_0}^{\Pi}[\int_0^T g(t)dt] = \limsup_{T\to\infty} \frac{E_{x_0}^{\Pi}[\int_0^T g(t)dt]}{T},$$

where $g(t)$ is the rate of accrual of cost at time t. Under suitable conditions (see, e.g., the verification theorems in this chapter), it can be shown that if the policy is unichain and stationary, the limsup above can be replaced by a lim and that the limit can be taken at the sequence transition times,

$$\mathcal{J}_{x_0}^{\Pi} = \lim_{N\to\infty} E_{x_0}^{\Pi}[\frac{\sum_{i=0}^{N} c(x_i, a_i)}{\sum_{i=0}^{N} t(x_i, a_i)}]$$

$$= \lim_{N\to\infty} E^{\Pi}_{x_0}[\frac{\frac{1}{N}\sum_{i=0}^{N} c(x_i, a_i)}{\frac{1}{N}\sum_{i=0}^{N} t(x_i, a_i)]}$$

where N is the number of decision epochs that have occured, $c(x_i, a_i)$ is the total cost accrued between the i^{th} and $i+1^{th}$ decision epochs, and $t(x_i, a_i)$ is the total time elapsed between the i^{th} and $i+1^{th}$ decision epochs. It can be shown ([35],[5]) that the above is the same as

$$\tilde{\mathcal{J}}^{\Pi}_{x_0} = \frac{\lim_{N\to\infty}\frac{1}{N}E^{\Pi}_{x_0}[\sum_{i=0}^{N} c(x_i, a_i)]}{\lim_{N\to\infty}\frac{1}{N}E^{\Pi}_{x_0}[\sum_{i=0}^{N} t(x_i, a_i)]},$$

i.e., the ratio of the limits is the limit of the ratio.

In Chapter 10, we will show that the expectation operator can be removed in these limits under the proper conditions.

Risk sensitive control was first described in [23], the discounted costs case was explored in [10], and a good survey is given in [30]. Solving the discrete time, discounted, risk sensitive cost case is difficult. In fact, it is shown in [10] that the solution in the discounted cost case is not stationary because the risk factor is different at every time, so policy iteration cannot be used. (See in particular PP. 56-57 of [10] and references therein.) Note that ([43]) the discounted costs case degenerates to the risk neutral case as $t \to \infty$.

In discrete time, the average risk sensitive cost over the infinite horizon is defined (see [30]) as

$$J'^{\Pi}_{x_0} = \lim_{T\to\infty}\frac{1}{\gamma T}\ln E^{\Pi}_{x_0}[e^{\gamma\sum_{t=0}^{T-1} c(x_t, a_t)}];\ \gamma > 0.$$

Results for this case are well understood, but the semi-Markov case does not appear to have been studied in the literature.

Define the risk-sensitive *ratio* objective function as follows:

$$\tilde{J}^{\Pi}_{x_0} = \limsup_{N\to\infty}\frac{1}{\gamma E^{\Pi}_{x_0}[\sum_{k=0}^{N-1} t_k]}\ln E^{\Pi}_{x_0} e^{\gamma\sum_{k=0}^{N-1} c(x_k, a_k)};\ \gamma > 0 \tag{4.1}$$

The actual risk sensitive objective function for an SMDP is

$$J_{x_0}^{\Pi} = \limsup_{T\to\infty} \frac{1}{\gamma T} \ln E_{x_0}^{\Pi}[e^{\int_0^T \gamma g(t)dt}]$$

and $J_{x_0}^{\Pi} \neq \tilde{J}_{x_0}^{\Pi}$ in general because the numerator and denominator cannot be separated in the average cost risk-sensitive semi-Markov case. (This is because the exponential and logarithm are not linear operators.)

The central focus of this thesis will be to examine the risk sensitive objective function $J_{x_0}^{\Pi}$.

Of course, other objective functions can be used. For example, in section 10.2 of chapter 10, we will examine the following objective function:

$$J_x^{\Pi}(\text{no } \gamma) = \lim_{N\to\infty} \ln E_x^{\Pi}[e^{\frac{1}{N}\sum_{k=0}^{N-1} c(x_k,a_k)}].$$

This objective function is similar to the risk sensitive objective function with the interesting property that the term $\frac{1}{N\gamma}$ has been taken inside the expectation and the exponential, resulting in the cancellation of the parameter γ. It turns out that it exists if and only if the risk sensitive objective function exists for at least one value of $\gamma > 0$, and its value is related to the value of the risk neutral objective function.

4.2 A dynamic program for the risk sensitive semi-Markov average cost case

In [18], the following dynamic programming equation is considered for the discrete time risk sensitive control problem:

$$e^{\lambda + W(x)} = \min_{a\in\alpha(x)} e^{\gamma c(x,a)} \int e^{W(y)} P(dy|x,a); \forall x \in S$$

where $\alpha(x)$ is the set of actions available in state x, $c(x,a)$ is the cost of taking action a when in state x, and $P(dy|x,a)$ is the transition probability density function for taking action a in state x.

In [18], sufficient conditions are found for the existence of a bounded solution $(\lambda, W(x))$ to the dynamic program, where λ is the average cost and $W(x)$ is the certainty equivalent of being in state x. Furthermore, the convergence properties of value and policy iteration have been explored in the literature (see, e.g., [30] for a survey, and [7] for a recent result.)

A natural extension of the discrete time dynamic programming equation into continuous time would be:

$$e^{W(x)} = \min_{a \in \alpha(x)} E[e^{\gamma\{c(x,a) - \lambda t(x,a)\}}] \int e^{W(y)} P(dy|x,a); \forall x \in S \tag{4.2}$$

Lemma 4.2.1 *If (4.2) holds, then*

$$E^{\Pi}_{x_0}[e^{\gamma \sum_{i=0}^{N} c(x_i,\Pi(x_i)) - \gamma\lambda \sum_{i=0}^{N} t(x_i,\Pi(x_i))}] \geq E^{\Pi}_{x_0}[\Pi_{i=0}^{N}\{\frac{e^{W(x_i)}}{\int e^{W(y)} P(dy|x_i, \Pi(x_i))}\}]. \tag{4.3}$$

Furthermore, if Π^* *is a stationary, Markov, deterministic policy such that* $\Pi^*(x)$ *minimizes (4.2) for each* $x \in S$*, then (4.3) holds with equality for* $\Pi = \Pi^*$.

Proof:

We use induction on N. By rearranging terms and replacing the minimum with the appropriate inequality in (4.2), we get

$$E[e^{\gamma\{c(x,a) - \lambda t(x,a)\}}] \geq \frac{e^{W(x)}}{\int e^{W(y)} P(dy|x,a)}; \forall x \in S \tag{4.4}$$

The fact that (4.3) holds for $N = 0$ follows directly from (4.4). Now, suppose that (4.3) holds for N.

$$E^{\Pi}_{x_0}[e^{\gamma \sum_{i=0}^{N+1} c(x_i,\Pi(x_i)) - \gamma\lambda \sum_{i=0}^{N+1} t(x_i,\Pi(x_i))}]$$

$$= E^{\Pi}_{x_0}[e^{\gamma \sum_{i=0}^{N} c(x_i,\Pi(x_i)) - \gamma\lambda \sum_{i=0}^{N} t(x_i,\Pi(x_i))}.$$

$$E[e^{\gamma\{c(x_{N+1},\Pi(x_{N+1})) - \lambda t(x_{N+1},\Pi(x_{N+1}))\}}|x_{N+1}]]$$

$$\geq E_{x_0}^{\Pi}[\Pi_{i=0}^{N}\{\frac{e^{W(x_i)}}{\int e^{W(y)}P(dy|x_i,\Pi(x_i))}\}\cdot E[e^{\gamma\{c(x_{N+1},\Pi(x_{N+1}))-\lambda t(x_{N+1},\Pi(x_{N+1}))\}}|x_{N+1}]]$$

$$\geq E_{x_0}^{\Pi}[\Pi_{i=0}^{N}\{\frac{e^{W(x_i)}}{\int e^{W(y)}P(dy|x_i,\Pi(x_i))}\}\cdot \frac{e^{W(x_{N+1})}}{\int e^{W(y)}P(dy|x_{N+1},\Pi(x_{N+1}))}].$$

Finally, all of the inequalities above are replaced by equality if $\Pi = \Pi^*$. □

The following assumptions are used in Theorem 4.2.1 to bound the costs between decision epochs:

Assumption 4.2.1 $\exists L > 0$ *such that* $L < E[e^{-\gamma\lambda t(x,a)}]$ $\forall x, a$.

Assumption 4.2.2 $\exists U < \infty$ *such that* $E[e^{\gamma c(x,a)}] < U$ $\forall x \in S, a \in \alpha(x)$.

The following theorem is an extension of Theorem 2.1 in [18], which covers the discrete time case.

Theorem 4.2.1 (Verification Theorem) *Suppose* $\{W(x), \lambda\}$ *is a bounded solution to the dynamic program (4.2) and that Assumptions 4.2.1 and 4.2.2 hold. Let* Π^* *be a stationary, Markov, deterministic policy such that* $\Pi^*(x)$ *minimizes the dynamic program for each* $x \in S$.

Then, $J^{\Pi^*}(x_0) = \lambda$ $\forall x_0$, *and furthermore* Π^* *is optimal with respect to the objective function* $J_{x_0}^{\Pi} \forall x_0 \in S$.

Proof:

First, we are guaranteed that $\Pi^*(x)$ exists since $\alpha(x)$ is compact and the joint density funtion on $(c(x,a), t(x,a))$ is continous in a.

From Lemma 4.2.1, (4.3) holds, and holds with equality for $\Pi = \Pi^*$.

In order to bound the right hand side of (4.3), we note that, by the Markov property of the underlying Markov chain,

$$E^{\Pi}_{x_0}[\int e^{W(y)}P(dy|x_N,\Pi(x_N))\cdot\Pi_{i=0}^{N}\{\frac{e^{W(x_i)}}{\int e^{W(y)}P(dy|x_i,\Pi(x_i))}\}]=e^{W(x_0)} \tag{4.5}$$

.

Since $W(\cdot)$ is bounded, $\exists -\infty < LB < UB < \infty$ such that $LB \leq W(x) \leq UB$ $\forall x \in S$.

Substituting $e^{LB} \leq \int e^{W(y)}P(dy|x_{N+1},\Pi(x_{N+1}))$ into (4.5), we obtain

$$E^{\Pi}_{x_0}[e^{LB}\cdot\Pi_{i=0}^{N}\{\frac{e^{W(x_i)}}{\int e^{W(y)}P(dy|x_i,\Pi(x_i))}\}]\leq e^{W(x_0)}\leq e^{UB}.$$

Noting that (4.3) holds with equality for $\Pi = \Pi^*$ and setting $B_2 = e^{UB-LB} < \infty$, we get

$$E^{\Pi^*}_{x_0}[e^{\gamma\sum_{i=0}^{N}c(x_i,\Pi^*(x_i))-\gamma\lambda\sum_{i=0}^{N}t(x_i,\Pi^*(x_i))}]\leq B_2;\ \forall N,x_0 \tag{4.6}$$

.

Now substituting $e^{UB} \geq \int e^{W(y)}P(dy|x_{N+1},\Pi(x_{N+1}))$ into (4.5), we obtain

$$E^{\Pi}_{x_0}[e^{UB}\cdot\Pi_{i=0}^{N}\{\frac{e^{W(x_i)}}{\int e^{W(y)}P(dy|x_i,\Pi(x_i))}\}]\geq e^{W(x_0)}\geq e^{LB}.$$

Combining with (4.3) and denoting $B_1 = e^{LB-UB} > 0$, we get

$$B_1 \leq E^{\Pi}_{x_0}[e^{\gamma\sum_{i=0}^{N}c(x_i,\Pi(x_i))-\gamma\lambda\sum_{i=0}^{N}t(x_i,\Pi(x_i))}];\ \forall\Pi,N,x_0. \tag{4.7}$$

Let t_N be the N^{th} transition time, i.e., $t_N = \sum_{i=0}^{N-1}t(x_i,\Pi^*(x_i))$. Therefore we have the equality

$$E^{\Pi^*}_{x_0}[e^{\gamma\sum_{i=0}^{N}c(x_i,\Pi^*(x_i))-\gamma\lambda\sum_{i=0}^{N}t(x_i,\Pi^*(x_i))}]=E^{\Pi^*}_{x_0}[e^{\gamma\{\int_{t=0}^{t_N}g(t)dt-\lambda T\}}]$$

.

We see from (4.7) and (4.6) that

$$B_1 \leq E^{\Pi^*}_{x_0}[e^{\gamma\{\int_{t=0}^{t_N}g(t)dt-\lambda T\}}]\leq B_2;\ \forall N,x_0$$

which describes the limiting behavior of the objective function for policy Π^* at transition times. (4.7) tells us that no other policy does better than Π^* at transition times, i.e., for the underlying discrete-time Markov chain. We have,

$$B_1 \leq [e^{\gamma\{\int_{t=0}^{t_N} g(t)dt - \lambda T\}}];\ \forall N, x_0 \tag{4.8}$$

.

Let $N(t)$ be the number of transitions that have occured prior to time t. Therefore, we have that

$$t_{N(t)} \leq t \leq t_{N(t)+1}$$

Similarly, we have

$$E_{x_0}^{\Pi^*}[e^{\gamma \sum_{i=0}^{N(T)} c(x_i,\Pi^*(x_i)) - \gamma\lambda \sum_{i=0}^{N(T)+1} t(x_i,\Pi^*(x_i))}]$$

$$\leq E_{x_0}^{\Pi^*}[e^{\gamma\{\int_{t=0}^{T} g(t)dt - \lambda T\}}]$$

$$\leq E_{x_0}^{\Pi^*}[e^{\gamma \sum_{i=0}^{N(T)+1} c(x_i,\Pi^*(x_i)) - \gamma\lambda \sum_{i=0}^{N(T)} t(x_i,\Pi^*(x_i))}]$$

And so by assumptions A1.1 and A1.2, we see that

$$L \cdot B_1 < E_{x_0}^{\Pi^*}[e^{\gamma\{\int_{t=0}^{T} g(t)dt - \lambda T\}}] < U \cdot B_2;\ \forall T.$$

Therefore,

$$L \cdot B_1 < \frac{E_{x_0}^{\Pi^*}[e^{\gamma \int_{t=0}^{T} g(t)dt}]}{e^{\gamma\lambda T}} < U \cdot B_2;\ \forall T.$$

Taking the natural log of all three sides of the inequality, then dividing by γT and then taking the limit as $T \to \infty$, we see that $J^{\Pi^*}(x_0) = \lambda\ \forall x_0$. And a similar argument from (4.8) shows that $J_{x_0}^{\Pi} \geq \lambda\ \forall x_0, \Pi$. $\Box$

Lemma 4.2.2 *If Assumption 4.2.1 holds, then* $\exists \kappa < \infty$ *such that* $E[e^{\gamma c(x,a)}] \leq \kappa \cdot E[e^{\gamma\{c(x,a) - \lambda t(x,a)\}}]$ *for all* x, a.

Proof:

Let $B_{x,a} \doteq E[e^{\gamma\{c(x,a)-\lambda t(x,a)\}}]$ and denote $z \doteq e^{-\gamma\lambda t(x,a)}$. We know from Assumption 4.2.1 that $L < E[z]$. And since $\infty > t(x,a) > 0$ w.p.1, we also know that $0 < z < 1$. Let $f(\cdot)$ be the probability density function for z; i.e., $P[a < z < b] = \int_a^b f(z)dz$. (Note that $f(\cdot)$ exists because it is given in terms of the joint density Z on $c(x,a)$ and $t(x,a)$.)

Claim:

$P[z > \frac{L}{4}] \geq \frac{L}{2}$.

Proof of claim:

Suppose that $P[z > \frac{L}{4}] < \frac{L}{2}$. Then,

$$E[z] = \int_0^1 zf(z)dz = \int_0^{\frac{L}{4}} zf(z)dz + \int_{\frac{L}{4}}^1 zf(z)dz$$

$$\leq \int_0^{\frac{L}{4}} \frac{L}{4}f(z)dz + \int_{\frac{L}{4}}^1 1f(z)dz \leq \frac{L}{4} + \frac{L}{2} < L,$$

contradicting Assumption 4.2.1.

Let $h(z) = E[e^{\gamma c(x,a)}|[e^{-\gamma\lambda t(x,a)} = z]$. By (3.7), $h(\cdot)$ is monotone increasing. Also, $E[zh(z)] = \int_0^1 zh(z)f(z)dz = B_{x,a}$ and $E[h(z)] = \int_0^1 h(z)f(z)dz = E[e^{\gamma c(x,a)}]$ by definition.

We have

$$E[e^{\gamma c(x,a)}] = \int_0^1 h(z)f(z)dz$$

$$= \int_0^{\frac{L}{4}} h(z)f(z)dz + \int_{\frac{L}{4}}^1 h(z)f(z)dz$$

$$\leq \int_0^{\frac{L}{4}} f(z)h(\frac{L}{4})dz + \frac{4}{L}\int_{\frac{L}{4}}^1 zh(z)f(z)dz$$

$$\leq P[z \leq \frac{L}{4}]h(\frac{L}{4}) + \frac{4}{L}\int_0^1 zh(z)f(z)dz$$

$$\leq \frac{P[z \leq \frac{L}{4}]}{P[z \geq \frac{L}{4}]} \int_{\frac{L}{4}}^{1} h(z)f(z)dz + \frac{4}{L}B_{x,a}$$

$$\leq \frac{1}{\frac{L}{2}} \int_{\frac{L}{4}}^{1} h(z)f(z)dz + \frac{4}{L}B_{x,a}$$

$$\leq \frac{1}{\frac{L}{2}} \frac{4}{L} \int_{\frac{L}{4}}^{1} zh(z)f(z)dz + \frac{4}{L}B_{x,a}$$

$$\leq \frac{8}{L^2} \int_{0}^{1} zh(z)f(z)dz + \frac{4}{L}B_{x,a} = (\frac{4}{L} + \frac{8}{L^2})B_{x,a}.$$

So the lemma is true for $\kappa = \frac{4}{L} + \frac{8}{L^2}$. □

The following assumption is called the *norm like* condition on the cost function in [7]:

Assumption 4.2.3 $\lim_{x \to \infty} \inf_{a \in \alpha(x)} E[e^{\gamma\{c(x,a) - \lambda t(x,a)\}}] = \infty$.

Theorem 4.2.2 (Verification Theorem for unbounded costs) *Suppose that $\{W(x), \lambda\}$ is a solution to the dynamic program (4.2) and that $\{W(x)\}$ is finite for each x and bounded below. Suppose furthermore that Assumption 4.2.1 and Assumption 4.2.3 hold. Let Π^* be a stationary, Markov, deterministic policy such that $\Pi^*(x)$ minimizes the dynamic program for each $x \in S$.*

Then, $J^{\Pi^*}(x_0) = \lambda \; \forall x_0$, and furthermore Π^* is optimal with respect to the objective function $J_{x_0}^{\Pi} \forall x_0 \in S$.

Proof:

Equations (4.3) and (4.5) still hold, and again (4.3) holds with equality for $\Pi = \Pi^*$. Since the term $\int e^{W(y)} P(dy|x_{N+1}, \Pi^*(x_{N+1}))$ is bounded below, (4.6) holds true, but with a different bound for each initial state:

$$E_{x_0}^{\Pi^*}[e^{\gamma\sum_{i=0}^{N} c(x_i,\Pi^*(x_i))-\gamma\lambda\sum_{i=0}^{N} t(x_i,\Pi^*(x_i))}] \leq B_2(x_0);\ \forall N. \tag{4.9}$$

$B_2(x_0) < \infty$ depends on the initial state, x_0, since $e^{W(x_0)}$ is no longer bounded above $\forall x_0$.

Because $W(\cdot)$ is not bounded above, (4.7) does not hold. Instead, we substitute the stopping time $v_n(A) \doteq$ the n^{th} visit time to set $A \subset S$ (i.e., the time of the n^{th} transition to A) into (4.3), yielding

$$E_{x_0}^{\Pi}[e^{\gamma\sum_{i=0}^{v_n(A)} c(x_i,\Pi(x_i))-\gamma\lambda\sum_{i=0}^{v_n(A)} t(x_i,\Pi(x_i))}]$$

$$\geq E_{x_0}^{\Pi}[\Pi_{i=0}^{v_n(A)}\{\frac{e^{W(x_i)}}{\int e^{W(y)}P(dy|x_i,\Pi(x_i))}\}]. \tag{4.10}$$

If A has finitely many elements, then $\int e^{W(y)}P(dy|x_i,\Pi(x_i))$ is finite by Assumption 4.2.3, so it is bounded above. Therefore, if we also have that $P_{x_0}^{\Pi}[\tau_A < \infty] = 1$, then we obtain this analog of (4.7):

$$B_1(x_0, A) \leq E_{x_0}^{\Pi}[e^{\gamma\sum_{i=0}^{v_n(A)} c(x_i,\Pi(x_i))-\gamma\lambda\sum_{i=0}^{v_n(A)} t(x_i,\Pi(x_i))}]. \tag{4.11}$$

Note: we need $P_{x_0}^{\Pi}[\tau_A < \infty] = 1$ in order to insure that $v_n(A) < \infty$ w.p.1, which causes the right hand side of (4.10) to be well defined.

Let $C = \{x|E[e^{\gamma\{c(x,\Pi^*(x))-\lambda t(x,\Pi^*(x))\}}] \leq 1$. By Assumption 4.2.3, C has finitely many elements. Therefore, (4.11) holds for $C = A$.

Let $r_A(i) = \max[t \leq i|s(t) \in A]$ with $r_A(i)$ defined to be -1 if the system has never taken a state in A. We see then that

$$E_{x_0}^{\Pi}[e^{\gamma\sum_{i=0}^{N} c(x_i,\Pi(x_i))-\gamma\lambda\sum_{i=0}^{N} t(x_i,\Pi(x_i))}]$$

$$= E_{x_0}^{\Pi}[e^{\gamma\sum_{i=0}^{r_C(N)} c(x_i,\Pi(x_i))-\gamma\lambda\sum_{i=0}^{r_C(N))} t(x_i,\Pi(x_i))}]+$$

$$E_{x_0}^{\Pi}[e^{\gamma\sum_{i=r_C(N)+1}^{N} c(x_i,\Pi(x_i))-\gamma\lambda\sum_{i=r_C(N)+1}^{N} t(x_i,\Pi(x_i))}].$$

The first term on the right hand side is bounded below by $B_1(x_0, C)$ and the second term is bounded below by 1 by the definition of C since the states through which the system evolves in the second summation are in C^c. Therefore, we have

$$B_1(x_0, C) \leq E_{x_0}^{\Pi}[e^{\gamma \sum_{i=0}^{N} c(x_i,\Pi(x_i)) - \gamma\lambda \sum_{i=0}^{N} t(x_i,\Pi(x_i))}]. \tag{4.12}$$

(4.12) and (4.9) give us the behavior of Π^* at at transition times and show that no policy does better than Π^* at transition times.

As in the proof of Theorem 4.2.1, let $N(t)$ be the number of transitions that have occured prior to time t. Therefore, we have that

$$t_{N(t)} \leq t \leq t_{N(t)+1}$$

Similarly, we have

$$\begin{aligned} &E_{x_0}^{\Pi}[e^{\gamma \sum_{i=0}^{N(T)} c(x_i,\Pi(x_i)) - \gamma\lambda \sum_{i=0}^{N(T)+1} t(x_i,\Pi(x_i))}] \\ &\leq E_{x_0}^{\Pi}[e^{\gamma\{\int_{t=0}^{T} g(t)dt - \lambda T\}}] \\ &\leq E_{x_0}^{\Pi}[e^{\gamma\{\sum_{i=0}^{N(T)+1} c(x_i,\Pi(x_i)) - \lambda \sum_{i=0}^{N(T)} t(x_i,\Pi(x_i))\}}] \end{aligned} \tag{4.13}$$

for any policy Π.

And so by assumption A1.1 and (4.12), we see that

$$L \cdot B_1(x_0) < E_{x_0}^{\Pi^*}[e^{\gamma\{\int_{t=0}^{T} g(t)dt - \lambda T\}}];\ \forall T.$$

Therefore,

$$L \cdot B_1(x_0, C) < \frac{E_{x_0}^{\Pi}[e^{\gamma \int_{t=0}^{T} g(t)dt}]}{e^{\gamma\lambda T}};\ \forall T.$$

for all policies Π.

Now, we have to bound the cost function above under Π^*. By (4.9), we know that

$$E_{x_0}^{\Pi^*}[e^{\gamma \sum_{i=0}^{N(T)+1} c(x_i,\Pi^*(x_i)) - \gamma\lambda \sum_{i=0}^{N(t)+1} t(x_i,\Pi^*(x_i))}] \leq B_2(x_0). \tag{4.14}$$

Let

$$E_T = e^{\gamma \sum_{i=0}^{N(T)} c(x_i,\Pi^*(x_i)) - \gamma\lambda \sum_{i=0}^{N(t)} t(x_i,\Pi^*(x_i))}$$

By (4.14), we see that

$$E_{x_0}^{\Pi^*}[E_T e^{\gamma c(x_{N(t)+1},\Pi^*(x_{N(t)+1}) - \gamma\lambda t(x_{N(t)+1},\Pi^*(x_{N(t)+1})}] \leq B_2(x_0).$$

Now, let us examine the behavior of

$$F_T = E_{x_0}^{\Pi^*}[E_T e^{\gamma c(x_{N(t)+1},\Pi^*(x_{N(t)+1})}].$$

We see that

$$F_T = E_{x_0}^{\Pi^*}[E_T E[e^{\gamma c(x_{N(t)+1},\Pi^*(x_{N(t)+1})}|E_T]$$

$$= E_{x_0}^{\Pi^*}[E_T \sum_{s\in S} P[x_{N(t)+1} = s|E_T]E[e^{\gamma c(s,\Pi^*(s))}]]$$

$$\leq E_{x_0}^{\Pi^*}[E_T \sum_{s\in S} P[x_{N(t)+1} = s|E_T]\kappa E[e^{\gamma c(s,\Pi^*(s))] - \gamma\lambda t(s,\Pi^*(s))}]$$

$$= \kappa E_{T+1} \leq B_2(x_0)$$

where the last inequality follows from Lemma 4.2.2.

Combining this with (4.13) gives us that

$$E_{x_0}^{\Pi}[e^{\gamma\{\int_{t=0}^{T} g(t)dt - \lambda T\}}] \leq \kappa B_2(x_0).$$

By taking logs and limits, etc., we see that $J^{\Pi^*}(x_0) = \lambda \ \forall x_0$ since x_0 was arbitrary. And also $J_{x_0}^{\Pi} \geq \lambda \ \forall x_0, \Pi$.

□

So the dynamic program (4.2) can be used to find an optimal policy. (4.2) is often referred to as the *optimality equation.* There is also an optimality inequality:

$$e^{W(x)} \geq \min_{a \in \alpha(x)} E[e^{\gamma\{c(x,a)-\lambda t(x,a)\}}] \int e^{W(y)} P(dy|x,a); \forall x \in S \tag{4.15}$$

The optimality inequality does not guarantee optimality of the policy it defines by minimizing its right hand side, but it does provide an upper bound on performance as the following corollary to Theorem 4.2.2 demonstrates.

Corollary 4.2.1 *Suppose that $\{W(x), \lambda\}$ is a solution to the optimality inequality (4.15) and that $\{W(x)\}$ is finite for each x and bounded below. Suppose furthermore that Assumption 4.2.1 and Assumption 4.2.3 hold. Let Π^* be a stationary, Markov, deterministic policy such that $\Pi^*(x)$ minimizes the right hand side of (4.15) for each $x \in S$.*

Then

$$J^{\Pi^*}(x_0) \leq \lambda \ \forall x_0. \tag{4.16}$$

In Chapter 5, we will find conditions under which the policy defined by the optimality inequality is optimal and has optimal cost λ, i.e., (4.16) holds with equality.

Lemma 4.2.3 *Under policy Π^* as defined in the statement of Theorem 4.2.2, $\exists B \subset \Omega^a \cap C$ such that $P_x^{\Pi}[\tau_B < \infty] = 1$.*

Proof:

Claim:

$P_{x_0}^{\Pi^*}[\eta_C = \infty] = 1 \ \forall x_0 \in S.$

Proof of claim:

Let $C_C^{min} = \inf_{x \in C} E[e^{\gamma\{c(x,\Pi^*(x))-\lambda t(x,\Pi^*(x))\}}]$. Since C has finitely many elements, the infimum is achieved. Thus, $C_C^{min} > 0$. Let

$$C_{C^c}^{min} = \inf_{x \in C^c} E[e^{\gamma\{c(x,\Pi^*(x))-\lambda t(x,\Pi^*(x))\}}].$$

Because the cost is norm-like (Assumption 4.2.3), there are finitely many states with cost less than M for any $M < \infty$. Therefore, the infimum is achieved. By definition of C, $C_{C^c}^{min} > 1$.

We have

$$E_{x_0}^{\Pi^*}[e^{\gamma \sum_{i=0}^{N} c(x_i,\Pi^*(x_i)) - \gamma\lambda \sum_{i=0}^{N} t(x_i,\Pi^*(x_i))}]$$

$$\geq E_{x_0}^{\Pi^*}[(C_C^{min})^{\sum_{i=0}^{N} I(x_i \in C)} \cdot (C_{C^c}^{min})^{\sum_{i=0}^{N} I(x_i \in C^c)}]$$

(4.9) implies that $\forall x_0 \in S$,

$$E_{x_0}^{\Pi^*}[(C_{C^c}^{min})^N (\frac{C_C^{min}}{C_{C^c}^{min}})^{\sum_{i=0}^{N} I(x_i \in C)}$$

$$= E_{x_0}^{\Pi^*}[(C_C^{min})^{\sum_{i=0}^{N} I(x_i \in C)} \cdot (C_{C^c}^{min})^{\sum_{i=0}^{N} I(x_i \in C^c)}] \leq B_2(x_0); \forall N.$$

or,

$$E_{x_0}^{\Pi^*}[(\frac{C_C^{min}}{C_{C^c}^{min}})^{\sum_{i=0}^{N} I(x_i \in C)}] \leq B_2(x_0)(C_{C^c}^{min})^{-N}; \forall N.$$

For N large enough, $B_2(x_0)(C_{C^c}^{min})^{-N}$ is arbitrarily small. We see that

$$P_{x_0}^{\Pi^*}[\sum_{i=0}^{N} I(x_i \in C) \leq \frac{N}{N-1} \frac{log(B_2(x_0)(C_{C^c}^{min})^{-N})}{log(\frac{C_C^{min}}{C_{C^c}^{min}})}] \leq \frac{N-1}{N}$$

And so, $P_{x_0}^{\Pi}[\eta_C = \infty] = 1$.

And the claim is proved.

The Lemma follows from Lemma 2.2.3 . □

4.3 Foundational assumptions for existence of optimal policies

Consider an SMDP with countable state space such that the set of all actions admissible in state x, $\alpha(x)$, is compact for all states x. The following theorem is found in [36]:

Theorem 4.3.1 (Tychonoff's Theorem) *Let $\{S_i\}$ for $i = 0, 1, 2, ...,$ denote a collection of compact sets. Then $S = \times_{i=0}^{\infty} S_i$ is compact.*

We therefore see that the set of all stationary, Markov, deterministic policies is compact.

Recall Assumption 3.4.1, taken from [35], which guarantees that there will be a finite number of transitions in any finite time interval. We rewrite it here with $t(x, a)$ in place of δt, to clarify its meaning in our context.

Assumption 3.4.1 (Restated) *There exist $\epsilon > 0$ and $\delta > 0$ such that*

$$P[t(x, a) \leq \delta] \leq 1 - \epsilon$$

$\forall x \in S$ and $a \in \alpha(x)$.

An immediate consequence of Assumption 3.4.1 is

$$E[e^{-\lambda t(x,a)}] \leq U_\lambda \doteq (1 - \epsilon) + \epsilon e^{-\lambda \delta};\ \forall x, a.$$

Assumption 4.3.1 *There exist $\epsilon^{'} > 0$ and $\delta^{'} < \infty$ such that*

$$P[t(x, a) \geq \delta^{'}] \leq \epsilon^{'}$$

$\forall x \in S$ and $a \in \alpha(x)$.

An immediate consequence of Assumption 4.3.1 is

$$E[e^{-\lambda t(x,a)}] \geq L_\lambda \doteq \epsilon^{'} e^{-\lambda \delta^{'}};\ \forall x, a.$$

In fact, Assumption 4.2.1 is equivalent to Assumption 4.3.1.

Assumption 4.3.2 $\lim_{x \to \infty} \min_{a \in \alpha(x)} E[c(x, a)] = \infty$.

Note: Assumption 4.3.2 is equivalent to Assumption 4.2.3 given that Assumption 4.3.1 is true.

Assumption 4.3.3 *$r(x)$, the set of all states reachable from x in one transition, is finite for each $x \in S$.*

The following Lemma assures us that we need only consider policies that induce a positive recurrent subclass.

Lemma 4.3.1 *Let $\Pi \in \Pi^{HR}$ be a stationary policy. If $P_x^\Pi[\tau_{M_\Pi} < \infty] < 1$, then*

$$\lim_{T\to\infty} \frac{1}{\gamma T} \ln E_x^\Pi[e^{\gamma \int_{t=0}^T g(t)dt}] = \infty;\ \forall x \in S.$$

Proof:

For $0 < \lambda < \infty$, define $A_\lambda = \{x | E[e^{\gamma\{c(x,\Pi(x))-\lambda t(x,\Pi(x))\}}] \leq 2\}$. By Assumption 4.3.2, A_λ has finitely many elements $\forall\lambda$. By Lemma 2.2.2, $\lim_{k\to\infty} P_x^\Pi[x_k \in A_\lambda^c \cup M_\Pi] = 1$, we see that $P_x^\Pi[\exists N < \infty | \forall k > N, x_k \in A_\lambda^c] \geq 1 - P_x^\Pi[\tau_{M_\Pi} < \infty] > p > 0$. Therefore, $P_x^\Pi[\lim_{k\to\infty} e^{\gamma \sum_{i=0}^N c(x_i,\Pi(x_i)) - \gamma\lambda \sum_{i=0}^N t(x_i,\Pi(x_i))}] \geq 1.5^N] > p > 0$ since $1.5 < 2$.

Therefore, $\lim_{N\to\infty} E_x^\Pi[e^{\gamma \sum_{i=0}^N c(x_i,\Pi(x_i)) - \gamma\lambda \sum_{i=0}^N t(x_i,\Pi(x_i))}] \geq 1$, and by the reasoning contained in the proofs of Theorems 1 and 2, the average cost is at least λ. Since λ was arbitrary, the Lemma is proved. □

Corollary 4.3.1 *Let $\Pi \in \Pi^{HR}$ be a stationary policy. If $P_x^\Pi[\tau_{M_\Pi} < \infty] < 1$, then*

$$\lim_{T\to\infty} \frac{1}{T} E_x^\Pi[\int_{t=0}^T g(t)dt] = \infty;\ \forall x \in S.$$

Chapter 5

Perron-Frobenius Eigenvalue

The main thrust of this chapter is to define the Perron-Frobenius eigenvalue and the round trip cost. As will be explained in Section 11.2, the difficulty of establishing the value of a policy lies in the fact that the state space is countable. By reducing the problem to a finite one, viz. the round trip cost, we reduce the problem to one that is tractable. This chapter focuses on the behavior within a positive recurrent class of an uncontrolled Markov chain. The results here will be used later on in establishing results for controlled Markov chains where the irreducibility assumption has been removed.

5.1 Defining the Perron-Frobenius Eigenvalue and the Round Trip Cost

In [2], a kernel was defined for use in determining the average cost in a risk sensitive MDP. Here, we adapt the kernel to the semi-Markov setting. Denote the kernel defined for $x, y \in S$ by

$$\hat{P}^{\Pi}_{\gamma,\lambda}(x,y) = E[e^{\gamma(c(x,\Pi(x))-\lambda t(x,\Pi(x))}]P(y|x,\Pi(x)).$$

Again adapting from [2], define the Perron Frobenius eigenvalue (*pfe*) $\lambda^{\Pi}_{C}(\gamma)$

for a policy Π that induces a positive recurrent subclass $C \subset S$ as

$$\lambda_C^\Pi(\gamma) \doteq \inf(\lambda \in \Re | \sum_{k=0}^{\infty} \hat{P}_{\gamma,\lambda}^{\Pi}(\theta,\theta) < \infty);\ \gamma > 0$$

for θ in a positive recurrent subclass C of Π. (Note that choice of θ is arbitrary, and the value of $\lambda_C^\Pi(\gamma)$ is the same for any $\theta \in C$. (This will become evident later on when it is proved that the optimal long term average cost starting from anywhere in C is given by $\lambda_C^\Pi(\gamma)$.) We set $\lambda_C^\Pi(\gamma) = \infty$ if the above infimum is over a null set. Equivalently, (see [2] and references therein)

$$\lambda_C^\Pi(\gamma) = \inf(\lambda \in \Re | E_\theta^\Pi[e^{\gamma \sum_{k=0}^{\tau_\theta - 1}\{c(x_k,\Pi(x_k)) - \lambda t(x_k,\Pi(x_k))\}} I(\tau_\theta < \infty)] \leq 1);\ \gamma > 0. \tag{5.1}$$

Define $D_C^\Pi(\lambda) = \{\gamma | \lambda_C^\Pi(\gamma) < \infty\}$. Also define $\bar{\gamma}_C^\Pi \doteq \sup(\gamma | \lambda_C^\Pi(\gamma) < \infty)$. If $\bar{\gamma}_C^\Pi < \infty$, then by Fatou's Lemma ([2]), we have that $D_C^\Pi = (-\infty, \bar{\gamma}_C^\Pi)$.

Note that the above is all defined with respect to θ and its positive recurrent subclass. Π may induce more than one positive recurrent subclass, and these results apply to each subclass separately.

Define

$$C^{\theta\to\theta}(\lambda) \doteq E_\theta^\Pi[e^{\gamma \sum_{k=0}^{\tau_\theta - 1}\{c(x_k,\Pi(x_k)) - \lambda t(x_k,\Pi(x_k))\}} I(\tau_\theta < \infty)].$$

We know by (5.1) that $\lambda_C^\Pi(\gamma) = \inf(\lambda \in \Re | C^{\theta\to\theta}(\lambda) \leq 1)$. Therefore, $C^{\theta\to\theta}(\lambda_C^\Pi(\gamma)) \leq 1$ and because $\frac{d}{d\lambda} C^{\theta\to\theta}(\lambda) < 0$, $C^{\theta\to\theta}(\lambda) < 1$ for $\lambda > \lambda_C^\Pi(\gamma)$.

But what is the behavior of $C^{\theta\to\theta}(\lambda)$ for $\lambda \leq \lambda_C^\Pi(\gamma)$?

Lemma 5.1.1 *If $C^{\theta\to\theta}(\lambda_C^\Pi(\gamma)) < 1$, then $C^{\theta\to\theta}(\lambda) = \infty\ \forall \lambda < \lambda_C^\Pi(\gamma)$.*

Proof:

Because the embedded Markov Chain induced by Π on C is recurrent, we know that $I(\tau_\theta < \infty) = 1$ w.p.1. Therefore

$$C^{\theta\to\theta}(\lambda) = E_\theta^\Pi[e^{\gamma \sum_{k=0}^{\tau_\theta - 1}\{c(x_k,\Pi(x_k)) - \lambda t(x_k,\Pi(x_k))\}} I(\tau_\theta < \infty)]$$

$$= E_\theta^\Pi[e^{\gamma \sum_{k=0}^{\tau_\theta - 1}\{c(x_k,\Pi(x_k)) - \lambda t(x_k,\Pi(x_k))\}}]$$

$$= E_\theta^\Pi[e^{\gamma \int_{t=0}^{T_\theta}\{g(t)-\lambda\}dt}]$$

$$= E_\theta^\Pi[e^{\gamma\{C(T_\theta)-\lambda T_\theta\}}],$$

where $C(T) \doteq \int_{t=0}^{T} g(t)dt$. Let the cumulative distribution function of T_θ be denoted $F_{T_\theta}(\tau) = P_\theta^\Pi[T_\theta \le \tau]$. We then get

$$C^{\theta\to\theta}(\lambda) = \int_{t=0}^{\infty} E_\theta^\Pi[e^{\gamma C(t)}|T_\theta = t]e^{-\gamma\lambda t}dF_{T_\theta}(t).$$

And therefore,

$$\frac{d}{d\lambda}C^{\theta\to\theta}(\lambda) = \int_{t=0}^{\infty} -\gamma E_\theta^\Pi[e^{\gamma C(t)}|T_\theta = t]te^{-\gamma\lambda t}dF_{T_\theta}(t),$$

which is negative and decreasing in λ.

Suppose that $C^{\theta\to\theta}(\lambda_C^\Pi(\gamma)) < 1$. Because $\frac{d}{d\lambda}C^{\theta\to\theta}(\lambda)$ is negative and decreasing in λ, we know that $\frac{d}{d\lambda}C^{\theta\to\theta}(\lambda) = -\infty$ for $\lambda \le \lambda_C^\Pi(\gamma)$. Therefore $C^{\theta\to\theta}(\lambda_C^\Pi(\gamma)) = \infty$ for $\lambda < \lambda_C^\Pi(\gamma)$.

□

Lemma 5.1.1 illustrates the fact that $C^{\theta\to\theta}(\lambda)$ is decreasing in λ, and that its rate of decrease is decreasing. Let us call $C^{\theta\to\theta}(\lambda)$ the *round trip cost* for θ at λ. *The reason we are concerned with the round trip cost is because it allows us to reduce an infinite problem (the long term average cost) to a finite problem (the cost to return to a state).* It is this property that makes the round trip cost so important, and so interesting.

Before we proceed with our development, let us pause for an explanation and a look ahead: The behavior of a Markov chain in a positive recurrent class can be classified by whether the round trip cost at $\lambda_C^{\Pi}(\gamma)$ is 1 or whether it is less than one. (Fact: this classification is the same for all $\theta \in C$.) We will show that the round trip cost is continuous from the right, decreasing, and has at most one point of discontinuity: the point where it jumps to ∞. Therefore the value of 1 is achieved if any finite value greater than or equal to 1 is achieved before the jump to infinity (if there is a jump to infinity.) *Now, let's continue with the development:*

The round trip cost at $\lambda = 0$ is either finite or infinite. If it is finite, then clearly it is greater than one since each state transition has positive cost. Therefore, if $C^{\theta \to \theta}(0) < \infty$, then the value $C^{\theta \to \theta}(\lambda) = 1$ is achieved since $C^{\theta \to \theta}(\lambda)$ is a smooth, decreasing function of λ. (See figure 5.1.)

If $C^{\theta \to \theta}(0) = \infty$, then because it is a decreasing function of λ, there is a value λ_i such that $C^{\theta \to \theta}(\lambda) = \infty$ for $\lambda < \lambda_i$ and $C^{\theta \to \theta}(\lambda) < \infty$ for $\lambda > \lambda_i$. If $C^{\theta \to \theta}(\lambda_i) = \infty$, then we say the semi-Markov chain is *Type I*, and if $C^{\theta \to \theta}(\lambda_i) < \infty$, we say the semi-Markov chain is *Type II*.

Figures 5.2 and 5.3 show the round trip cost as a function of λ for a Type I and a Type II chain, respectively.

Lemma 5.1.1 (which corresponds to figure 5.4) covers the case of a semi-Markov chain with $\lambda_i = \lambda_C^{\Pi}(\gamma)$; that is the case in which $C^{\theta \to \theta}(\lambda_C^{\Pi}(\gamma)) < 1$. In all other cases, (i.e., the cases shown in figures 5.1, 5.2 and 5.5,) the value $C^{\theta \to \theta}(\lambda) = 1$ will be achieved. Note: it is also possible that $\lambda_i = \lambda_C^{\Pi}(\gamma)$ and $C^{\theta \to \theta}(\lambda_C^{\Pi}(\gamma)) = 1$. This would correspond to figure 5.3 in which the round trip cost at λ_i is 1.

When does a semi-Markov chain exhibit Type II behavior? In order to answer that question, we will look at some examples and solve a matrix equation for $C^{\theta \to \theta}(\lambda)$.

In figures 5.6, 5.7, and 5.8, transition probabilities refer to the probability of the transition being made in a round trip, not the probability of the transition being

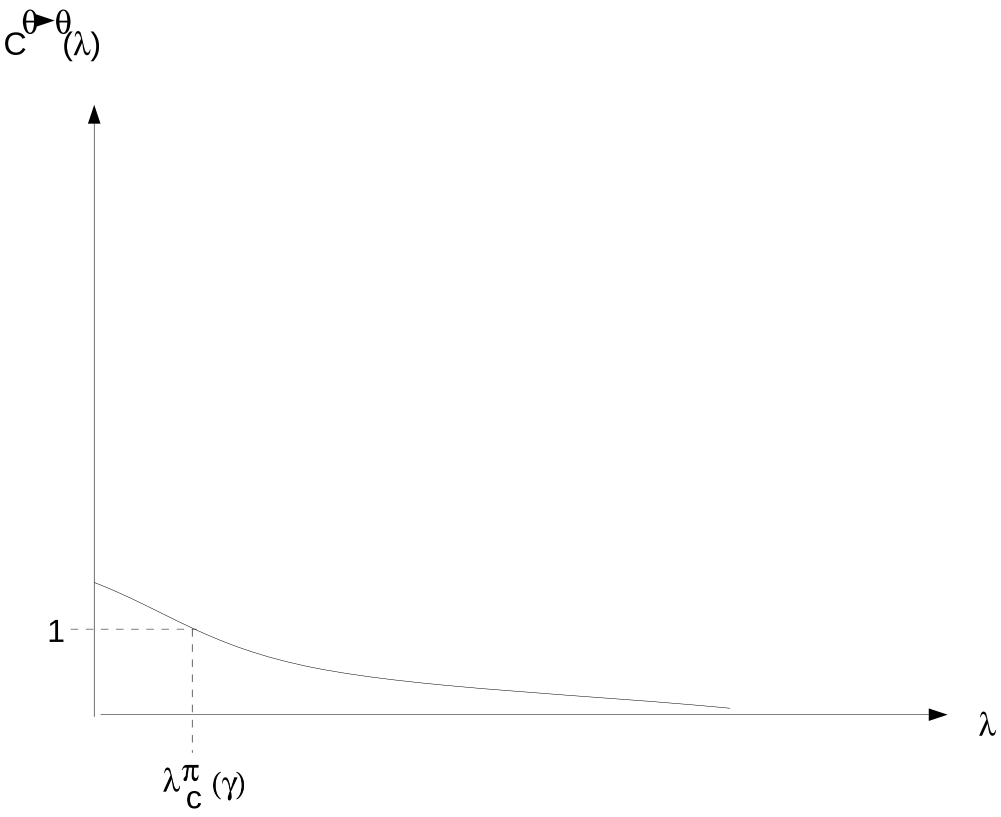

Figure 5.1: Type I semi-Markov chain – round trip cost is finite at $\lambda = 0$.

made given that the current state is the state from which the transition occurs. For example, in figure 5.7 P_5 is the probability that it takes exactly 5 transitions to return to the base state.

The values inside the circles denote the occupancy cost of a state. For example, in figure 5.8, the round trip cost at $\lambda = 0$ is 2 with probability P_1, 6 with probability P_2, 12 with probability P_3, etc..

Let us look at figure 5.6. It is clear from inspection that $C^{\theta\to\theta}(10) = e^{\gamma(5-10)} = e^{-5\gamma}$. This is because $c - \lambda = -5$ at the first step and 0 at each

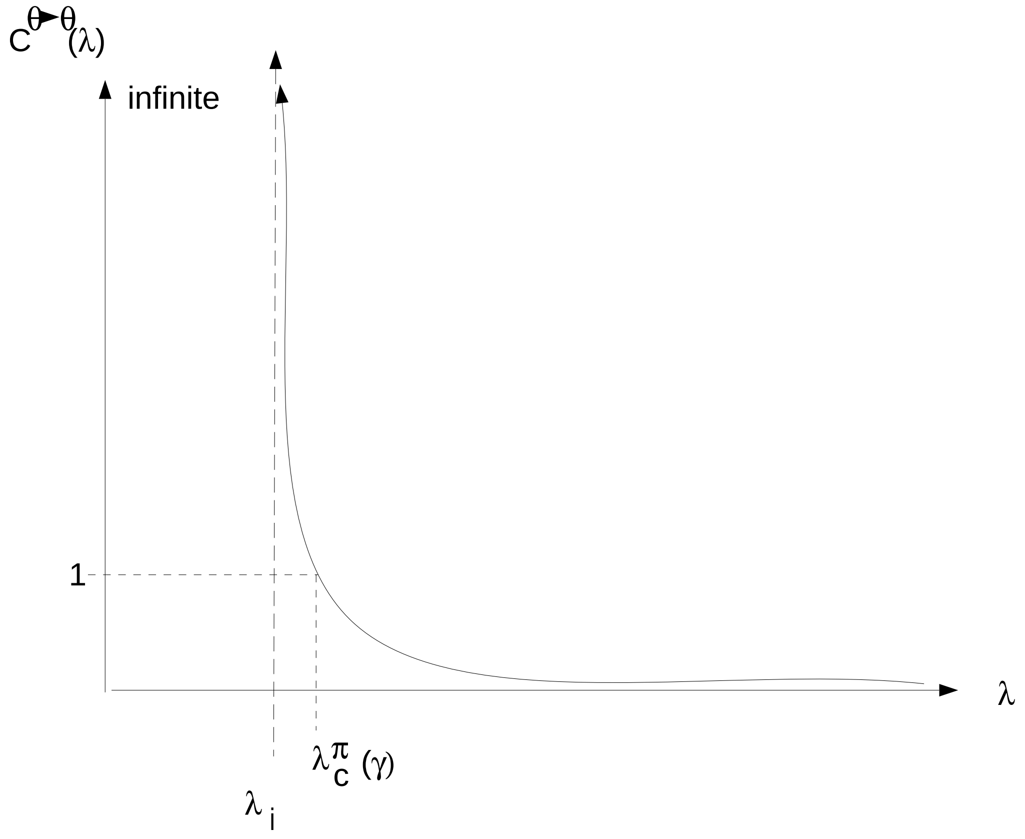

Figure 5.2: Type I semi-Markov chain – round trip cost grows asymptotically.

subsequent step (if there is more than one step) in a round trip. Let $\lambda < 10$ be given and let us determine the round trip cost:

$$C^{\theta\to\theta}(\lambda) = \sum_{k=1}^{\infty} E[e^{\sum_{i=0}^{k-1}\gamma(c(i)-\lambda)}|\tau_\theta = k] \cdot P[\tau_\theta = k]$$

$$= \sum_{k=1}^{\infty} e^{\gamma\{5+10\cdot(k-1)-\lambda\cdot k\}} \cdot P[\tau_\theta = k]$$

$$= e^{-5\gamma} \cdot \{1 - (\frac{1}{4} + \frac{1}{9} + \frac{1}{16} + ...\} + \sum_{k=2}^{\infty} e^{-5\gamma} \cdot e^{(k-2)\gamma(10-\lambda)} \cdot \frac{1}{k^2}$$

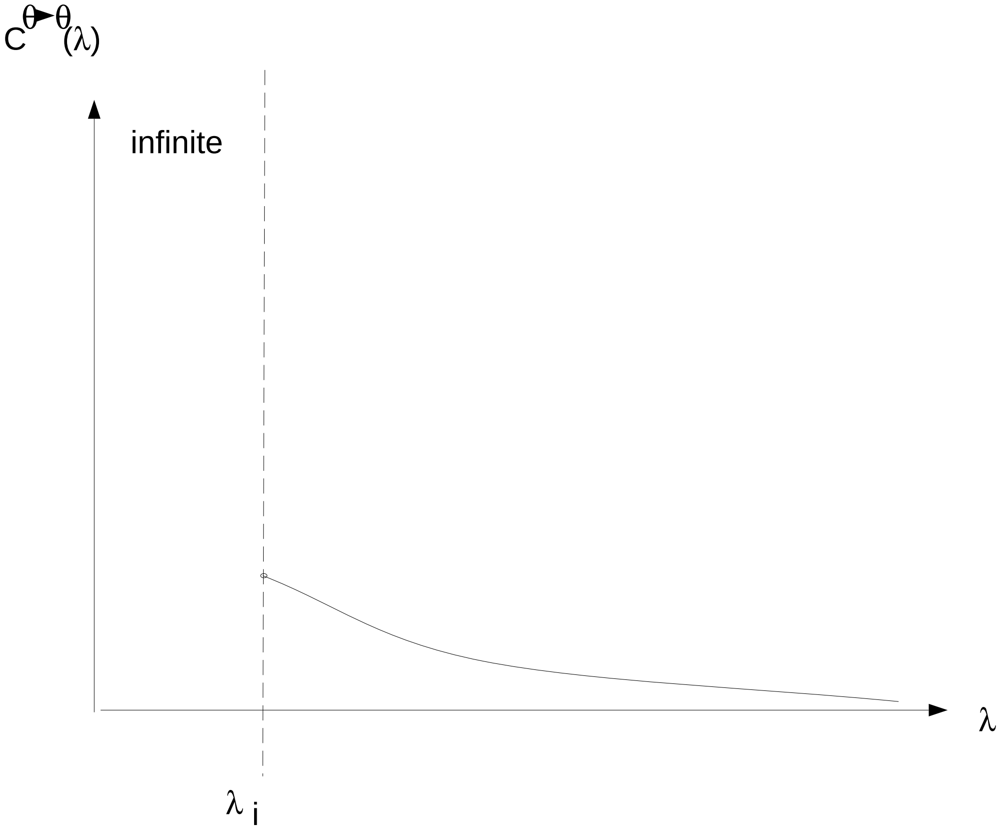

Figure 5.3: Type II semi-Markov chain – round trip cost grows discontinuously. $(C^{\theta\to\theta}(\lambda_i) < \infty)$

$$> e^{-5\gamma} \sum_{k=2}^{\infty} e^{(k-2)\gamma(10-\lambda)} \cdot \frac{1}{k^2},$$

and because the exponential dominates $\frac{1}{k^2}$, we see that $C^{\theta\to\theta}(\lambda) = \infty$ for $\lambda < 10$.

The Markov chain in figure 5.6 is therefore an example of a Markov chain of Type II.

The Markov chain depicted in figure 5.7 has norm-like costs. But we can still get the same behavior for $C^{\theta\to\theta}(\lambda)$ as in figure 5.6 by setting the value of P_i

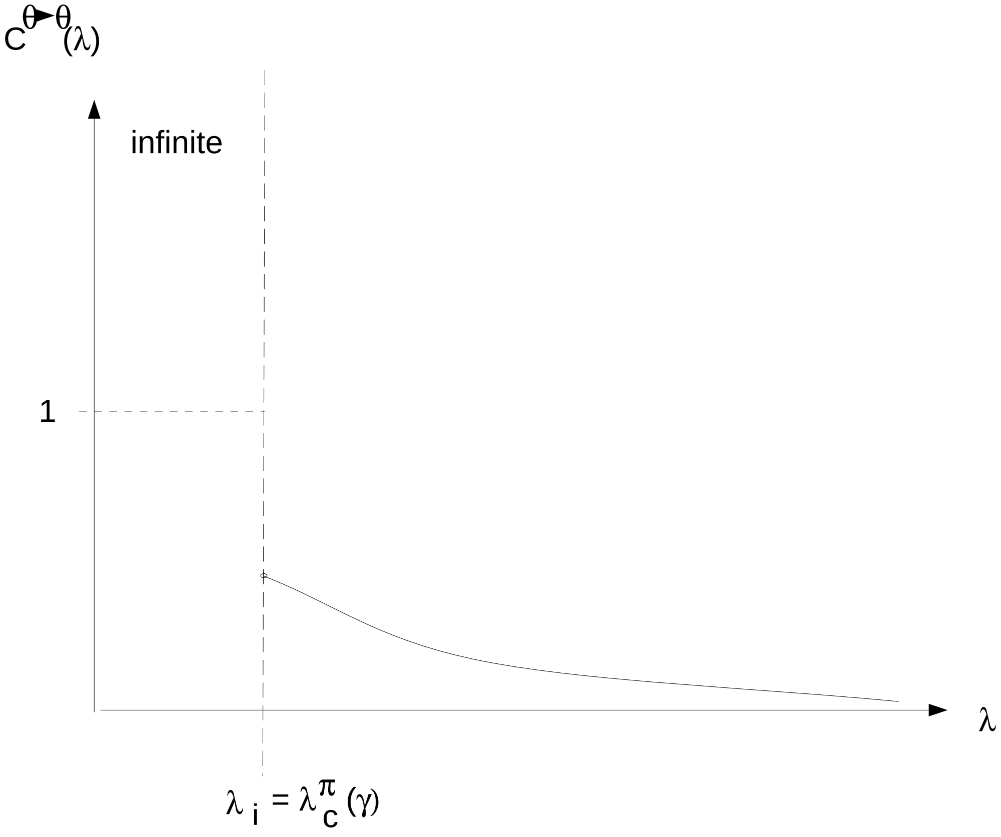

Figure 5.4: Type II semi-Markov chain – ($\lambda_i = \lambda_C^{\Pi}(\gamma)$)

appropriately. If we set $[\sum_{k=1}^{i} k] \cdot P_i \doteq 5 + 10 \cdot (k-1) \cdot I[k > 1] \cdot \frac{1}{k^2}$, then the behavior of $C^{\theta\to\theta}(\lambda)$ is identical in figures 5.6 and 5.7 and the Markov chain in figure 5.7 is therefore of Type II.

The Markov chain depicted in figure 5.8 has a finite number of possible transitions into and out of each state in addition to having norm-like costs. Furthermore, if the value of P_i is the same $\forall i$ in figures 5.7 and 5.8 and the value of the risk sensitivity parameter γ is twice as big in figure 5.7 as in figure 5.8, then the values of $C^{\theta\to\theta}(\lambda)$ are identical in each figure. This is because the Markov chain in figure 5.8 has an equal probability of having a round trip of twice the

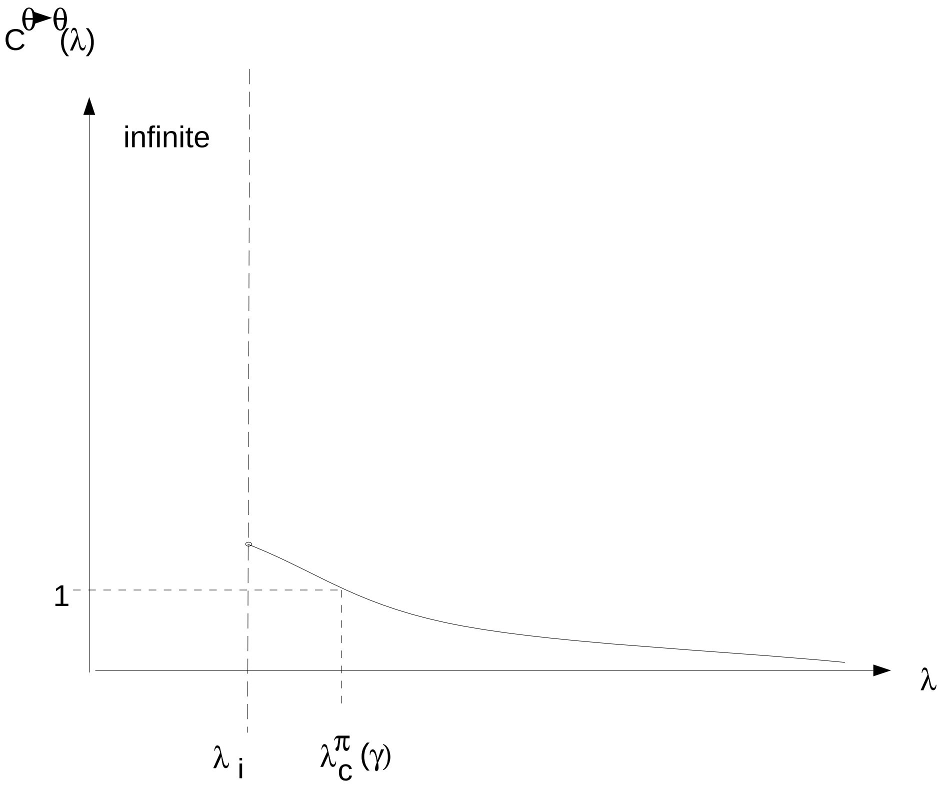

Figure 5.5: Type II semi-Markov chain – $(\lambda_i < \lambda_C^{\Pi}(\gamma))$

length and twice the cost as the Markov chain in figure 5.7. Therefore, if we set $[\sum_{k=1}^{i} k] \cdot P_i \doteq 5 + 10 \cdot (k-1) \cdot I[k > 1] \cdot \frac{1}{k^2}$, then the Markov chain of figure 5.8 is of Type II.

5.2 Round trip cost when the state space is finite

We have seen examples in which Markov chains with a countable state space exhibit behavior of Type II. Let us examine what happens in a Markov chain with finite state space. Because the state space is finite, we can derive a matrix formula for $C^{\theta \to \theta}(\lambda)$. In order to do so, we will have to introduce some notation. In this subsection a vector will be denoted with a ˆ above; a matrix will be denoted with

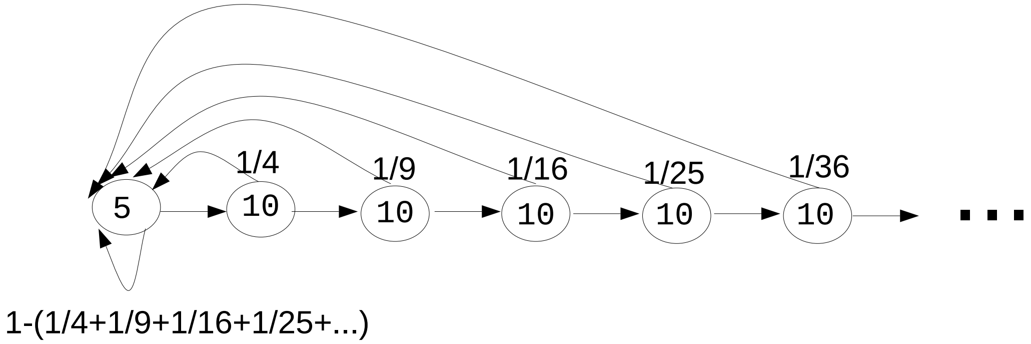

Figure 5.6: Example of Type II Markov Chain.

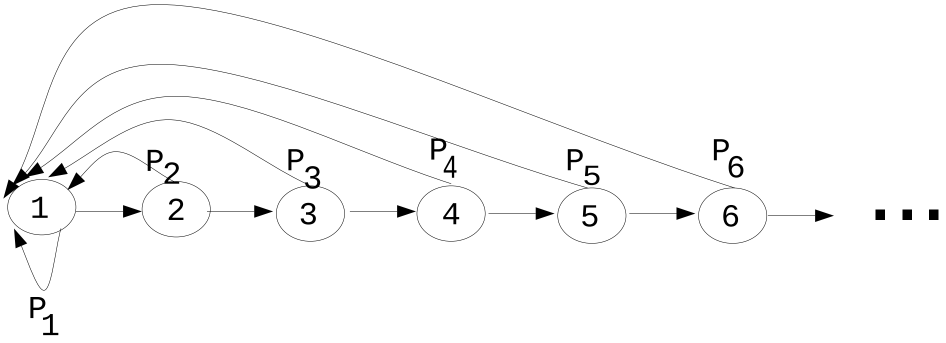

Figure 5.7: A Markov Chain with norm-like costs.

a $^-$ above, and a scalar will have neither above. I.e., $\hat{x}$ is a vector; $\bar{x}$ is a matrix, and x is a scalar. Define the vector operator $\odot$ as follows: $\hat{c} = \hat{a} \odot \hat{b}$ if $\hat{c}_i = \hat{a}_i \cdot \hat{b}_i$, i.e., $\odot$ represents element-wise multiplication. Let an irreducible, time-invariant semi-Markov chain be given with $n+1$ states, labeled θ and $\{z_1, ..., z_n\}$.

Denote $w(\lambda) = C^{\theta \to \theta}(\lambda)$; $p^\theta = P[x_{k+1} = \theta | x_k = \theta]$; and $\hat{p}_i = P[x_{k+1} = z_i | x_k = \theta]$, i.e., the i^{th} element of $\hat{p}$ is $P[x_{k+1} = z_i | x_k = \theta]$.

Denote $\hat{W}_i(\lambda) = E_{z_i}[e^{\gamma \sum_{k=0}^{\tau_\theta - 1} c(x_k) - \lambda t(x_k)}]$; $\hat{P}_i^\theta = P[x_{k+1} = \theta | x_k = z_i]$; and $\bar{P}_{ij} = P[x_{k+1} = z_j | x_k = z_i]$.

We denote the transition costs as $\Delta c(\lambda) = E[e^{\gamma\{c(\theta) - \lambda t(\theta)\}}]$ and $\hat{\Delta C}_i(\lambda) = E[e^{\gamma\{c(z_i) - \lambda t(z_i)\}}]$. Define the diagonal matrix $CC(\lambda)$ such that

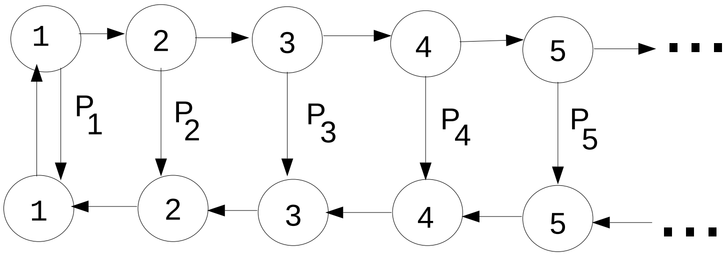

Figure 5.8: A Markov Chain with norm-like costs and finite transitions into and out of each state.

$$CC_{ij}(\lambda) = \begin{cases} \hat{\Delta C}_i(\lambda) \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}$$

We then obtain the following equations to solve for $w(\lambda)$ and $\hat{W}(\lambda)$:

$$w(\lambda) = \Delta c(\lambda)[p^{\theta} + \hat{p}^T \hat{W}_i(\lambda)]; \tag{5.2}$$

$$\hat{W}(\lambda) = \hat{\Delta C}(\lambda) \odot [\bar{P}\hat{W}(\lambda) + \hat{P}^{\theta}]. \tag{5.3}$$

Equation (5.3) can be rearranged to give

$$\hat{\Delta C}(\lambda) \odot \hat{P}^{\theta} = [I - CC(\lambda)\bar{P}]\hat{W}(\lambda), \tag{5.4}$$

yielding

$$\hat{W}(\lambda) = [I - CC(\lambda)\bar{P}]^{-1}[\hat{\Delta C}(\lambda) \odot \hat{P}^{\theta}], \tag{5.5}$$

if $[I - CC(\lambda)\bar{P}]$ is nonsingular.

Equation (5.2) then gives us the value of $w(\lambda)$. For a given value of λ, in order for the round trip cost $w(\lambda)$ to be between 0 and ∞, we need for the solution $\hat{W}(\lambda)$ to equation (5.4) to exist and be such that $0 < \hat{W}_i(\lambda) < \infty \; \forall i$.

We make the following assumptions:

Assumption 5.2.1 $\forall x \in S,\ 0 < E[e^{\gamma c(x)}] < \infty.$

Assumption 5.2.2 $\forall x \in S,\ E[t(x)] > 0.$

Note: Assumption 5.2.2 is identical to Assumption 3.4.1 restricted to a semi-Markov chain instead of an SMDP.

Under these assumptions, $\Delta c(\lambda)$ and $\hat{\Delta C}(\lambda)$ are smooth functions of λ, and are bounded away from 0 and ∞ for $\lambda \geq 0$. Because the elements of $[I - CC(\lambda)\bar{P}]$ vary continuously as a function of λ, equation (5.4) has a solution for $\hat{W}(\lambda)$ that varies continuously with λ. Furthermore, due to this continuous variation, the following two statements are true:

1: If (5.4) has a solution for $\hat{W}(\lambda)$ for $\lambda = \lambda'$, then (5.4) has a solution for $\hat{W}(\lambda)$ $\forall \lambda$ in an open interval containing λ'.

2: If (5.4) has a solution for $\lambda = \lambda'$ such that $\hat{W}(\lambda') > 0$, then $\exists$ an open interval L containing λ' such that $\hat{W}(\lambda) > 0\ \forall \lambda \in L$.

From the above two statements, it is clear that if $0 < w(\lambda') < \infty$, then there is an open interval L containing λ' such that $0 < w(\lambda) < \infty\ \forall \lambda \in L$. Therefore, the semi-Markov chain either has finite round trip cost at $\lambda = 0$ (as in figure 5.1) or is of Type I (as in figure 5.2.)

The above argument becomes even simpler in the discrete time case because the matrix $[I - CC(\lambda)\bar{P}]$ has all constant entries with $-\lambda$s on the diagonal. Its inverse then has entries which are rational functions of λ, so it has a pole at the largest eigenvalue of $[I - CC(0)\bar{P}]$. This pole causes the round trip cost to grow to infinity asymptotically, making the Markov chain type I if it does not have finite round trip cost at $\lambda = 0$.

In the countable state space case, the same matrix equations ((5.2) and (5.4)) hold, but due to the fact that they are infinite matrices, the argument in the above paragraph no longer holds. We have already shown examples of Markov chains with countable state space that are of Type II. Of course, there are Markov chains (and SMDPs) with countable state space that are of type I. For an SMDP with infinite round trip cost at $\lambda = 0$, there is a simple characteristic of the distribution of the round trip cost that determines whether the SMDP is of type I or type II.

The round trip cost is given by

$$C^{\theta\to\theta}(\lambda) = \int_{t=0}^{\infty} E[e^{\gamma\int_{\tau=0}^{t} g(\tau)d\tau}|T_\theta = t]dP[T_\theta \leq t]e^{-\gamma\lambda t}.$$

If the round trip cost at $\lambda = 0$ is infinite, then

$$E[e^{\gamma\int_{\tau=0}^{t} g(\tau)d\tau}|T_\theta = t]dP[T_\theta \leq t] = e^{\gamma\lambda_i}f(t),$$

where $f(t)$ is a sub-exponential function, i.e.

$$\int_{t=0}^{\infty} f(t)e^{at}dt = \infty \ \forall a > 0.$$

It is easy to show that the SMDP is of Type I if

$$\int_{t=0}^{\infty} f(t)dt = \infty,$$

and Type II if

$$\int_{t=0}^{\infty} f(t)dt < \infty.$$

Furthermore, if the SMDP is of Type II, then $\exists\lambda'$ such that $C^{\theta\to\theta}(\lambda) = 1$ iff

$$\int_{t=0}^{\infty} f(t)dt \geq 1$$

because $C^{\theta\to\theta}(\lambda)$ is a continuous and decreasing function of λ from the right. (If the SMDP is of Type I, then there is also such a λ'. Finally, if the round trip cost at $\lambda = 0$ is finite, then it is also greater than one because costs are positive. Therefore, $\exists\lambda'$ such that $C^{\theta\to\theta}(\lambda) = 1$.

In summary, the only condition under which $\nexists\lambda'$ such that $C^{\theta\to\theta}(\lambda') = 1$ is when the positive recurrent class induced by Π has countably many elements, the SMDP is of Type II, and the round trip cost at λ_i is strictly less than 1, as in figure 5.4.

This is important because when there is such a λ', it is the Perron-Frobenius eigenvalue and the 'nice' recursive equation (4.2) (with only one admissible action per state – this is an uncontrolled Markov chain) holds.

5.3 Average cost on the infinite horizon

Now that we have explored the behavior of the round trip cost, the basis for determination of average cost, we can proceed with analysis of the average cost.

Lemma 5.3.1 is proved by Balaji and Meyn in [2] (Proposition 3.3 on page 9) for the discrete time case. Their proof is built on the foundation of Kingman's subadditive ergodic theorem. ([24] and [25]). However, Kingman's subadditive ergodic theorem does not apply to the semi-Markov case. As he says ([24], P. 499): "In this paper T will be taken as the set of non-negative integers, although *interesting problems arise* when T is in the interval $(0, \infty)$." [Emphasis added] Furthermore, Kingman discusses the continuous parameter process in [25] and explains why his ergodic theorem no longer applies in that case.

In the following lemma, we state and prove the semi-Markov case without use of Kingman's theorem. We do this by using separate techniques to bound the limit above and below, to the same value.
Note: unless stated otherwise, Π is a stationary, Markov policy and C is a positive recurrent class induced by Π.

Lemma 5.3.1 *If $x \in C$ and $\gamma < \bar{\gamma}_C^{\Pi}$, then $J_x^{\Pi} = \lambda_C^{\Pi}(\gamma)$.*

Proof:

Choose $\theta \in C$. Define

$$V(x) = E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_\theta - 1}\{c(x_k,a_k) - \lambda_C^{\Pi}(\gamma)t(x_k,a_k)\}}],\ x \in C.$$

By Fatou's Lemma, $V(\theta) \leq 1$. If $V(\theta) = 1$, then $V(\cdot)$ satisfies the following recursive equation:

$$V(x) = E[e^{\gamma\{c(x,\Pi(x)) - \lambda_C^{\Pi}(\gamma)t(x,\Pi(x))\}}] \cdot \sum_{z \in r(x,\Pi(x))} P(z|x,\Pi(x))V(z). \tag{5.6}$$

If $V(\theta) \leq 1$, then $V(\cdot)$ satisfies

$$V(x) = E[e^{\gamma\{c(x,\Pi(x)) - \lambda_C^{\Pi}(\gamma)t(x,\Pi(x))\}}] \cdot \sum_{z \in r(x,\Pi(x))} P(z|x,\Pi(x))max\{V(z), I(\theta)\}. \tag{5.7}$$

Claim:

$V(x)$ is bounded away from zero (I.e., $\ln V(x)$ is bounded below) on C.

Suppose that $V(\cdot)$ is not bounded away from zero on C.

Define $CC = \{x|E[e^{\gamma\{c(x,\Pi(x)) - \lambda_C^{\Pi}(\gamma)t(x,\Pi(x))\}}] \leq 1$. We know by (5.7) that if $x \in C - CC$, then $\exists y \in C$ such that $V(y) < V(x)$. Therefore, $\inf_{x \in C} V(x) = \inf_{x \in CC} V(x)$. And since CC has finitely many elements by Assumption 4.3.2, we see that the infimum is achieved. Since $V(\cdot)$ is not bounded below, there must be a $z \in CC$ such that $V(z) = 0$.

Let $ZZ = \{x \in C|V(x) = 0\}$. By (5.7), if $x \in ZZ$, then $\forall y \in r(x, \Pi(x))$, $max\{V(y), I(\theta)\} = 0$. This imples that $y \in ZZ$ and that $y \neq \theta$.

Since the embedded Markov Chain induced by Π on C is communicating (recurrent implies communicating), we know that θ can be reached from z. But this is a contradiction, and the claim is proved.

Claim:

$V(x) < \infty\ \forall x \in C$.

Proof of claim:

Suppose $\exists y \in C$ such that $V(y) = \infty$. Because the embedded Markov chain induced by Π on C is recurrent, $\exists h < \infty$ and $p > 0$ such that $P_\theta^\Pi[\{x_h = y\} \cap \{\tau_\theta > h\}] = p$. Therefore,

$$V(\theta) = E_\theta^\Pi[e^{\gamma\sum_{k=0}^{\tau_\theta-1}\{c(x_k,a_k)-\lambda_C^\Pi(\gamma)t(x_k,a_k)\}}]$$

$$\geq E_\theta^\Pi[e^{\gamma\sum_{k=0}^{\tau_\theta-1}\{c(x_k,a_k)-\lambda_C^\Pi(\gamma)t(x_k,a_k)\}}|\{x_h = y\} \cap \{\tau_\theta > h\}]P_\theta^\Pi[\{x_h = y\} \cap \{\tau_\theta > h\}]$$

$$= E_\theta^\Pi[e^{\gamma\sum_{k=0}^{\tau_\theta-1}\{c(x_k,a_k)-\lambda_C^\Pi(\gamma)t(x_k,a_k)\}}|\{x_h = y\} \cap \{\tau_\theta > h\}] \cdot p$$

$$= E_\theta^\Pi[e^{\gamma\sum_{k=0}^{h}\{c(x_k,a_k)-\lambda_C^\Pi(\gamma)t(x_k,a_k)\}}|\{x_h = y\} \cap \{\tau_\theta > h\}] \cdot E_y^\Pi[e^{\gamma\sum_{k=0}^{\tau_\theta-1}\{c(x_k,a_k)-\lambda_C^\Pi(\gamma)t(x_k,a_k)\}}] \cdot p$$

$$= E_\theta^\Pi[e^{\gamma\sum_{k=0}^{h-1}\{c(x_k,a_k)-\lambda_C^\Pi(\gamma)t(x_k,a_k)\}}|\{x_h = y\} \cap \{\tau_\theta > h\}] \cdot V(y) \cdot p$$

$$\geq (\inf_{x \in C, a \in \alpha(x)} E[e^{\gamma\{c(x,a)-\lambda_C^\Pi(\gamma)t(x,a)\}}])^h \cdot V(y) \cdot p = \infty,$$

where the last equality follows because $p > 0$, $V(y) = \infty$, and the fact that we know from Assumption 4.3.2 and Assumption 4.3.1 that $\inf_{x \in C, a \in \alpha(x)} E[e^{\gamma\{c(x,a)-\lambda_C^\Pi(\gamma)t(x,a)\}}] > 0$ is achieved.

But this is a contradiction of the fact that $V(\theta) \leq 1$ and the claim is proved.

If $V(\theta) = 1$, then policy Π is the policy Π^* named in the statement of Theorem 4.2.2 for the trivial MDP with $\Pi(x)$ being the only admissible action in state x if we substitute $\lambda = \lambda_C^\Pi(\gamma)$ and use (5.6) in place of the dynamic program (4.2). Furthermore, the claim showed that $V(\cdot)$ is bounded below. Also, Assumption 4.2.1 follows from Assumption 4.3.2, and Assumption 4.2.3 follows from Assumption 4.3.1 and Assumption 4.3.2. Therefore, Theorem 4.2.2 gives the desired result.

In general (for $V(\theta) \leq 1$), we observe the following: Since (5.7) holds, we know that (4.15) holds with $W(x) \doteq \ln[\max\{V(z), I(\theta)\}]$. Therefore, we know by appealing to Corollary 4.2.1 with an argument analogous to the one in the paragraph above that $J_x^\Pi \leq \lambda_C^\Pi(\gamma)$.

Claim:

$J_x^\Pi \geq \lambda_C^\Pi(\gamma)$.

Proof of claim:

If $V(\theta) = 1$, then the claim is true by the earlier argument. If $V(\theta) < 1$, then we know by Lemma 5.1.1 that $C^{\theta\to\theta}(\lambda) = \infty \ \forall \ \lambda < \lambda_C^\Pi(\gamma))$.

sub-claim:

If $V(\theta) < 1$ and $\lambda < \lambda_C^\Pi(\gamma)$, then

$$\lim_{N\to\infty} E_x^\Pi[e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] = \infty.$$

Proof of sub-claim:

$$\lim_{N\to\infty} E_x^\Pi[e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]$$

$$= \lim_{N\to\infty} E_x^\Pi[e^{\sum_{k=0}^{\tau_\theta-1} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}} \cdot e^{\sum_{k=\tau_\theta}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]$$

$$= E_x^\Pi[e^{\sum_{k=0}^{\tau_\theta-1} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]\cdot \sum_{M=1}^{\infty} P[\tau_\theta = M]\cdot \lim_{N\to\infty} E_x^\Pi[e^{\sum_{k=M}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}|\tau_\theta = M]$$

$$= V(x) \cdot \sum_{M=1}^{\infty} P[\tau_\theta = M] \cdot \lim_{N\to\infty} E_x^\Pi[e^{\sum_{k=M}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}|\tau_\theta = M].$$

For any $M < \infty$, we know that

$$\lim_{N\to\infty} E_x^\Pi[e^{\sum_{k=M}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}|\tau_\theta = M] = \lim_{N\to\infty} E_\theta^\Pi[e^{\sum_{k=0}^{N-M} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]$$

$$= \lim_{N\to\infty} E_\theta^\Pi[e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}].$$

Therefore,

$$\lim_{N\to\infty} E_x^{\Pi}[e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]$$

$$= V(x)\cdot \lim_{N\to\infty} E_\theta^{\Pi}[e^{\sum_{k=0}^{\tau_\theta -1} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}} \cdot e^{\sum_{k=\tau_\theta}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]$$

$$= V(x)\cdot E_\theta^{\Pi}[e^{\sum_{k=0}^{\tau_\theta -1} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]\cdot \sum_{M=1}^{\infty} P[\tau_\theta = M]\cdot$$

$$\lim_{N\to\infty} E_\theta^{\Pi}[e^{\sum_{k=M}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}|\tau_\theta = M]$$

$$= V(x)\cdot C^{\theta\to\theta}(\lambda)\cdot \sum_{M=1}^{\infty} P[\tau_\theta = M]\cdot \lim_{N\to\infty} E_\theta^{\Pi}[e^{\sum_{k=M}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}|\tau_\theta = M].$$

For any $M < \infty$, we know that

$$\lim_{N\to\infty} E_\theta^{\Pi}[e^{\sum_{k=M}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}|\tau_\theta = M] = \lim_{N\to\infty} E_\theta^{\Pi}[e^{\sum_{k=1}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}.$$

Therefore,

$$\lim_{N\to\infty} E_x^{\Pi}[e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] = V(x)\cdot C^{\theta\to\theta}(\lambda)\cdot \lim_{N\to\infty} E_x^{\Pi}[e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}].$$

Since $V(x) > 0$ and $C^{\theta\to\theta}(\lambda) = \infty$, we must have

$$\lim_{N\to\infty} E_x^{\Pi}[e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] = \infty,$$

and the sub-claim is proved.

By the sub-claim and Lemma 8.2.4 (2), we know that if $\lambda < \lambda_C^{\Pi}(\gamma))$, then $J_x^{\Pi} \geq \lambda$. And the claim is proved.

The Lemma follows from the claim and the argument just preceding the claim.

□

Lemma 5.3.2 *If $\bar{\gamma}_C^\Pi > 0$, then $\lambda_C^\Pi(\gamma)$ is a nondecreasing function of γ over $\gamma \in (0, \bar{\gamma}_C^\Pi)$.*

Outline of Proof:

Let γ_1, γ_2 be given such that $0 < \gamma_1 < \gamma_2 < \infty$. Jensen's inequality can be used to show that $\lambda_C^\Pi(\gamma_2) \geq \lambda_C^\Pi(\gamma_1)$.

□

Lemma 5.3.3 *If $\bar{\gamma}_C^\Pi > 0$, then $\lambda_C^\Pi(\gamma)$ is continuous over $\gamma \in (0, \bar{\gamma}_C^\Pi)$.*

Proof:

Claim:

$\forall \theta \in C$, $E_\theta^\Pi[e^{\sum_{k=0}^{\tau_\theta - 1} \gamma(c(x_k,\Pi(x_k)) - \lambda t(x_k,\Pi(x_k)))}]$ is increasing in γ for $\gamma \in (0, \bar{\gamma}_C^\Pi)$.

Proof of claim:

The claim follows because the exponential function is increasing.

Claim:

$\frac{d}{d\gamma}[E_\theta^\Pi[e^{\sum_{k=0}^{\tau_\theta - 1} \gamma(c(x_k,\Pi(x_k)) - \lambda t(x_k,\Pi(x_k)))}]$ exists and is finite and increasing for $\gamma \in (0, \bar{\gamma}_C^\Pi)$.

Proof of claim:

Denote

$$C(\lambda) \doteq \sum_{k=0}^{\tau_\theta - 1} c(x_k, \Pi(x_k)) - \lambda t(x_k, \Pi(x_k)).$$

By taking the Taylor series of e^x and the fact that the expected value of the sum is the sum of the expected values, we have

$$\frac{d}{d\gamma}[E_\theta^\Pi[e^{\gamma C(\lambda)}]] = E_\theta^\Pi[C(\lambda) \cdot e^{\gamma C(\lambda)}]. \tag{5.8}$$

$C(\lambda) > 0$ w.p.1 since $T_\theta < \infty$ w.p.1 by positivity of the chain induced by Π on C. Therefore, $\frac{d}{d\gamma}[E_\theta^\Pi[e^{\gamma C(\lambda)}]]$ is greater than zero for $\gamma \in (0, \bar{\gamma}_C^\Pi)$. Since the derivatives of both $E_\theta^\Pi[C(\lambda)]$ and $E_\theta^\Pi[e^{\gamma C(\lambda)}]$ are positive, they are both increasing in γ. Furthermore, since $e^{\gamma C(\lambda)}$ and $C(\lambda)$ are increasing in terms of each other, we have that $E_\theta^\Pi[C(\lambda) \cdot e^{\gamma C(\lambda)}]$ is increasing. Therefore, by (5.8), we see that $\frac{d}{d\gamma}[E_\theta^\Pi[e^{\gamma C(\lambda)}]$ is increasing for $\gamma \in (0, \bar{\gamma}_C^\Pi)$.

All that remains to be shown is that $\frac{d}{d\gamma}[E_\theta^\Pi[e^{C(\lambda)}] < \infty$ for $\gamma \in (0, \bar{\gamma}_C^\Pi)$. Suppose not, i.e., suppose $\exists \gamma'' < \bar{\gamma}_C^\Pi$ such that $\frac{d}{d\gamma}[E_\theta^\Pi[e^{\gamma C(\lambda)}]|_{\gamma = \gamma''} = \infty$. Then for any $\gamma > \gamma''$, we have that $E_\theta^\Pi[e^{C(\lambda)}] = \infty$, which contradicts $\gamma'' < \bar{\gamma}_C^\Pi$. And the claim is proved!

The Lemma is a consequence of this claim.

□

Define $\lambda_C^\Pi(0) \doteq \inf(\lambda \in \Re | E_\theta^\Pi[\sum_{k=0}^{\tau_\theta - 1}(c(x_k, \Pi(x_k)) - \lambda t(x_k, \Pi(x_k)))I(\tau_\theta < \infty)] \leq 0)$.

Lemma 5.3.4 *Suppose that $\bar{\gamma}_C^\Pi > 0$. Then,*

$$\lim_{\gamma \downarrow 0} \lambda_C^\Pi(\gamma) = \lambda_C^\Pi(0).$$

Proof:

Since $\bar{\gamma}_C^\Pi > 0$, we know by (5.1) that $\exists \gamma' > 0$ and $0 < \lambda' < \infty$ such that

$$E_\theta^\Pi[e^{\gamma' \sum_{k=0}^{\tau_\theta - 1}\{c(x_k, \Pi(x_k)) - \lambda' t(x_k, \Pi(x_k))\}} I(\tau_\theta < \infty)] \leq 1. \tag{5.9}$$

By the fact that C is a positive recurrent subclass under Π, we know that $I(\tau_\theta < \infty) = 1$ w.p.1. And by Jensen's inequality, we obtain

$$E_\theta^\Pi[\gamma' \sum_{k=0}^{\tau_\theta - 1}\{c(x_k, \Pi(x_k)) - \lambda' t(x_k, \Pi(x_k))\}] \leq \ln(1) = 0.$$

Since this holds true for any $\gamma' > 0, \lambda' < \infty$ such that (5.9) holds, we see that

$$\inf(\lambda \in \Re | E_\theta^\Pi[\sum_{k=0}^{\tau_\theta - 1}(c(x_k, \Pi(x_k)) - \lambda t(x_k, \Pi(x_k)))I(\tau_\theta < \infty)] \leq 0) \leq \lambda_C^\Pi(\gamma) \forall \gamma > 0$$

and therefore

$$\inf(\lambda \in \Re | E_\theta^\Pi[\sum_{k=0}^{\tau_\theta - 1}(c(x_k, \Pi(x_k)) - \lambda t(x_k, \Pi(x_k)))I(\tau_\theta < \infty)] \leq 0) \leq \lim_{\gamma \downarrow 0} \lambda_C^\Pi(\gamma).$$

Suppose that

$$\inf(\lambda \in \Re | E_\theta^\Pi[\sum_{k=0}^{\tau_\theta - 1}(c(x_k, \Pi(x_k)) - \lambda t(x_k, \Pi(x_k)))I(\tau_\theta < \infty)] \leq 0) < \lim_{\gamma \downarrow 0} \lambda_C^\Pi(\gamma).$$

Then $\exists \bar{\lambda} > 0$ such that

$$E_\theta^\Pi[\sum_{k=0}^{\tau_\theta - 1}\{c(x_k, \Pi(x_k)) - \bar{\lambda} t(x_k, \Pi(x_k))\}] \leq 0 \tag{5.10}$$

and $\exists \lambda^{'} > \bar{\lambda}$ such that

$$E_\theta^\Pi[e^{\gamma \sum_{k=0}^{\tau_\theta - 1}\{c(x_k, \Pi(x_k)) - \lambda^{'} t(x_k, \Pi(x_k))\}} I(\tau_\theta < \infty)] \geq 1 \forall \gamma > 0, \tag{5.11}$$

where the second inequality follows since $\lambda_C^\Pi(\gamma)$ is increasing $\forall 0 < \gamma < \bar{\gamma}_C^\Pi$. (We select $\lambda^{'} \leq \lim_{\gamma \downarrow 0} \lambda_C^\Pi(\gamma)$.)

Recall the notation of Lemma 5.3.3 and the Taylor series in terms of the expectation of the moments of $C(\lambda)$. From the expansion, we get

$$\lim_{\gamma \downarrow 0} \frac{E_\theta^\Pi[e^{\gamma C(\lambda^{'})}] - 1}{\gamma} = E_\theta^\Pi[C(\lambda^{'})] \geq 0,$$

where the inequality follows from (5.11). But since $E_\theta^\Pi[C(\lambda)]$ is decreasing in λ and $\lambda^{'} > \bar{\lambda}$, we have a contradiction of (5.10).

□

Lemma 5.3.5 *If* Π *induces a null recurrent subclass* $C \subset S$*, then* $\lambda_C^\Pi(\gamma) = \infty$ $\forall \gamma \in [0, \infty)$*.*

Proof:

Since the induced Markov Chain over C is null recurrent, the long term average risk neutral cost, $\lambda_C^\Pi(0) = \infty$. Thus by Lemma 5.3.2 and Lemma 5.3.4, $\lambda_C^\Pi(\gamma) = \infty$ $\forall \gamma > 0$.

□

The above lemma is important because it shows that no policy that does not induce a positive recurrent class can have a finite risk sensitive average cost starting from any state.

5.4 Performance for a large risk sensitivity parameter

For a discrete time, finite horizon MDP, it is well known that the cost of a policy approaches the 'maximum cost' as the risk sensitivity parameter approaches ∞, where the 'maximum cost' is the cost of the most expensive realization that occurs with nonzero probability.

In this thesis, we have generalized the cost structure in two ways: we consider average cost over the infinite horizon instead of finite horizon cost, and we consider an SMDP instead of an MDP. Determining what happens in an SMDP as $\gamma \to \infty$ is a very tricky technical problem that we will not explore further. Instead, we will generalize the result to the average cost case over the infinite horizon for an MDP.

Before stating the result, we need to introduce some notation:

Suppose that stationary policy $\Pi \in \Pi^{MD}$ induces a finite irreducible class $C \subset S$. Define an *admissible cycle* for policy Π as a finite sequence of states, starting and ending at the same state, such that each transition occurs with nonzero probability under policy Π, i.e., $\psi = \{x_0, x_1, x_2, ..., x_n\}$ is an admissible cycle if

1. $x_0 = x_n$
2. $x_{i+1} \in r(x_i, \Pi(x_i))$; $i = 0, 1, 2, ..., n-1$

Because we are considering MDPs, we assume that the cost of transition out of a state is fixed with probability one. Denote the cost of transitioning out of a state by $\Delta C(x)$.

(Alternately, we could allow transition costs to be non-deterministic as long as there is an upper bound B such that $0 < c(x, \Pi(x)) < B$ w.p.1 $\forall x \in C$. Then it is easy to show that $\lim_{\gamma \to \infty} \frac{1}{\gamma} \ln E[e^{\gamma c(x,\Pi(x))}] = \sup\{C \in \Re^+ | P[c(x, \Pi(x)) \geq C] > 0\}$, and for the purpose of determining performance with large sensitivity parameter we can just set $c(x, \Pi(x)) \doteq \sup\{C \in \Re^+ | P[c(x, \Pi(x)) \geq C] > 0\}$.)

Define the average cycle cost $C(\{x_0, x_1, x_2, ..., x_n\}) = \frac{1}{\gamma n} \ln[\Pi_{i=0}^{n-1} \Delta C(x_i)]$. The average cycle cost can also be expressed as $C(\psi) = \frac{1}{\gamma n} \sum_{i=0}^{n-1} \ln[\Delta C(x_i)]$.

Denote the set of all admissible cycles for policy Π as Ψkl^{Π} (where Ψkl is pronounced 'cycle',) and define the maximum average cycle cost ξ as follows:

$$\xi = \sup_{\psi \in \Psi kl^{\Pi}} C(\psi).$$

A cycle $\{x_0, x_1, x_2, ..., x_n\}$ is called *non-redundant* if $x_i \neq x_j$ $\forall i \neq j$ such that $i, j \in \{1, 2, ..., n\}$. I.e., a cycle is non-redundant if it contains no sub-cycles. Denote the set of all non-redundant cycles admissible under policy Π as Ψkl_{nr}^{Π}.

Lemma 5.4.1

$$\xi = \sup_{\psi \in \Psi kl_{nr}^{\Pi}} C(\psi).$$

Proof:

$\Psi kl_{nr}^{\Pi} \subset \Psi kl^{\Pi}$, so $\xi \geq \sup_{\psi \in \Psi kl_{nr}^{\Pi}} C(\psi)$.

Suppose that $\psi = \{x_0, x_1, x_2, ..., x_n\} \in \Psi kl^{\Pi}$.

Claim:

$\exists \psi' \in \Psi kl_{nr}^{\Pi}$ such that $C(\psi') \geq C(\psi)$.

(sketch of) Proof of claim:

The following procedure will terminate in finite time and generate a $\psi' \in \Psi kl_{nr}^{\Pi}$ that satisfies the claim:

1. Let set $\bar{\psi} = \psi$.

2. If $\bar{\psi} \in \Psi kl_{nr}^{\Pi}$, then set $\psi' = \bar{\psi}$ and terminate the procedure.

3. Because $\bar{\psi} = \{x_0, x_1, x_2, ..., x_n\} \notin \Psi kl_{nr}^{\Pi}$, $\exists a, b \in \{1, 2, ..., n\}$, $a < b$ such that $x_a = x_b$.

Set

$$\bar{\psi}^1 = \{x_0, x_1, ..., x_{a-1}, x_a, x_{b+1}, x_{b+2}, ..., x_n\}$$

and set

$$\bar{\psi}^2 = \{x_a, x_{a+1}, ..., x_{b-1}, x_b\}.$$

4. If $C(\bar{\psi}^1) > C(\bar{\psi}^2)$, then set $\bar{\psi} = \bar{\psi}^1$. Otherwise, set $\bar{\psi} = \bar{\psi}^2$.

5. Go to step **2**.

And the claim can be seen to be true.

The claim shows that $\xi \leq \sup_{\psi \in \Psi kl^{\Pi}_{nr}} C(\psi)$.

□

Lemma 5.4.1 shows that the supremum in the definition of ξ is achieved because there are only a finite number of non-redundant cycles for $|C| < \infty$.

For $z \in C$, define $\Psi kl^{\Pi}_z \subset \Psi kl^{\Pi}$ as the set of all admissible policies for policy Π that start and end at state z. Clearly, $\Psi kl^{\Pi}_z \cap \Psi kl^{\Pi}_w = \emptyset$ for $z \neq w$ and $\Psi kl^{\Pi} = \cup_{z \in C} \Psi kl^{\Pi}_z$.

Lemma 5.4.2 *For any $z \in C$,*

$$\xi = \sup_{\psi \in \Psi kl^{\Pi}_z} C(\psi).$$

Proof:

By Lemma 5.4.1, there is a nonredundant cycle ψ^* such that $C(\psi^*) = \xi$. $\psi^* \in \Psi kl^{\Pi}_w$ for some $w \in C$. If $z = w$, we are done. Suppose $z \neq w$. By irreducibility, there is a cycle $\psi_{zw} = \{x_0, x_1, ..., x_n\}$ such that $x_0 = x_n = z$ and $x_i = w$ for some i.

Define a sequence of cycles as follows: $\psi^1 = \psi^*$. $\psi^2 = \psi^* \psi^*$, or the concatenation of ψ^* with itself. $\psi^{k+1} = \psi^k \psi^*$. I.e., ψ^k is the concatenation of ψ^* with itself k times. Clearly $C(\psi^k) = \xi \ \forall k$.

We will now prove the Lemma by construction: we will define a sequence of cycles $\{\psi_{zw}^k | k = 0, 1, 2, 3, ...\}$ such that $\psi_{zw}^k \in \Psi kl_z^\Pi \; \forall k$ and $\lim_{k\to\infty} C(\psi_{zw}^k) = \xi$.

Define $\psi_{zw}^0 = \psi_{zw}$. Define ψ_{zw}^k by taking ψ_{zw}, removing an instance of w, and replacing that instance with ψ^k.

Suppose that ψ_{zw} has length n_1 and ψ^* has length n_2. Algebra tells us that $C(\psi_{zw}^k) = \frac{n_1 C(\psi_{zw}) + k \cdot n_2 \xi}{n_1 + k \cdot n_2}$. Therefore $\lim_{k\to\infty} C(\psi_{zw}^k) = \xi$.

□

Lemma 5.4.3 *Suppose that stationary policy $\Pi \in \Pi^{MD}$ induces a positive recurrent subclass $C \subset S$ with $|C| < \infty$, that all transition times $t(x, \Pi(x)) \equiv 1$ w.p.1, and that all transition costs are deterministic; i.e., the process is a discrete time Markov chain.*

Then $\lim_{\gamma\to\infty} J_x^\Pi(\gamma) = \xi \; \forall x \in C$.

Proof:

We know that

$$J_x^\Pi(\gamma) = \lim_{N\to\infty} \frac{1}{\gamma N} \ln E_x^\Pi[e^{\gamma \sum_{k=1}^N c(x_k)}] \leq \lim_{N\to\infty} \frac{1}{\gamma N} \ln[e^{\gamma \sum_{k=1}^N c(e_k)}],$$

where $\{e_1, e_2, ...\}$ is the most expensive admissible sample path.

$$\lim_{N\to\infty} \frac{1}{\gamma N} \ln[e^{\gamma \sum_{k=1}^N c(e_k)}] = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^N c(e_k) = \xi,$$

so we see that $\lim_{\gamma\to\infty} J_x^\Pi(\gamma) \leq \xi$.

We now must show that $\lim_{\gamma\to\infty} J_x^\Pi(\gamma) \geq \xi$, which is equivalent to the following statement:

$$\lim_{\gamma\to\infty} \lambda_C^\Pi(\gamma) \leq \xi.$$

Because γ is no longer fixed, we augment the notation for $C^{\theta\to\theta}(\lambda)$ by changing it to $C_\gamma^{\theta\to\theta}(\lambda)$.

By definition (and adapting for the deterministic costs and uniform, discrete transition times),

$$C_\gamma^{\theta\to\theta}(\lambda) = \sum_{\psi\in\Psi kl_\theta^\Pi} P[\psi]e^{\gamma C(\psi)}e^{-\gamma\lambda}. \tag{5.12}$$

Let $\{\psi_1, \psi_2, ...\}$ be a sequence of cycles in Ψkl_x^Π such that $\lim_{i\to\infty} C(\psi_i) = \xi$.

Define a sequence of positive real numbers as follows: $\lambda_i = \frac{i}{i+1}C(\psi_i)$. We see that $\lim_{i\to\infty}\lambda_i = \xi$.

From (5.12), we obtain

$$\lim_{\gamma\to\infty} C_\gamma^{\theta\to\theta}(\lambda_i) = \lim_{\gamma\to\infty} \sum_{\psi\in\Psi kl_\theta^\Pi} P[\psi]e^{\gamma C(\psi)}e^{-\gamma\lambda_i} \geq$$

$$\lim_{\gamma\to\infty} P[\psi_i]e^{\gamma C(\psi_i)}e^{-\gamma\lambda_i} = P[\psi_i]\lim_{\gamma\to\infty} e^{\gamma\frac{1}{i+1}C(\psi_i)} = \infty.$$

Therefore we must have $C_\gamma^{\theta\to\theta}(\lambda_i) > 1$ for γ large enough.

Recall that $\lambda \leq \lambda_C^\Pi(\gamma)$ if $C_\gamma^{\theta\to\theta}(\lambda) \geq 1$. Therefore $\lambda_i \leq \lambda_C^\Pi(\gamma)$ for γ large enough. Since $\lim_{i\to\infty}\lambda_i = \xi$, that means that $\xi \leq \lambda_C^\Pi(\gamma)$ and the Lemma is proved.

□

Corollary 5.4.1 *If Assumption 4.2.1 and* **(A2.2)** *hold and a policy* Π *induces a recurrent class* C *such that* $|C| = \infty$*, then*

$$\lim_{\gamma\to\infty} J_x^\Pi(\gamma) = \infty.$$

Proof:

Lemma 5.4.3 shows that as $\gamma \to \infty$, the long term average risk sensitive cost approaches the worst cycle cost. If there are infinitely many states and norm-like costs, the worst cycle cost is infinity. This holds for a semi-Markov process as well as a Markov process.

□

5.5 The risk neutral case and its relation to the risk sensitive case

The distinctions between semi-Markov chains with finite round trip cost at $\lambda = 0$, of Type I, and of Type II, all hold in the risk neutral case as well. When looking at figures 5.1 through 5.5, simply change '1' to '0', since in the risk neutral case we are seeking a round trip cost of zero. All else remains the same. (We state this without proof because its proof is quite similar to the proof in the risk sensitive case.)

The following lemma is the risk neutral version of Lemma 5.3.1.

Lemma 5.5.1 *Suppose that* Π *induces a positive recurrent subclass* $C \subset S$ *and that* $\bar{\gamma}_C^\Pi > 0$.

Then,

$$\lim_{T \to \infty} \frac{1}{T} E_x^\Pi [\int_{t=0}^T g(t)dt] = \lambda_C^\Pi(0); x \in C.$$

Proof:

The proof of this lemma mirrors precisely the proof of Lemma 5.3.1, through appropriate modifications to cover the risk neutral, instead of the risk sensitive, case. For that reason, the proof is omitted.

□

It is worth noting that if the risk neutral round trip cost at $\lambda_C^\Pi(0)$ is zero, then the dynamic program (risk neutral version of 4.2) has a solution. The result then follows from Theorem 11.4.6 and Proposition 11.4.7 in [35]. (To see this, note that policy Π confined to C is *unichain* according to Puterman's definition.) Of course the case where the round trip cost at $\lambda_C^\Pi(0)$ is less than zero (i.e., the semi-Markov chain is of Type II with risk neutral costs and $\lim_{\lambda \downarrow \lambda_i} C^{\theta \to \theta}(\lambda) < 0$) must be covered differently, as in the proof of Lemma 5.3.1.

Example 5.5.1

In order to illustrate the fact that under the conditions of Lemma 5.3.4 and Lemma 5.5.1, the long term average risk sensitive cost approaches the long term risk neutral cost as the risk sensitivity parameter approaches zero from above, let us examine a simple example. Suppose we have a Markov chain with two states: s_1 and s_2. Suppose that $c(s_1) =$ the cost of a transition from $s_1 = 1$, and $c(s_2) = 2$. Suppose furthermore that $p(s_1|s_1) =$ the probability of transitioning from s_1 to itself $= p(s_2|s_1) = .5$. Suppose furthermore that $p(s_1|s_2) = 1$, so that the system always transitions to state s_1 from state s_2. Clearly, the Markov chain is irreducible and positive recurrent. Therefore the conditions of Lemma 5.5.1 hold. If we can show that the risk sensitive average cost is defined for some $\gamma > 0$, then the conditions of Lemma 5.3.4 also hold.

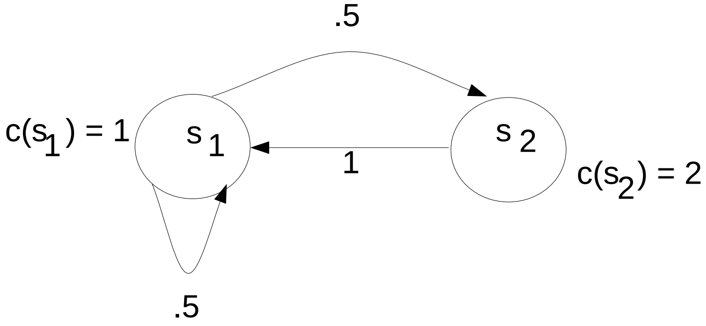

Figure 5.9: A simple Markov Chain.

It is well known ([1]) that the average risk neutral cost is the expected value over the ergodic distribution of the transition cost. The balance equations are:

$$P(s_1) = P(s_1) \cdot p(s_1|s_1) + P(s_2) \cdot p(s_1|s_2)$$

$$P(s_2) = P(s_1) \cdot p(s_2|s_1) + P(s_2) \cdot p(s_2|s_2),$$

which have solution

$$P(s_1) = \frac{2}{3}; P(s_2) = \frac{1}{3}.$$

Therefore $\lambda(0) = P(s_1) \cdot c(s_1) + P(s_2) \cdot c(s_2) = \frac{4}{3}$, and the risk neutral, long term average cost is $\frac{4}{3}$.

Let us now determine the risk sensitive long-term average cost through a technique we will call *recursive computation*:

By (4.2), we obtain:

$$e^{W(s_1)} = e^{\gamma\{c(s_1)-\lambda(\gamma)\}}[p(s_1|s_1)e^{W(s_1)} + p(s_2|s_1)E^{W(s_2)}]$$

$$e^{W(s_2)} = e^{\gamma\{c(s_2)-\lambda(\gamma)\}}[p(s_1|s_2)e^{W(s_1)} + p(s_2|s_2)E^{W(s_2)}]$$

Since only relative values of the value function $(W(\cdot))$ are relevant, we arbitrarily set $W(s_1) = 0$. Then, we substitute in values to obtain:

$$1 = e^{\gamma\{1-\lambda(\gamma)\}}[\frac{1}{2} + \frac{1}{2}e^{W(s_2)}]$$

$$e^{W(s_2)} = e^{\gamma\{2-\lambda(\gamma)\}}[1 \cdot 1].$$

Now we substitute the second equation into the first, yielding

$$\frac{1}{2} + \frac{1}{2}e^{\gamma\{2-\lambda(\gamma)\}} = e^{\gamma\{\lambda(\gamma)-1\}}.$$

Now let $x = e^{\gamma\lambda(\gamma)}$ and obtain a quadratic for x:

$$x^2 e^{-\gamma} - \frac{1}{2}x - \frac{1}{2}e^{2\gamma} = 0,$$

yielding by the Pythagorean Theorem

$$x = \frac{\frac{1}{2} \pm \sqrt{\frac{1}{4} + 2e^{\gamma}}}{2e^{-\gamma}}.$$

The negative root doesn't make sense because the exponential in the definition of x guarantees that $x > 0$. Therefore, we obtain

$$\lambda(\gamma) = \frac{1}{\gamma}\ln[\frac{\frac{1}{2} + \sqrt{\frac{1}{4} + 2e^{\gamma}}}{2e^{-\gamma}}].$$

Clearly then, $\bar{\gamma}_C = \infty > 0$. Application of L'hopital's rule gives us that $\lim_{\gamma \downarrow 0} \lambda(\gamma) = \frac{4}{3}$, which is the risk-neutral long-term average cost, as predicted by Lemma 5.3.4.

Another application of L'hopital's rule shows us that $\lim_{\gamma \uparrow 0} \lambda(\gamma) = \frac{3}{2}$.

So Lemma 5.4.3 predicts that the long-term average maximum cost is $\frac{3}{2}$. Let us examine figure 5.9. It is clear that the worse transition from state s_1 is to state s_2. This yields a Markov Chain that alternates deterministically between s_1 and s_2, yielding a cost of $\frac{3}{2}$, as expected.

This simple example illustrates the application of Lemmas 5.3.4 and 5.4.3, and further shows the difficulty of solving exactly for the long term average risk sensitive cost. Had there been more than a few states, solution of equation (4.2) would have required the use of approximation techniques.

5.6 A dynamic program for the case $C^{\theta \to \theta}(\lambda_C^{\Pi}(\gamma)) < 1$

This dynamic program covers both the case where round trip cost at $\lambda_C^{\Pi}(\gamma)$ is 1 and the case where round trip cost at $\lambda_C^{\Pi}(\gamma)$ is less than 1. It can be extended in the obvious way to cover the risk neutral case. (Note: this theorem and its proof are related to Lemma 5.3.1 and its proof.)

Theorem 5.6.1 *Assume that Assumption 4.3.2 holds. If for some $\theta \in S$, policy Π induces a positive recurrent subclass C containing θ, and $\lambda_C^{\Pi}(\gamma) < \infty$, then $\exists$ a solution $\{W_\theta(\cdot)\}$, finite for each $x \in C$ and bounded below, to the following functional equation:*

$$e^{W_\theta(x)} = E[e^{\gamma\{c(x,\Pi(x)) - \lambda_C^{\Pi}(\gamma) t(x,\Pi(x))\}}] \int \max\{e^{W_\theta(y)}, I(y=\theta)\} P(dy|x,\Pi(x)); \, \forall x \in C, \tag{5.13}$$

with $W_\theta(\theta) = \ln[C^{\theta\to\theta}(\lambda_C^\Pi(\gamma))] \le 1$.

Proof: Define $W_\theta(x) \doteq \ln\{E_x^\Pi[e^{\sum_{k=0}^{\tau_\theta}\gamma\{c(x,a)-\lambda_C^\Pi(\gamma)t(x,a)\}}\}]$. By definition, (5.13) holds.

Claim $W_\theta(\theta) \le 0$.

Proof of claim:

Since the embedded Markov chain induced by Π on C is positive recurrent, we know that $E_\theta^\Pi[\tau_\theta] < \infty$. Therefore, $\tau_\theta < \infty$ w.p.1. By (5.1) and Fatou's Lemma, we have that $W_\theta(\theta) \le 0$.

Claim

$W_\theta(x) < \infty \; \forall x \in C$.

Proof of claim:

Suppose $\exists x \in C$ s.t. $W_\theta(x) = \infty$. (So we know that $e^{W_\theta(x)} = \infty$. Then, by (5.13), any $y \in C$ such that $x \in r(y, \Pi(y))$ must also have $W_\theta(y) = \infty$. By induction, any $z \in C$ such that x is reachable in finitely many steps without first hitting θ with nonzero probability from z must also have $W_\theta(z) = \infty$. Since the embedded Markov chain induced by Π on C is recurrent, we must have that either $W_\theta(\theta) = \infty$. This contradicts $W_\theta(\theta) \le 0$ and the claim follows.

Because of the norm-like cost assumption Assumption 4.3.2, we know that $CC_h = \{x \in C | E[e^{\gamma c(x,\Pi(x)) - \lambda_C^\Pi(\gamma)t(x,\Pi(x))}] \le h\}$ has finitely many members for any $h > 0$.

Claim

$\inf_{x\in CC_2} W_\theta(x) = \inf_{x\in C} W_\theta(x)$.

Proof of claim:

Choose $x \notin CC_2$. Since $x \notin CC_2$, we know that $E[e^{\gamma c(x,\Pi(x)) - \lambda_C^\Pi(\gamma)t(x,\Pi(x))}] > 2$. So by (5.13) we get

$$e^{W_\theta(x)} > 2\int \max\{e^{W_\theta(y)}, I(\theta = y)\}P(dy|x,\Pi(x)) \ge 2\int e^{W_\theta(y)}P(dy|x,\Pi(x)).$$

And so $e^{W_\theta(x)} > 2\inf_{x\in C} W_\theta(x)$. And the claim is proved.

Claim

$\{W_\theta(x)|x \in C\}$ is bounded below.

Proof of claim:

Suppose $\exists z \in C$ s.t. $W_\theta(z) = -\infty$. Therefore

$$E_z^\Pi[e^{\gamma \sum_{k=0}^{\tau_\theta - 1}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] = 0.$$

We know by the norm-like cost Assumption 4.3.2 that $\exists B > 0$ such that $E[e^{\gamma\{c(x,a)-\lambda t(x,a)\}}] \geq B\ \forall x, a$.

Therefore

$$0 = E_z^\Pi[e^{\gamma \sum_{k=0}^{\tau_\theta - 1}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] \geq E_z^\Pi[B^{\tau_\theta}]$$

$$= \sum_{k=0}^{\infty} B^k P_z^\Pi[\tau_\theta = k],$$

which implies that $P_z^\Pi[\tau_\theta = k] = 0\ \forall k$, and therefore $P_z^\Pi[\tau_\theta < \infty] = 0$, which contradicts irreducibility of C under Π, and therefore there cannot be any $z \in C$ with $W_\theta(z) = -\infty$.

The previous claim tells us that $\inf_{x\in CC_2} W_\theta(x) = \inf_{x\in S} W_\theta(x)$. This infimum must be finite since CC_2 has finitely many members.

□

Note: If $W_\theta(\theta) = 0$ (i.e, if $C^{\theta\to\theta}(\lambda_C^\Pi(\gamma)) = 1$), then (5.13) reduces to

$$e^{W_\theta(x)} = E[e^{\gamma\{c(x,\Pi(x))-\lambda_C^\Pi(\gamma)t(x,\Pi(x))\}}] \int e^{W_\theta(y)} P(dy|x,\Pi(x));\ \forall x \in C, \qquad (5.14)$$

Lemma 5.6.1 *Suppose that stationary $\Pi \in \Pi^{MD}$ induces a positive recurrent class $C \subset S$ and that $W(\cdot)$ satisfies (5.13).*

Then $\forall x, y \in C$

$$E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_y}[c(x_k,a)-\lambda t(x_k,a)]}] \leq e^{W(x)-W(y)},$$

with equality if $W(\theta)=0$, *i.e., if* $C^{\theta\to\theta}(\lambda_C^{\Pi})=1$.

Proof:

We rewrite (5.13) using **(AH)** and the fact that the state space is countable as

$$e^{W(x)} = E[e^{\gamma\{c(x,\Pi(x))-\lambda_C^{\Pi}(\gamma)t(x,\Pi(x))\}}].$$

$$\sum_{y\in r(x,\Pi(x))} \max\{e^{W(y)}, I(y=\theta)\}P(y|x,\Pi(x));\ \forall x \in C. \tag{5.15}$$

This is a set of linear equations in $\{e^{W(x)}|x\in C\}$. For notational simplicity, we will denote $P(j|i,\Pi(i)) \doteq p_{ij}$ and $E[e^{\gamma\{c(x,\Pi(x))-\lambda_C^{\Pi}(\gamma)t(x,\Pi(x))\}}] \doteq c_i$. And if $j \notin r(i,\Pi(i))$, we set $p_{ij}=0$. We can then rewrite (5.15) as

$$e^{W(i)} = c_i \sum p_{ij} \max\{e^{W(j)}, I(y=\theta)\}. \tag{5.16}$$

Because C is a positive recurrent subclass induced by Π, we know that $\forall i,j \in C$, $P_i^{\Pi}[\tau_m<\infty]=1$. So if we let C_s^{im} denote the set of all finite sequences of states in C that start with i, contain m only once, and end with m; then we have the following identities:

$$\sum_{ss\in C_s^{im}} P(ss) = 1,$$

and

$$E_i^{\Pi}[e^{\gamma\sum_{k=0}^{\tau_m}[c(x_k,\Pi(x_k))-\lambda t(x_k,\Pi(x_k))]}] = \sum_{ss\in C_s^{im}} P(ss)C(ss). \tag{5.17}$$

where $ss=\{ss(N);ss_0,ss_1,...,ss_{ss(N)-1}\}$ denotes an element of C_s^{ij} of length $ss(N)$ with $ss_0=i$ and $ss_{ss(N)-1}=j$; $P(ss)=\Pi_{k=0}^{ss(N)-1}p_{ss_k,ss_{k+1}}$ is the probability of sequence ss, and $C(ss)=\Pi_{k=0}^{ss(N)-1}c_{ss_k}$ is the cost of sequence ss.

Claim:

$$e^{W(i)-W(m)} \geq \sum_{ss \in C_s^{im}} P(ss)C(ss);\ i, m \in C,$$

with equality if $W(\theta) = 1$.

Proof of claim:

We proceed by induction.

We apply (5.16) to each $e^{W(j)}$ on the right hand side. Repeating this process recursively yields

$$e^{W(i)} = [\sum_{ss \in C_s^{im}} P(ss)C(ss)] \max\{e^{W(j)}, I(y = \theta)\} e^{-N_\theta(ss) \cdot W(\theta)},$$

where $N_\theta(ss) \doteq$ the number of times θ appears in ss except the first and last elements of ss. I.e., $N_\theta(ss) + I(i = \theta) + I(j = \theta) =$ the number of times θ appears in ss.

And the claim follows because $W(\theta) \leq 1$.

The above claim, combined with (5.17) yields that

$$E_i^\Pi[e^{\gamma \sum_{k=0}^{\tau m}[c(x_k, \Pi(x_k)) - \lambda t(x_k, \Pi(x_k))]}] = e^{W(i)-W(m)} e^{N_\theta(ss) \cdot W(\theta)}.$$

□

Corollary 5.6.1 *If $W(\theta) = 1$, then $\forall \theta_1, \theta_2 \in C$ and all $x \in C$, $W_{\theta_1}(x) = W_{\theta_2}(x) \cdot W_{\theta_1}(\theta_2)$.*

From this it can be inferred that if the round trip cost at λ_C^Π is 1 for θ, then it is 1 for any state; and conversely, if it is ≤ 1 for θ, it therefore must be ≤ 1 for any state.

In fact, we can extract similar results to Corollary 4.3 if (5.13) holds not just over a positive recurrent class, but over all of S; i.e.,

$$e^{W(x)} = E[e^{\gamma\{c(x,\Pi(x))-\lambda t(x,\Pi(x))\}}]\int \max\{e^{W(y)}, I(y=\theta)\}P(dy|x,\Pi(x));\ \forall x \in S. \tag{5.18}$$

For $D, F \subset S$, define $\rho_x^\Pi(E,F) = P_x^\Pi[\tau_E < \tau_F]$. If either E or F contain only one element, they may be replaced in the notation by that single element.

Lemma 5.6.2 *Suppose that (5.18) holds. Let $A \subset S$ and let Π be a Markov, stationary policy. Suppose that $P_x^\Pi[\tau_A < \infty] = 1$. Then, $E_x^\Pi[e^{\gamma\sum_{k=0}^{\tau_A} c(x,\Pi(x))-\lambda t(x,\Pi(x))}] \leq \sum_{y\in A}\rho_x^\Pi(y, A-y)e^{W(x)-W(y)}$, with equality if $W(\theta)=1$.*

Proof:

We follow the same notation used in Lemma 5.6.1 and extend it slightly. For $x \in s$, $A \subset S$, and $B \subset S$, let $C_s^{x,A,B}$ denote the set of all finite sequences of states in S that start with x, contain no state in $A \cup B$ except possibly for the last state, and have a last state contained in A.

Because we are given that $P_x^\Pi[\tau_A < \infty] = 1$, we know that

$$\sum_{y\in A}\rho_x^\Pi(y, A-y) = \sum_{y\in A}\sum_{ss\in C_s^{x,y,A-y}} P(ss) = \sum_{ss\in C_s^{x,A,A}} P(ss) = 1,$$

and

$$E_x^\Pi[e^{\gamma\sum_{k=0}^{\tau_A}[c(x_k,\Pi(x_k))-\lambda t(x_k,\Pi(x_k))]}] = \sum_{ss\in C_s^{x,A,A}} P(ss)C(ss) =$$

$$\sum_{y\in A}\sum_{ss\in C_s^{x,y,A-y}} P(ss)C(ss) \leq \sum_{y\in A}\rho_x^\Pi(y, A-y)e^{W(x)-W(y)},$$

where the final inequality can be proven through the same procedure used to prove the claim within the proof of Lemma 5.6.1. □

Chapter 6

An Optimal Policy Under the Assumption of Reachability with Finite Expected Cost

6.1 Reachability with finite expected cost

The following assumption will be needed to help guarantee the existence of an optimal policy in Theorem 6.1.1. In words, it means that one can get from any state to any other state with finite expected cost for the risk parameter γ.

Assumption 6.1.1 (γ) *$\forall x, y \in S$, $\exists$ a policy $\Pi^{x \to y}$ such that $E_x^{\Pi^{x\to y}}[e^{\gamma \sum_{k=0}^{\tau_y} c(x_k, a_k)}] < \infty$.*

Note: Because $E[c(x,a)]$ is bounded below away from zero, we also have that $E_x^{\Pi^{x\to y}}[\tau_y] < \infty$.

This assumption is designed to prevent the accrual of infinite expected costs going between states in the optimal policy. Without it, as pointed out in [9], the cost of a stationary Markov policy may depend on the initial state. The following example is adapted from [9]:

Example 6.1.1 (Infinite cost to escape a state)

Suppose a discrete time Markov chain has 2 states: x_1 and x_2. The cost of being in state x_1 is $c(x_1) = c$, and the cost of being in state x_2 is $c(x_2) = 0$. The state x_2 is absorbing, i.e., $p(x_2|x_2) = 1$. Therefore we know that $J(x_2) = 0$. Suppose that $p(x_1|x_1) = p$ and $p(x_2|x_1) = 1 - p$. The expected risk sensitive cost to get from x_1 to x_2 is given by

$$E_{x_1}[e^{\gamma \sum_{k=0}^{\tau_{x_2}-1} c}] = \sum_{k=1}^{\infty} e^{k\gamma c} P[\tau_{x_2} = k]$$

$$= \sum_{k=1}^{\infty} \frac{1-p}{p} e^{k\gamma c} p^k = \sum_{k=1}^{\infty} \frac{1-p}{p} (pe^{\gamma c})^k,$$

which is finite only if $pe^{\gamma c} < 1$. Suppose that $pe^{\gamma c} > 1$. Then we get

$$J(x_2) = \lim_{N\to\infty} \frac{1}{N\gamma} \ln E_{x_1}[e^{\gamma \sum_{k=0}^{N-1} c(x_k)}]$$

$$= \lim_{N\to\infty} \frac{1}{N\gamma} \ln E_{x_1}[e^{\gamma \sum_{k=0}^{\tau_{x_2}-1} c}] = \lim_{N\to\infty} \frac{1}{N\gamma} \ln \sum_{k=1}^{N-1} \frac{1}{p} (pe^{\gamma c})^k$$

$$= \lim_{N\to\infty} \frac{1}{N\gamma} \ln[\frac{1}{p} \frac{pe^{\gamma c} - (pe^{\gamma c})^N}{1 - pe^{\gamma c}}]$$

$$= \lim_{N\to\infty} \frac{1}{N\gamma} \ln[(pe^{\gamma c})^N] = \ln[pe^{\gamma c}] > J(x_1) = 0,$$

and so the value of the objective function depends on the initial state.

Definition 6.1.1 *A policy Π' is called 'shortest path optimal' to reach state y if $\forall \Pi \in \Pi^{HR}$,*

$$E_x^{\Pi'}[e^{\gamma \sum_{k=0}^{\tau_y - 1} c(x_k, a_k)}] \leq E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_y - 1} c(x_k, a_k)}]; \forall x \in S.$$

If the above inequality only holds for a particular $x \in S$, then Π' is called 'shortest path optimal' from x to y.

The following lemma demonstrates that Assumption 6.1.1 guarantees the existence of shortest path optimal policies.

Lemma 6.1.1 *Under assumption 6.1.1(γ) and given Assumptions 4.3.3 and 2.0.2; for any $\theta \in S$, there exists a Markov, stationary policy Π_γ^θ such that*

$$E_x^{\Pi_\gamma^\theta}[e^{\gamma \sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)}] < \infty; \forall x \in S.$$

Furthermore, $\forall \Pi \in \Pi^{HR}$,

$$E_x^{\Pi_\gamma^\theta}[e^{\gamma \sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)}] \leq E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)}]; \forall x \in S,$$

i.e., Π_γ^θ is shortest path optimal.

Proof:

Given $x \in S$, we know by assumption 6.1.1(γ) that there exists a policy that drives the system to θ with finite cost. Let

$$Q(x,a) \doteq \inf_{\Pi \in \Pi^{HR}} E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)} | a(0) = a]; x \neq \theta,$$

and let

$$v(x) \doteq \inf_{\Pi \in \Pi^{HR}} E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)}] = \inf_{a \in \alpha(x)} Q(x,a); x \neq \theta. \tag{6.1}$$

We have

$$Q(x,a) = E[e^{\gamma c(x,a)}] \cdot \sum_{z \in r(x,a)} P(z|x,a) v(z); x \neq \theta. \tag{6.2}$$

where we define $Q(\theta, a) \doteq 0 \; \forall a \in \alpha(\theta)$. Since the transition law $P(x|s,a)$ is continuous in a for a fixed s and by Assumption 4.3.3, we see that $Q(x,a)$ is a continuous function of a. Therefore, it achieves its infimum at $a^*(x)$. Define the policy $\Pi_\gamma^\theta(x) \doteq a^*(x) \; \forall x$.

Claim:

Π_γ^θ minimizes $E_x^\Pi[e^{\gamma \sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)}]$ over all policies for each $x \in S$.

Proof of claim:

We see from (6.1) and (6.2) that

$$v(x) = \inf_{a \in \alpha(x)} E[e^{\gamma c(x,a)}] \cdot \sum_{z \in r(x,a)} P(z|x,a)v(z); x \neq \theta. \tag{6.3}$$

And the infimum is achieved since $\cup_{a \in \alpha(x)} r(x,a)$ is finite by Assumption 4.3.3; and $\alpha(x)$ is compact and $P(y|x,a), P(c|x,a)$ are continuous in a by Assumption 2.0.2. The policy Π_γ^θ is the policy Π^* named in the statements of Lemma 4.2.1 and Theorem 4.2.2 if we substitute $\lambda = 0$ and use (6.3) in place of the dynamic program (4.2). By Lemma 4.2.1, (4.3) then holds.

We know from Theorem 4.2.2 that (4.9) holds for policy Π_γ^θ for any value of N.

Because costs accrue until θ is reached, we have

$$E_{x_0}^{\Pi^*}[e^{\gamma \sum_{i=0}^N c(x_i, \Pi^*(x_i)) - \gamma\lambda \sum_{i=0}^N t(x_i, \Pi^*(x_i))}] \geq (\inf_{x \in S-\theta} E[e^{\gamma c(x,a)}])^N \cdot P_x^{\Pi_\gamma^\theta}[\tau_\theta \geq N].$$

By (4.9), the right hand side must be finite $\forall N$. Since $\inf_{x \in S-\theta} E[e^{\gamma c(x,a)}] > 0$, that requires that $P_x^{\Pi_\gamma^\theta}[\tau_\theta < \infty] = 1$.

Therefore, we can substitute the stopping time τ_θ in place of N in (4.3). We get the desired result, that Π_γ^θ is optimal and that it has value function $v(x)$.

The claim is proved.

Finally, because $v(x)$ is optimal, we have that

$$E_x^\Pi[e^{\gamma \sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)}] = v(x) \leq E_x^{\Pi^{x \to \theta}}[e^{\gamma \sum_{k=0}^{\tau_\theta} c(x_k, a_k)}] < \infty.$$

□

The following Theorem is the *major result of this chapter and the foundational result for the later developments* in this thesis. It means that if Assumption 6.1.1(γ) holds, then there is a single stationary, Markov, deterministic policy that is optimal from any initial state and the value of the objective function under that policy is the same at any initial state. It corresponds to the classical risk neutral average costs result (see, e.g., [35]) that if a chain is communicating, then there is a single stationary, Markov deterministic policy that is optimal and the value of the objective function is the same at any state.

Theorem 6.1.1 *Suppose that assumptions 3.4.1, 4.3.1, 4.3.2, and 4.3.3 hold. Suppose furthermore that assumption 6.1.1(γ) holds for all $\gamma < \bar{\gamma}$. Then, for any $\gamma < \bar{\gamma}$, there exists a stationary, Markov, deterministic policy Π_* such that*

$$\lambda^* = \lim_{T\to\infty} \frac{1}{\gamma T} \ln E_{x_0}^{\Pi_*}[e^{\int_{t=0}^{T} g(t)dt}] \leq \lim_{T\to\infty} \frac{1}{\gamma T} \ln E_{x_0}^{\Pi}[e^{\int_{t=0}^{T} g(t)dt}];\ \forall \Pi \in \Pi^{HR}, x_0 \in S.$$

Proof:

Let $\gamma < \bar{\gamma}$ be given.

By Lemma 4.3.1, if a policy does not enter M in finite time w.p.1, then it has infinite average cost. So let us consider the policies that induce a nonempty M and enter it in finite time w.p.1 from any initial state. Let $\rho(x) = \{\Pi \in \Pi^{HR} | x \in M(\Pi)\}$.

For $\Pi \in \rho(x)$, define $AC_x(\Pi) = \inf(\lambda \in \Re | E_x^{\Pi} e^{\sum_{k=0}^{\tau_x - 1} \gamma\{c(x_k,a_k) - \lambda t(x_k,a_k)\}} \leq 1)$.

Define $H(x) = \inf_{\Pi \in \rho(x)} AC_x(\Pi)$.

Claim:

If $\inf_{x\in S} H(x) < \infty$, then it is achieved.

Proof of Claim:

Select an arbitrary $x^{'} \in S$ such that $H(x^{'}) < \infty$. We know that any policy that achieves $AC_x(\Pi) < H(x^{'})$ must induce a recurrent class that has a nonempty intersection with the set $Z_{H(x^{'})}$, where

$$Z_a = \{x | \min_{a\in\alpha(x)} E[e^{\gamma(c(x,a) - a\cdot t(x,a))}] \leq 1\}.$$

Therefore, $\inf_{x\in S} H(x) = \inf_{x \in Z_{H(x^{'})}} H(x)$.

By Assumption 4.3.2, $Z_{H(x')}$ has finitely many elements, so the infimum is achieved.

And the claim is proved.

Let $\theta = \arg\min_{x \in S} H(x)$. And define $\lambda^* \doteq H(\theta)$.

By definition of $H(x)$, $\exists$ a sequence of policies $\{\Pi_i\}_{i=1}^{\infty}$ such that $\lim_{i\to\infty} AC_\theta(\Pi_i) = \lambda^*$.

Furthermore, by the logic contained in the proof of the above claim, each policy Π_i must induce a positive recurrent class that has a nonempty intersection with Z_a for some a large enough. Let us say that policy Π_i induces a positive recurrent class containing $x_i^{'} \in Z_a$. Therefore, by definition of $AC_x(\Pi)$ and Lemma 5.6.1,

$$E_{x_i^{'}}^{\Pi_i}[e^{\sum_{k=0}^{\tau_{x_i^{'}}-1} \gamma\{c(x_k,a_k)-AC_\theta(\Pi_i)t(x_k,a_k)\}}] \leq 1.$$

We will now define a stationary, Markov, deterministic policy $\Pi_i^{'}$ such that

$$E_{x_i^{'}}^{\Pi_i^{'}}[e^{\sum_{k=0}^{\tau_{x_i^{'}}-1} \gamma\{c(x_k,a_k)-AC_\theta(\Pi_i)t(x_k,a_k)\}}] \leq 1. \tag{6.4}$$

Let

$$V^{AC_\theta(\Pi_i)}(x,a) \doteq \inf_{\Pi \in \Pi^{HR}} E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_{x_i^{'}}-1} \gamma\{c(x_k,a_k)-AC_\theta(\Pi_i)t(x_k,a_k)\}} | a(0) = a],$$

and let

$$v^{AC_\theta(\Pi_i)}(x) = \inf_{a \in \alpha(x)} V^{AC_\theta(\Pi_i)}(x,a). \tag{6.5}$$

We have

$$V^{AC_\theta(\Pi_i)}(x,a)$$

$$= E[e^{\gamma\{c(x_k,a_k)-AC_\theta(\Pi_i)t(x_k,a_k)\}}] \cdot \sum_{z \in r(x,a)} P(z|x,a) \max\{I(z=\theta), v^{AC_\theta(\Pi_i)}(z)\}. \tag{6.6}$$

Since the transition law $P(x|s, a)$ is continuous in a for a fixed s and by Assumption 4.3.3, we see that $V^{AC_\theta(\Pi_i)}(x, a)$ is a continuous function of a. Therefore, it achieves its infimum at $a^*(x)$. Define the policy $\Pi_i^{'}(x) \doteq a^*(x)\ \forall x$.

Claim:

$\Pi_i^{'}$ minimizes $E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_{x_i^{'}}-1}\{c(x_k,a_k)-AC_\theta(\Pi_i)t(x_k,a_k)\}}]$ over all policies for each $x \in S$.

Proof of Claim:

We have from (6.5) and (6.6) that

$$v^{AC_\theta(\Pi_i)}(x)$$

$$= \inf_{a \in \alpha(x)} E[e^{\gamma\{c(x_k,a_k)-AC_\theta(\Pi_i)t(x_k,a_k)\}}] \cdot \sum_{z \in r(x,a)} P(z|x, a) \max\{I(z = \theta), v^{AC_\theta(\Pi_i)}(z)\}. \tag{6.7}$$

And the infimum is achieved by $a^*(x) \doteq \Pi_i^{'}(x)$. It can be seen by an argument analogous to the proof of Lemma 4.2.1 that led to (4.3), that the following inequality holds:

$$E_x^{\Pi}[e^{\gamma \sum_{i=0}^{N} c(x_i,\Pi(x_i))-\gamma AC_\theta(\Pi_i)\sum_{i=0}^{N} t(x_i,\Pi(x_i))}]$$

$$\geq E_x^{\Pi}[\Pi_{i=0}^{N}\{\frac{v^{AC_\theta(\Pi_i)}(x_i)}{\int \max\{I(z = \theta), v^{AC_\theta(\Pi_i)}(y)\}P(dy|x_i, \Pi(x_i))}\}],$$

with equality for $\Pi = \Pi_i^{'}$.

If we substitute for N the stopping time τ_θ, we obtain (by recursive cancellation of the numerator and denominator on the right hand side of the inequality),

$$E_x^{\Pi}[e^{\gamma \sum_{i=0}^{\tau_\theta} c(x_i,\Pi(x_i))-\gamma AC_\theta(\Pi_i)\sum_{i=0}^{\tau_\theta} t(x_i,\Pi(x_i))}] \geq v^{AC_\theta(\Pi_i)}(x),$$

with equality for $\Pi = \Pi_i^{'}$. And the claim is proved.

Since $\Pi_i^{'}$ must do at least as well as Π_i, (6.4) holds.

Since $x_i^{'} \in Z_\theta \ \forall i$ and Z_a has finitely many elements, there must be a subsequence $\{o(i)\}$ such that $x_{o(i)}^{'} = \omega \ \forall i$.

By Tychonoff's Theorem (Theorem 4.3.1), $\{\Pi_{o(i)}^{'}\}$ has a limit point: the stationary, Markov, deterministic policy $\Pi_*^{'}$.

Because $\lambda_\gamma^{\Pi_i^{'}} \geq \lambda^* \ \forall i$, there exists a subsequence of $\{o(i)\}$, $\{h(i)\}$, such that $\lambda_\gamma^{\Pi^{'}_{h(i+1)}} \leq \lambda_\gamma^{\Pi^{'}_{h(i)}} \ \forall i$.

Therefore, by Fatou's lemma we see that

$$E_\omega^{\Pi_*^{'}} e^{\gamma \sum_{k=0}^{\tau_\omega} c(x_k,a_k) - \lambda^* t(x_k,a_k)}] \leq 1. \tag{6.8}$$

Claim:

$\exists \omega^{'} \in S$ such that

$$E_{\omega^{'}}^{\Pi_*^{'}}[\tau_{\omega^{'}}] < \infty$$

and

$$E_{\omega^{'}}^{\Pi_*^{'}} e^{\gamma \sum_{k=0}^{\tau_{\omega^{'}}} c(x_k,a_k) - \lambda^* t(x_k,a_k)}] \leq 1.$$

Proof:

If $P_\omega^{\Pi_*^{'}}[\tau_\omega < \infty] = 1$, then we know from (6.8) and Lemma 5.3.1 that $J_\omega^{\Pi_*^{'}} \leq \lambda^*$. Therefore by Lemma 5.3.5 we know that the equivalence class containing ω must be positive recurrent, and the claim follows by setting $\omega^{'} = \omega$.

Suppose $P_\omega^{\Pi_*^{'}}[\tau_\omega < \infty] < 1$. Therefore there must be a state $\omega^{'} \in S$ that is in a positive recurrent class induced by $\Pi_*^{'}$ and such that $P_\omega^{\Pi_*^{'}}[\tau_{\omega^{'}} < \infty] > 0$ and $P_{\omega^{'}}^{\Pi_*^{'}}[\tau_\omega < \infty] = 0$.

(6.8) then implies that

$$E_\omega^{\Pi_*^{'}} e^{\gamma \sum_{k=0}^{\tau_\omega} c(x_k,a_k) - \lambda^* t(x_k,a_k)} | \tau_{\omega^{'}} < \tau_\omega] < \infty.$$

Therefore we have

$$\infty > E_{\omega^{'}}^{\Pi_*^{'}} e^{\gamma \sum_{k=0}^{\tau_\omega} c(x_k,a_k) - \lambda^* t(x_k,a_k)}] = E_{\omega^{'}}^{\Pi_*^{'}} e^{\gamma \sum_{k=0}^{\infty} c(x_k,a_k) - \lambda^* t(x_k,a_k)}]$$

Just as in the proof of the sub-claim in Lemma 5.3.1, it can be shown that

$$E_{\omega^{'}}^{\Pi_*^{'}} e^{\gamma \sum_{k=0}^{\infty} c(x_k,a_k)-\lambda^* t(x_k,a_k)}]$$

$$= E_{\omega^{'}}^{\Pi_*^{'}} e^{\gamma \sum_{k=0}^{\tau_{\omega^{'}}-1} c(x_k,a_k)-\lambda^* t(x_k,a_k)}] \cdot E_{\omega^{'}}^{\Pi_*^{'}} e^{\gamma \sum_{k=0}^{\infty} c(x_k,a_k)-\lambda^* t(x_k,a_k)}].$$

Since this is less than ∞, we must have

$$E_{\omega^{'}}^{\Pi_*^{'}} e^{\gamma \sum_{k=0}^{\tau_{\omega^{'}}-1} c(x_k,a_k)-\lambda^* t(x_k,a_k)}] \leq 1,$$

and the claim is proved.

Let $QQ =$ the recurrent class of $\Pi_*^{'}$ containing $\omega^{'}$.

Clearly, $\lambda_{QQ}^{\Pi_*}(\gamma) = \lambda^*$, so by Lemma 5.3.1, $J_x^{\Pi_*} = \lambda^* \ \forall x \in QQ$.

In the proof of Lemma 5.3.1 it was shown that $\exists$ a function $V(x)$, bounded away from zero and finite for each $x \in S$, such that (5.7) holds (with $\theta \doteq \omega^{'}$). Define $W(x) = \ln V(x)$ in (5.7).

Define

$$\Pi_*(x) = \begin{cases} \Pi_*^{'}(x) \text{ if } x \in QQ \\ \Pi_\gamma^{\omega^{'}} \text{ if } x \notin QQ \end{cases}$$

where $\Pi_\gamma^{\omega^{'}}$ is as defined in Lemma 6.1.1.

Let us extend (5.7) by defining

$$W(x) \doteq \ln\{E_x^{\Pi_*}[e^{\sum_{k=0}^{\tau_\theta - 1} \gamma\{c(x,a)-\lambda^* t(x,a)\}}\}]; \ \forall x \in S,$$

with

$$e^{W(x)} = E[e^{\gamma\{c(x,\Pi_*^{'}(x))-\lambda^* t(x,\Pi_*(x))\}}] \int \max\{e^{W(y)}, I(y = \omega^{'})\} P(dy|x, \Pi(x)); \ \forall x \in S. \tag{6.9}$$

First, let us show that $W(x) < \infty \ \forall x \in S$. Because QQ is absorbing under Π_*, $W(x)$ is the same as the value given by (5.7) (with $\theta \doteq \omega^{'}$) for $x \in QQ$. For $x \in QQ^c$, we have

$$e^{W(x)} = E_x^{\Pi_*}[e^{\sum_{k=0}^{\tau_\theta - 1} \gamma\{c(x,a) - \lambda^* t(x,a)\}}]$$

$$= E_x^{\Pi_*}[e^{\sum_{k=0}^{\tau_{QQ} - 1} \gamma\{c(x,a) - \lambda^* t(x,a)\}} \cdot e^{\sum_{k=\tau_{QQ}}^{\tau_\theta - 1} \gamma\{c(x,a) - \lambda^* t(x,a)\}}]$$

$$= E_x^{\Pi_\gamma^\omega}[e^{\sum_{k=0}^{\tau_{QQ} - 1} \gamma\{c(x,a) - \lambda^* t(x,a)\}} \cdot E_{x_{\tau_{QQ}}}^{\Pi_*'}[e^{\sum_{k=\tau_{QQ}}^{\tau_\theta - 1} \gamma\{c(x,a) - \lambda^* t(x,a)\}} | \{x_0, x_1, ..., x_{\tau_{QQ}}\}]]$$

$$\leq E_x^{\Pi_\gamma^\omega}[e^{\sum_{k=0}^{\tau_{QQ} - 1} \gamma\{c(x,a) - \lambda^* t(x,a)\}} \cdot E_{x_{\tau_{QQ}}}^{\Pi_\gamma^\omega}[e^{\sum_{k=\tau_{QQ}}^{\tau_\theta - 1} \gamma\{c(x,a) - \lambda^* t(x,a)\}} | \{x_0, x_1, ..., x_{\tau_{QQ}}\}]]$$

$$= E_x^{\Pi_\gamma^\omega}[e^{\sum_{k=0}^{\tau_\theta - 1} \gamma\{c(x,a) - \lambda^* t(x,a)\}}]$$

$$\leq E_x^{\Pi_\gamma^\omega}[e^{\sum_{k=0}^{\tau_\theta - 1} \gamma c(x,a)}] < \infty.$$

Claim: $W(x)$ is bounded below over S.

Proof of Claim:

Suppose that $W(\cdot)$ is not bounded below. Define $CC = \{x | E[e^{\gamma\{c(x,\Pi_*(x)) - \lambda^* t(x,\Pi_*(x))\}}] \leq 1$. We know by (6.9) that if $x \in CC^c$, then $\exists y \in S$ such that $W(y) < W(x)$. Therefore, $\inf_{x \in S} W(x) = \inf_{x \in CC} W(x)$. And since CC has finitely many elements by Assumption 4.3.2, we see that the infimum is achieved. Since $W(\cdot)$ is not bounded below, there must be a $z \in CC$ such that $W(z) = -\infty$.

By definition of $W(\cdot)$, this means that

$$E_z^{\Pi_*}[e^{\sum_{k=0}^{\tau_{\omega'}} \gamma\{c(x,a) - \lambda^* t(x,a)\}} = 0.$$

By the norm-like cost assumption Assumption 4.2.3, we know that there is a lower bound $B > 0$ such that

$$E[e^{\gamma\{c(x,a) - \lambda^* t(x,a)\}} \geq B, \ \forall x, a.$$

Therefore

$$E_z^{\Pi_*}[e^{\sum_{k=0}^{\tau_{\omega'}}\gamma\{c(x,a)-\lambda^* t(x,a)\}} \geq E_z^{\Pi_*}[\Pi_{k=0}^{\tau_{\omega'}} B$$

$$= \sum_{k=0}^{\infty} B^k P_z^{\Pi_*}[\tau_{\omega'} = k],$$

and therefore

$$0 = \sum_{k=0}^{\infty} B^k P_z^{\Pi_*}[\tau_{\omega'} = k],$$

which means that $P_z^{\Pi_*}[\tau_{\omega'} = k] = 0 \ \forall k$, or $P_z^{\Pi_*}[\tau_{\omega'} < \infty] = 0$, which contradicts irreducibility of the embedded Markov chain, and the claim is proved.

And so we see that $WW(x) = \max\{W(x), \ln(I(x = \omega^{'}))\}$ satisfies the conditions of Corollary 4.2.1, and we get that

$$J_x^{\Pi_*} \leq \lambda^*; \ \forall x \in S.$$

Also, by definition of θ and λ^*, we know that for any policy $\Pi \in \Pi^{HR}$ and any $x \in S$, $AC_x(\Pi) \geq \lambda^*$. Therefore, $J_x^{\Pi} \geq \lambda^*$.

Therefore $\Pi_* \in \Pi^{MD}$ is an optimal stationary policy with cost λ^*, and the Theorem is proved.

□

Define the optimality inequality as

$$e^{W(x)} = \inf_{a \in \alpha(x)} E[e^{\gamma\{c(x,a)-\lambda^* t(x,a)\}}] \int \max\{e^{W(y)}, I(y = \omega^{'})\} P(dy|x,a); \ \forall x \in S. \tag{6.10}$$

Corollary 6.1.1 *Suppose that assumptions 3.4.1, 4.3.1, 4.3.2, and 4.3.3 hold. Suppose furthermore that assumption 6.1.1(γ) holds for all $\gamma < \bar{\gamma}$.*

Then, for any $\gamma < \bar{\gamma}$, there exists a stationary, Markov, deterministic policy Π^* such that Π^* solves the optimality inequality (6.10) and

$$\lambda^* = \lim_{T\to\infty} \frac{1}{\gamma T} \ln E_{x_0}^{\Pi^*}[e^{\int_{t=0}^{T} g(t)dt}] \leq \lim_{T\to\infty} \frac{1}{\gamma T} \ln E_{x_0}^{\Pi}[e^{\int_{t=0}^{T} g(t)dt}]; \forall \Pi \in \Pi^{HR}, x_0 \in S.$$

Furthermore, if $W(\omega^{'}) = 1$, then then policy Π^* solves the dynamic program (4.2).

Proof:

We saw in the proof of Theorem 6.1.1 that Π_* achieves, starting from any state, the smallest cost that any policy can achieve. is satisfied. Furthermore, we know by definition of $\Pi_*^{'}$ that Π_* minimizes the optimality inequality (6.10) $\forall x \in QQ$. But Π_* may not minimize the optimality inequality $\forall x \in S$ (that is for $x \notin QQ$ – i.e. there may be a bias). So we define Π^* as the policy that minimizes the optimality inequality. The existence and properties of Π_* guarantee that Π^* exists and satisfies (6.10).

If $W(\omega^{'}) = 1$, then (6.10) reduces to (4.2)

□

Note: As shown in Chapter 5, $W(\omega^{'}) = 1$ if any of the following conditions are met:

1. $|S| < \infty$.
2. The round trip cost (of policy Π^* to $\omega^{'}$) at $\lambda = 0$ is finite.
3. The Semi-Markov chain induced by Π^* is of Type I.

6.2 Borkar's Convex Analytic Approach

In Section 5.3 of [1] (see also references cited therein), section 5.7 of [19], and [26]; Borkar's approach is described. Under a simple continuity assumption in the transition kernel and the assumption of norm-like costs (see Assumption 4.3.2) and under the (major) assumption that all policies have S as their sole recurrent class, there is a stationary, Markov risk neutral optimal policy, and furthermore that policy is sample path optimal. But the strong irreducibility assumption can be removed. In particular, as shown by Lasserre in [26], if that assumption is completely removed, then *there is an initial state x_0 and a stationary Markov policy Π_{sp} such that the optimal average risk neutral cost starting from any state under any policy is achieved w.p.1 by every sample path starting from x_0 under policy Π_{sp}.* To fully understand this (italicized) statement, it would be helpful to read the rest of this thesis. In particular, the reader is referred to Subsection 10.1.1.

Let us look again at Theorem 6.1.1. All that was added (except assumptions to handle the problems induced by covering the semi-Markov case) was Assumption 6.1.1. This assumption, in a sense, puts the risk sensitive problem on the same footing as the risk neutral problem. This is because in the risk neutral objective function, one need not worry about a state which takes finite expected cost just to transition out of the state. Other than that assumption, our strong result (Theorem 6.1.1) has assumptions no stronger than Borkar's and Lasserre's results.

Chapter 7

Reachability and Probabilistic Reachability

In this chapter we do not concern ourselves with costs, but with reachability: is there a policy that can take the system from state x to state y w.p.1? Barring that, is there one that can do it with nonzero probability? Also, for a given policy, how do the states communicate?

Define the logical relationship $R(C, D)$; $C, D \subset S$ to be true if $\forall x \in C$, $\exists \Pi_x^{'} \in \Pi^{HR}$ such that $P_x^{\Pi_x^{'}}[\tau_D < \infty] = 1$. For simplicity, if either C or D contains only one element, we may substitute that element in the notation. For example, if $C = \{x\}$ and $D = \{y\}$, we have that $R(x, y)$ is true if $\exists \Pi^{'} \in \Pi^{HR}$ such that $P_x^{\Pi^{'}}[\tau_y < \infty] = 1$. Also, define the logical relationship $R^{\Pi}(C, D)$ to be true if $P_x^{\Pi}[\tau_D < \infty] = 1 \ \forall x \in C$.

Lemma 7.0.1 *If $R(x, y)$ and $R(y, z)$, then $R(x, z)$.*

Proof:

We define $\Pi_{x \to z}$ as the policy which follows $\Pi_{x \to y}$ until y is reached and $\Pi_{y \to z}$ thereafter.

Note: Lemma 7.0.1 also holds true if x, y, and/or z are sets rather than single elements of the state place.

Define the set of all self-reachable states SR as $x \in SR$ iff $R(x, x)$. Note that Lemma 7.0.1 holds true if $R(\cdot, \cdot)$ is replaced by $R^{\Pi}(\cdot, \cdot)$. Define SR^{Π} as $x \in SR^{\Pi}$ if $R^{\Pi}(x, x)$. Define the relation $\sim$ on SR as $x \sim y$ iff $R(x, y)$ and $R(y, x)$.

Lemma 7.0.2 $\sim$ *is an equivalence relation.*

Proof:

$\sim$ is reflexive because it is defined on SR.

$\sim$ is symmetric by definition.

$\sim$ is transitive by Lemma 7.0.1.

Therefore, $X = SR \cup SR^c$. Furthermore, SR is the union of (at most countably many) disjoint equivalence classes under $\sim$.

Definition 7.0.1 *An equivalence class* $\aleph$ *under* $\sim$*, i.e. a set such that for some* $x \in S$, $\aleph = \{y \in S | x \sim y\}$ *is called a 'strongly communicating class'.*

Assumption 7.0.1 *If not* $R(x, y)$*, then* $\forall$ *policies* Π, $P^{\Pi}[s(t_2) = y | s(t_1) = x] = 0$ $\forall$ *times* t_1, t_2 *such that* $t_1 < t_2$.

In other words, under Assumption 7.0.1, the system can either reach y from x in finite expected time under some policy or with probability 1, or the system will not reach y from x under any policy. In the development to follow, we will not be assuming Assumption 7.0.1, because it is a restrictive (although convenient) assumption.

Assumption 7.0.2 $\forall s \in S$, $R(s, SR)$.

In other words, every state can reach SR.

Note: if both Assumption 7.0.1 and Assumption 7.0.2 are true, then $\forall s \in S$, $\exists x \in SR$ such that $R(s, x)$.

For $x \in SR$, denote the strongly communicating class containing x as $\aleph(x)$.

Lemma 7.0.3 *if $z \in \aleph(x)$, $w \in \aleph(y)$, and $R(x, y)$, then $R(z, w)$.*

Proof:

Repeated application of Lemma 7.0.1 is sufficient.

An immediate corollary to Lemma 7.0.3 is that for $x, y \in SR$, $R(x, y)$ implies $R(\aleph(x), \aleph(y))$.

If $x, y \in SR$, $R(x, y)$ and not $R(y, x)$, then we denote $\aleph(x) \preceq \aleph(y)$. Also, define $\aleph(x) \preceq \aleph(x)$.

Lemma 7.0.4 *$\preceq$ is a partial ordering on the strongly communicating classes induced by $\sim$ on SR.*

Proof:

We are given that $x, y, z \in SR$ and none are in the same strongly communicating class.

$R(x, y)$ and $R(y, z)$ implies $R(x, z)$ by Lemma 7.0.1. Also, $R(x, z)$ implies not $R(z, x)$ since x and z are in different strongly communicating classes.

Therefore, $\aleph(x) \preceq \aleph(y)$ and $\aleph(y) \preceq \aleph(z)$ implies $\aleph(x) \preceq \aleph(z)$.

Also, $\preceq$ is reflexive by definition.

□

Lemma 7.0.5 *If Assumption 7.0.1 and Assumption 7.0.2 hold and $\nexists y$ such that $\aleph(x) \preceq \aleph(y)$, then the set $\aleph(x)$ is invariant, i.e., under any policy Π and any $z \in \aleph(x)$, $P_z^{\Pi}[x_n \in \aleph(x)] = 1 \ \forall n > 0$.*

Proof: Suppose that the conclusion is false, i.e., suppose that $\exists z \in \aleph(x)$, $y \notin \aleph(x)$ and $a \in \alpha(z)$ such that $P(y|z, a) > 0$. Therefore, by Assumption 7.0.1, $R(z, y)$.

If $y \in SR$, then we have $\aleph(x) \preceq \aleph(y)$, which contradicts an assumption. If $y \notin SR$, then by Assumption 7.0.2 combined with Assumption 7.0.1, $\exists w \in SR$ s.t. $R(y, w)$. If $w \in \aleph(x)$, then $y \in \aleph(x)$, contradicting our supposition. If not, then $\aleph(x) \preceq \aleph(w)$, contradicting an assumption.

Define T, the set of all transient states, as follows: $x \in T$ if $\forall \Pi$, $P_x^\Pi[x_k \to \infty] > 0$, i.e. if not $R(x, SR)$. T is empty iff **(J3)** is true. Also, T is empty if S is finite.

Define $x < y$ if $R(x, y)$ but not $R(y, x)$. For $x, y \in SR$, we see that $x < y$ iff $\aleph(x) \preceq \aleph(y)$.

Define $R^p(x, y)$ to be true if $R(x, y)$ is not true and $\exists \Pi \in \Pi^{HR}$ such that $P_x^\Pi[\tau_y < \infty] > 0$. It is interesting to note that cases can be constructed in which $R^p(x, y)$ is true and

$$\sup_{\Pi \in \Pi^{HR}} P_x^\Pi[\tau_y < \infty] = 1.$$

Clearly, $R^p(x, y)$ implies not $R(x, y)$. Also, the following condition holds iff Assumption 7.0.1 is not true: $\exists s, y \in S$ such that $R^p(x, y)$.

We say that $x \to y$ is true if $\exists \Pi \in \Pi^{HR}$ such that $x \overset{\Pi}{\to} y$ is true. By definition, we have that $x \to y$ iff either $R(x, y)$ or $R^p(x, y)$.

We say that $R^{'}(x, C)$ is true if $R(x, C)$ is true and $\forall D \subset C$ such that $D \neq C$, $R(x, D)$ is not true. Clearly, if $R^{'}(x, C)$ is true, then $R^p(x, D)$ is true for all proper subsets D of C. Also, define $R^{'\Pi}(x, C)$ to be true if $R^\Pi(x, C)$ is true and $\forall D \subset C$ such that $D \neq C$, $R^\Pi(x, D)$ is not true. And define $R^{p\Pi}(x, y)$ to be true if $0 < P_x^\Pi[\tau_y < \infty] < 1$. Finally, define $R^{''}(x, A)$ to be true if $\exists \Pi \in \Pi^{HR}$ such that $R^{'\Pi}(x, A)$ is true. Clearly, $R^{''}(x, A)$ implies $R(x, A)$.

Lemma 7.0.6 *If $R(x, C)$ and $|C| < \infty$, then $\exists D \subset C$ such that $R^{'}(x, D)$.*

Proof:

The following procedure will construct $D \subset C$ with the desired property:

1. Set $D_0 = C$ and $i = 0$.

2. If there is an element $z \in D_i$ such that $R(x, D_i - \{z\})$ is true, then set $D_{i+1} = D_i - \{z\}$ and increment i. If there is no such element z, then set $D = D_i$ and terminate the procedure.

3.

Repeat step **2**.

Step 2 guarantees that $R^{'}(x, D)$. The procedure is guaranteed to terminate since $|C| < \infty$.

□

Lemma 7.0.7 *If $R^{'}(x, D)$, then $\forall y \in D$, not $R(y, D - \{y\})$.*

Proof:

Suppose $\exists y \in D$ such that $R(y, D - \{y\})$. Then $\exists \Pi_1$ such that $R^{\Pi_1}(x, D)$ and $\exists \Pi_2$ such that $R^{\Pi^2}(y, D - \{y\})$. Define policy Π^3 as the policy that follows policy Π_1 until τ_y and policy Π_2 is afterwards. We then have $R(x, D - \{y\})$, which contradicts $R^{'}(x, D)$.

□

Lemma 7.0.8 *If $R^{\Pi}(x, C)$, then $\exists D \subset C$ such that $R^{'\Pi}(x, D)$.*

Proof:

Order the states in C from 1 to N (i.e., $\{z_i\}_{i=1}^{N}$), where $N = \infty$ if $|C| = \infty$. We construct D according to the following procedure:

1. $F_0 = C; j = 0$

2. If $R^{\Pi}(x, F_j - \{z_j\})$, then set $F_{j+1} = F_j - \{z_j\}$. Otherwise, set $F_{j+1} = F_j$.

3. Increment j. If $j \leq |C|$, then go to step **2**.

If $|C| < \infty$, then the procedure terminates when $j = |C| + 1$. Set $D = F_{|C|+1}$.

If $|C| = \infty$, then the procedure does not terminate. Define $D \subset C$ as follows: $z_i \in D$ if $z_i \in F_{i+1}$. (I.e., $D = \lim_{i \to \infty} F_i$.)

It is clear that $D \neq \emptyset$ in either case.

Claim:

$$P_x^{\Pi}[\tau_D < \infty] = 1.$$

Proof of claim:

If $|C| < \infty$, then the claim is evident by inspection of step **2**. If $|C| = \infty$, then

$$P_x^{\Pi}[\tau_D < \infty] = P_x^{\Pi}[\tau_{\lim_{i \to \infty} F_i} < \infty].$$

The procedure guarantees that $F_{i+1} \subset F_i \ \forall i < \infty$. Therefore,

$$P_x^{\Pi}[\tau_{F_i} < \infty] \geq P_x^{\Pi}[\tau_{F_{i+1}} < \infty].$$

Therefore, by the monotone convergence theorem, we have

$$P_x^{\Pi}[\tau_{\lim_{i \to \infty} F_i} < \infty] = \lim_{i \to \infty} P_x^{\Pi}[\tau_{F_i} < \infty].$$

Again, by inspection of step **2**, we know that $P_x^{\Pi}[\tau_{F_i} < \infty] = 1 \ \forall i < \infty$.

And the claim is proved.

Claim:

If $z_i \in D$ for some i, then $R^{\Pi}(x, D - \{z_i\})$ is not true.

Proof of claim:

By step **2** of the procedure for constructing D, we know that $R^{\Pi}(x, F_i - \{z_i\})$ is not true, which means that $P_x^{\Pi}[\tau_{F_i - \{z_i\}} < \infty] < 1$. Since $D \subset F_i$, we know that

$$P_x^{\Pi}[\tau_{D - \{z_i\}} < \infty] \leq P_x^{\Pi}[\tau_{F_i - \{z_i\}} < \infty] < 1.$$

And the claim is proved.

The Lemma follows directly from the two preceding claims.

□

Lemma 7.0.9 *If* $x \in D \subset S$ *and* $\exists \Pi \in \Pi^{HR}$ *such that* $R^{\Pi}(x, D)$ *and* $P_x^{\Pi}[\tau_D = \tau_x] < 1$*, then* $R(x, D - \{x\})$*.*

Proof:

If $R^{\Pi}(x, D - \{x\})$, then we are done. If not, then define $\Pi^{'}$ as the policy that follows Π until x is reached. Upon reaching x, the history is erased and Π is again followed. Each time that x is reached, the history is reinitialized and Π is again followed.

Since $P_x^{\Pi}[\tau_D = \tau_x] < 1$, eventually w.p.1 D will be reached before x. Therefore, $R^{\Pi^{'}}(x, D - \{x\})$ and we are done.

□

Lemma 7.0.10 *If* $A \subset B \subset S$*,* $R(A, x)$*, and* $\exists$ *policy* Π_1 *such that* $R^{\Pi_1}(x, B)$ *and* $P_x^{\Pi_1}[x_{\tau_B} \notin A] > 0$*; then* $R(x, B - A)$*.*

Proof:

Because $R(A, x)$, $\exists$ $\Pi_2 \in \Pi^{HR}$ such that $R^{\Pi_2}(y, x)$ $\forall y \in A$.

Define policy $\Pi^{'}$ as the policy that follows Π_1 until B is reached. Then, if $x(\tau_B) \in A$, follow policy Π_2 until x is reached. Then repeat.

Because $P^{\Pi_1}[x(\tau_B) \notin A] > 0$, policy $\Pi^{'}$ will eventually hit $B - A$, starting from x.

□

Chapter 8

The Not Strongly Communicating Case

An SMDP is called *not strongly communicating* if $\exists x, y \in S$ such that not $R(x, y)$. In a not strongly communicating SMDP, it is not necessarily the case that the optimal average (risk sensitive or risk neutral) cost is the same starting from any state. Furthermore, the value function in the not strongly communicating case exhibits minimax behavior. The following example illustrates both of these points:

Example 8.0.1

Suppose we have a Markov chain with 2 states, $S = \{x_1, x_2\}$. Each state transitions back to itself w.p.1, and the cost is 1 in state x_1 and 2 in state x_2. So the average cost for any value of γ is 1 from state x_1 and 2 from state x_2, i.e., $J(x_1) = 1$, $J(x_2) = 2$.

Now, let's introduce control to the situation. Suppose there is a third state, x_3, where there are 2 possible actions in state x_3, $\alpha(x_3) = \{a_1, a_2\}$. Suppose that $c(x_3, a_1) = 1.6$, $p(x_3|x_3, a_1) = 1$, $c(x_3, a_2) = 5$, and define $p_i \doteq p(x_i|x_3, a_2); i = 1, 2, 3$. Suppose that we set $p_i = \frac{1}{3}; i = 1, 2, 3$. Then, it is clear that the optimal risk-neutral policy is to choose action a_2 (called policy Π_2), and the expected average cost is $\mathcal{J}^{\Pi_2}(x_3) = 1.5$. Choosing action a_1 (policy Π_1) yields an average cost of $\mathcal{J}^{\Pi_1}(x_3) = 1.6$. If the cost criterion is risk-sensitive average cost with $\gamma > 0$, then again we see that the $J^{\Pi_1}(x_3) = 1.6$. If action a_2 is chosen, then there is a transient period during which the state remains x_3, followed by the MDP settling into either state x_1 or x_2 with equal probability. Even if we assume that the transient period in state x_3 does not raise the average cost (N.B.: this assumption is true for γ small enough, as will be shown later.) we see that the average risk-sensitive cost is

$$J^{\Pi_2}(x_3) = \lim_{N\to\infty} \frac{1}{\gamma N} \ln[\frac{1}{2}e^{\gamma \sum_{k=0}^{N} 1} + \frac{1}{2}e^{\gamma \sum_{k=0}^{N} 2}]. \tag{8.1}$$

Clearly, the term $\frac{1}{2}e^{\gamma \sum_{k=0}^{N} 1}$ becomes insignificant compared to the term $\frac{1}{2}e^{\gamma \sum_{k=0}^{N} 2}$ for N large, so we have

$$J^{\Pi_2}(x_3) = \lim_{N\to\infty} \frac{1}{\gamma N} \ln[\frac{1}{2}e^{\gamma \sum_{k=0}^{N} 2}] = 2.$$

It should be clear that as long as $p_2 > 0$, we will have $J^{\Pi_2}(x_3) = 2$ and the optimal policy will be Π^1. We can see that (8.1) is equivalent to

$$J^{\Pi_2}(x_3) = \max_{x \in r(x_3, a_2)} J(x).$$

So we see that in the not strongly communicating case, the control action chosen in a state which can lead to two or more states that can't reach each other is the action that minimizes the maximum of the average cost for any of the states reached in one step. This is a minimax behavior enforced by a risk-sensitive criterion with finite risk parameter.

8.1 On the classification of Markov chains

We classify SMDPs into two categories: *strongly communicating* and *not strongly communicating*. An SMDP is called strongly communicating if $R(x, y)\ \forall x, y$, and it is called not strongly communicating otherwise. Theorem 6.1.1 gives sufficient conditions for the existence of an optimal policy with cost independent of initial state for an infinite horizon, risk sensitive objective function for a strongly communicating SMDP. Theorem 9.1.1 gives sufficient conditions for the existence of an optimal policy for an infinite horizon, risk sensitive objective function for the general case including the not strongly communicating case. As already shown in Example 8.0.1, there is no guarantee in the not strongly communicating case that the optimal cost is independent of the initial state. We classify SMDPs by whether all states are reachable w.p.1 from all other states. The reason we do this is, as illustrated in Example 8.0.1, if a state is only probabilistically reachable from another state, a minimax rule applies. (I.e., the optimal controller will minimize the worst case strongly communicating class in which it can (with nonzero probability) end up.) This is different from the risk neutral case, in which an averaging rule applies to the value function.

Our classification scheme differs from the scheme Puterman uses for MDPs (see [35], P. 348). (Note: Puterman's scheme is directly comparable to ours since our classifications apply to the embedded Markov chain, and therefore apply to MDPs.) We will not use Puterman's classification scheme, but we describe it here and contrast it with our scheme. The reason our classification scheme is different than Puterman's is because he is concerned with risk neutral costs. He classifies MDPs in two ways ([35]):

1. On the basis of probabilistic reachability (i.e., the property $x \to y$).

2. On the basis of probabilistic reachability under any stationary policy (i.e., $\forall \Pi$, $x \overset{\Pi}{\to} y$).

The first property is important to Puterman because in the risk neutral case, each recurrent class that is probabilistically reachable from a given state under the optimal policy contributes to the long run costs starting from that state. (In contrast, minimaxing is used in the risk sensitive case we are concerned with.)

The second property is important to Puterman because if every policy induces a single recurrent class, certain classical analytical techniques become useable. By contrast, we are extending certain results to the case where a policy may induce multiple recurrent classes.

That said, here are Puterman's definitions ([35], P. 348):

An MDP is called

recurrent if $P^{\Pi}(x,y) > 0$ $\forall$ stationary $\Pi \in \Pi^{MD}; x, y \in S$.

unichain if $\forall$ stationary $\Pi \in \Pi^{MD}$, $PSR^{\Pi} \neq \emptyset$ and $\forall x \in PSR^{\Pi}$, $\Omega^{\Pi}(x) = PSR^{\Pi}$.

communicating if $x \to y$ $\forall x, y \in S$.

weakly communicating if $PSR \neq \emptyset$, PSR absorbing, and $\forall x \in PSR$, $\Omega(x) = PSR$.

multichain if $\exists$ a stationary $\Pi \in \Pi^{MD}$ such that $PSR^{\Pi} \neq \emptyset$ and $\exists x, y \in PSR^{\Pi}$ such that $\Omega^{\Pi}(x) \neq \Omega^{\Pi}(y)$ and both $\Omega^{\Pi}(x)$ and $\Omega^{\Pi}(y)$ are absorbing under Π.

From now on, we will use our own definitions, not Puterman's.

8.2 Optimal policies in the not strongly communicating case

This section, at over 30 pages, is the longest section in the thesis. It is also the heart of the thesis, where all the other results come together. The optimal policy for the strong communicating case determined in Theorem 6.1.1 is combined with the reachability (w.p.1) properties found in the last chapter. The sequence of lemmas and increasingly complex notation in this section culminates in final result of this section, Theorem 8.2.1, which is a strong and nontrivial result. Theorem 8.2.1 completely eliminates the irreducibility assumption and finds an optimal policy for a given initial state. This theorem is the centerpiece of the thesis.

The following assumption guarantees the existence of optimal policies in the not strongly communicating case as will be shown later in the strong results Theorem 8.2.1 and Theorem 9.1.1. In words, Assumption 8.2.1γ means that if a set can be reached with probability 1, then it can be reached with finite expected risk sensitive cost when the risk sensitivity parameter is γ (or less).

Assumption 8.2.1 (γ) *If $R(x, C)$ is true, then $\exists \Pi \in \Pi^{HR}$ such that $R^{\Pi}(x, C)$ is true and*

$$E_x^{\Pi}[e^{\sum_{k=0}^{\tau_C} \gamma c(x_k, a_k)}] < \infty.$$

Lemma 8.2.1 *If $x, y \in SR$, $y \in \aleph(x)$, and Assumption 8.2.1(γ) is true, then*

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \inf_{\Pi \in \Pi^{HR}} J_y^{\Pi}.$$

Proof:

Suppose that $\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} < \inf_{\Pi \in \Pi^{HR}} J_y^{\Pi}$. Then, $\exists$ a policy Π_x such that $\forall \Pi \in \Pi^{HR}$, $J_x^{\Pi_x} < J_y^{\Pi}$.

Since $y \in \aleph(x)$, we know that $R(y, x)$ is true. So by Assumption 8.2.1(γ), $\exists \Pi_{y \to x} \in \Pi^{HR}$ such that $E_y^{\Pi_{y \to x}}[e^{\sum_{k=0}^{\tau_x} \gamma c(x_k, a_k)}] \doteq F < \infty$.

Define policy Π_y as follows. (Where $\Pi_y(k)(x)$ is the action taken if the system is in state x at the k^{th} decision epoch.)

$$\Pi_y(k) = \begin{cases} \Pi_{y \to x}(x) \text{ if } k < \tau_x \\ \Pi_x \text{ if } k \geq \tau_x \end{cases}$$

where the policy Π_x begins its history at time τ_x. I.e., the behavior of policy Π_y subsequent to reaching state x does not depend on how x was reached.

We have

$$J_y^{\Pi_y} = \limsup_{T \to \infty} \frac{1}{\gamma T} \ln E_y^{\Pi_y}[e^{\int_0^T \gamma g(t) dt}]$$

$$= \limsup_{T\to\infty} \frac{1}{\gamma T} \ln\{E_y^{\Pi_y}[e^{\int_0^{T\vee T_x} \gamma g(t)dt} \cdot E_x^{\Pi_x}[e^{\int_{T\vee T_x}^{T} \gamma g(t)dt} | T_x]]\}$$

$$\leq \limsup_{T\to\infty} \frac{1}{\gamma T} \ln\{E_y^{\Pi_y}[e^{\int_0^{T\vee T_x} \gamma g(t)dt} \cdot E_x^{\Pi_x}[e^{\int_0^{T} \gamma g(t)dt}]]\}$$

$$= \limsup_{T\to\infty} \frac{1}{\gamma T} \ln\{E_y^{\Pi_y}[e^{\int_0^{T\vee T_x} \gamma g(t)dt}] \cdot E_x^{\Pi_x}[e^{\int_0^{T} \gamma g(t)dt}]\}$$

$$\leq \limsup_{T\to\infty} \frac{1}{\gamma T} \ln\{E_y^{\Pi_y}[e^{\int_0^{T_x} \gamma g(t)dt}] \cdot E_x^{\Pi_x}[e^{\int_0^{T} \gamma g(t)dt}]\}$$

$$= \limsup_{T\to\infty} \frac{1}{\gamma T} \ln\{F \cdot E_x^{\Pi_x}[e^{\int_0^{T} \gamma g(t)dt}]\}$$

$$= \limsup_{T\to\infty} \frac{1}{\gamma T} \ln E_x^{\Pi_x}[e^{\int_0^{T} \gamma g(t)dt}] = J_x^{\Pi_x},$$

where for $a, b \in \Re$, we define

$$a \vee b = \begin{cases} a \text{ if } a < b \\ b \text{ if } a \geq b \end{cases}$$

and T_x is the first hitting time of state x in continuous time, i.e. $T_x = \sum_{k=0}^{\tau_x - 1} t(x, a)$.

So we have $J_y^{\Pi_y} \leq J_x^{\Pi_x}$, which is a contradiction.

□

For $x \in SR$, define $\beta(x) \subset \alpha(x)$ as follows:

$a \in \beta(x)$ iff $R(x, a) \subset \aleph(x)$.

From the definition of $\aleph(x)$ it can be deduced that $\beta(x) \neq \emptyset\ \forall x \in SR$.

For $x \notin SR$, define $\beta(x) = \alpha(x)$.

Let us define a new SMDP, called the *restricted SMDP*, by restricting allowable actions in state x to those contained in $\beta(x)$. We denote the restricted SMDP by putting a $^-$ over the P or E operator. If $x \in SR$, then $\forall \Pi \in \Pi^{HR}$ and all $0 \leq k < \infty$, $\bar{P}_x^{\Pi}[x_k \in \aleph(x)] = 1$.

Because each strongly communicating class in SR is communicating, we know from Theorem 6.1.1 that if assumptions 3.4.1, 4.3.1, 4.3.2, and 4.3.3 hold and assumption 6.1.1(γ) (Note: 6.1.1(γ) is a consequence of Assumption 8.2.1(γ)) holds for all $\gamma < \bar{\gamma}$ with $S \doteq \aleph(x)$ for the *restricted SMDP*, then for any $\gamma < \bar{\gamma}$, there exists a stationary, Markov, deterministic policy $\Pi^*_{\aleph(x)}$ and a constant $0 < \lambda^*_{\aleph(x)} < \infty$ such that

$$\lambda^*_{\aleph(x)} = \lim_{T\to\infty} \frac{1}{\gamma T} \ln \bar{E}_{x_0}^{\Pi^*_{\aleph(x)}}[e^{\int_{t=0}^{T} g(t)dt}] \leq \lim_{T\to\infty} \frac{1}{\gamma T} \ln \bar{E}_{x_0}^{\Pi}[e^{\int_{t=0}^{T} g(t)dt}];\ \forall \Pi \in \Pi^{HR}, x_0 \in \aleph(x).$$

Furthermore, if we define $W(x)$ as in (5.13) with $\Pi = \Pi^*_{\aleph(x)}$, then policy $\Pi^*_{\aleph(x)}$ solves the dynamic program (4.2) for all $s \in S \doteq \aleph(x)$.

The following Lemma is clearly true:

Lemma 8.2.2 *For $x \in SR$,*

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} \leq \lambda^*_{\aleph(x)}.$$

Proof:

See the preceding discussion.

Assumption 8.2.2 *$\forall x \in S$, $\exists D \subset SR$ such that $|D| < \infty$ and $R(x, D)$.*

For $\Pi^{'} \in \Pi^{HR}$ and $A \subset S$, we define the set of restricted policies $\Pi^r(\Pi^{'}, A)$ as follows:

$\Pi \in \Pi^r(\Pi^{'}, A)$ if $\Pi(k) = \Pi^{'}(k)\ \forall k < \tau_A$.

I.e., $\Pi \in \Pi^r(\Pi^{'}, A)$ if it is identical to $\Pi^{'}$ prior to the first hitting time of set A.

We define $\Pi^R(x, A)$ as follows:

$$\Pi^R(x, A) = \cup_{\{\Pi | R^\Pi(x,A)\}} \Pi^r(\Pi, A).$$

I.e., $\Pi \in \Pi^R(x, A)$ if $P_x^\Pi[\tau_A < \infty] = 1$.

Lemma 8.2.3 *Let $D \subset SR$, and let $R(x, D)$ and Assumption 8.2.1(γ) be true. Then,*

$$\inf_{\Pi \in \Pi^R(x,D)} J_x^\Pi \leq \sup_{y \in D} \inf_{\Pi \in \Pi^{HR}} J_y^\Pi.$$

Proof:

Assumption 8.2.1(γ) assures us of the existence of a $\Pi^{'} \in \Pi^R(x, D)$ such that

$$E_x^{\Pi^{'}}[e^{\sum_{k=0}^{\tau_D} \gamma c(x_k,a_k)}] \doteq F < \infty.$$

Since $\Pi^{'} \in \Pi^R(x, D)$, we know that $P_x^{\Pi^{'}}[\tau_D < \infty] = 1$. Let

$$F_y \doteq \begin{cases} E_x^{\Pi^{'}}[e^{\sum_{k=0}^{\tau_D} \gamma c(x_k,a_k)} | \tau_D = \tau_y] \text{ if } P_x^{\Pi^{'}}[\tau_D = \tau_y] > 0 \\ 0 \text{ if } P_x^{\Pi^{'}}[\tau_D = \tau_y] = 0 \end{cases}.$$

Therefore, we have $F = \sum_{y \in D} P_x^{\Pi^{'}}[\tau_D = \tau_y] F_y$. So clearly, $F_y < \infty \ \forall y \in D$. For each $y \in D$, let $\{\Pi_m^y\}_{m=1}^\infty$ be a sequence of policies such that

$$\lim_{m \to \infty} J_y^{\Pi_m^y} = \inf_{\Pi \in \Pi^{HR}} J_y^\Pi.$$

Define $\Pi_m^{'} \in \Pi^r(\Pi^{'}, D)$ as the policy that follows Π_m^y (and erases the history – i.e., starts fresh with no history) upon reaching $y \in D$ if $\tau_y = \tau_D$. Since $\Pi^{'}$ takes the system to D w.p.1, $\Pi_m^{'}$ is well defined.

Claim:

$$\lim_{m \to \infty} J_x^{\Pi_m^{'}} \leq \sup_{y \in D} \inf_{\Pi \in \Pi^{HR}} J_y^\Pi.$$

Proof of claim:

Sub-claim:

If $P_x^{\Pi'}[\tau_y = \tau_D] > 0$, then

$$\lim_{T\to\infty} \frac{1}{\gamma T} \ln\{E_x^{\Pi'_m}[e^{\int_{t=0}^{T} \gamma g(t)dt} | T_y = T_D]\} \leq \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{E_y^{\Pi_m^y}[e^{\int_{t=0}^{T} \gamma g(t)dt}]\}.$$

Proof of Sub-claim:

$$\lim_{T\to\infty} \frac{1}{\gamma T} \ln\{E_x^{\Pi'_m}[e^{\int_{t=0}^{T} \gamma g(t)dt} | T_y = T_D]\}$$

$$= \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{E_x^{\Pi'_m}[e^{\int_{t=0}^{T_y} \gamma g(t)dt} \cdot e^{\int_{t=T_y}^{T} \gamma g(t)dt} | T_y = T_D]\}$$

$$\leq \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{E_x^{\Pi'_m}[e^{\int_{t=0}^{T_y} \gamma g(t)dt} | T_y = T_D] \cdot E_y^{\Pi_m^y}[e^{\int_{t=0}^{T} \gamma g(t)dt}]\}$$

$$= \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{F_y \cdot E_y^{\Pi_m^y}[e^{\int_{t=0}^{T} \gamma g(t)dt}]\}$$

$$= \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{E_y^{\Pi_m^y}[e^{\int_{t=0}^{T} \gamma g(t)dt}]\},$$

where the last equality follows since $F_y < \infty$.

And the sub-claim is proved.

$$J_x^{\Pi'_m} = \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{E_x^{\Pi'_m}[e^{\int_{t=0}^{T} \gamma g(t)dt}]$$

$$= \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{\sum_{y\in D} P_x^{\Pi'_m}[\tau_y = \tau_D] \cdot E_x^{\Pi'_m}[e^{\int_{t=0}^{T} \gamma g(t)dt} | T_y = T_D]\}$$

$$\leq \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{\sum_{y\in D} P_x^{\Pi'_m}[\tau_y < \infty] \cdot \sup_{y\in D} E_x^{\Pi'_m}[e^{\int_{t=0}^{T} \gamma g(t)dt} | T_y = T_D]\}$$

$$= \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{\sup_{y\in D} E_x^{\Pi'_m}[e^{\int_{t=0}^{T} \gamma g(t)dt} | T_y = T_D]\}$$

$$= \sup_{y \in D} \lim_{T \to \infty} \frac{1}{\gamma T} \ln\{E_x^{\Pi'_m}[e^{\int_{t=0}^{T} \gamma g(t) dt} | T_y = T_D]\}$$

$$\leq \sup_{y \in D} \lim_{T \to \infty} \frac{1}{\gamma T} \ln\{E_y^{\Pi_m^y}[e^{\int_{t=0}^{T} \gamma g(t) dt}]\},$$

where the last inequality follows from the sub-claim.

The claim follows by taking the limit as $m \to \infty$.

The lemma follows directly from the claim since $\Pi' \in \Pi^R(x, D)$.

□

Lemma 8.2.4 *The following three implications are true:*

1. *If Assumption 4.2.1 (equivalently, Assumption 4.3.1) holds and*

$$\lim\sup_{N \to \infty} E_x^{\Pi}[e^{\sum_{k=0}^{N} \gamma\{c(x_k, a_k) - \lambda t(x_k, a_k)\}}] = 0,$$

(equivalently, $\limsup_{T \to \infty} e^{-\gamma \lambda T} E_x^{\Pi}[e^{\int_{t=0}^{T} g(t) dt}] = 0$*)*

then $J_x^{\Pi} \leq \lambda$.

2. *If Assumption 3.4.1 holds and*

$$\lim\sup_{N \to \infty} E_x^{\Pi}[e^{\sum_{k=0}^{N} \gamma\{c(x_k, a_k) - \lambda t(x_k, a_k)\}}] = \infty,$$

(equivalently, $\limsup_{T \to \infty} e^{-\gamma \lambda T} E_x^{\Pi}[e^{\int_{t=0}^{T} g(t) dt}] = \infty$*)*

then $J_x^{\Pi} \geq \lambda$.

3. *If Assumption 4.3.1 and Assumption 3.4.1 hold and*

$$0 < \lim\sup_{N \to \infty} E_x^{\Pi}[e^{\sum_{k=0}^{N} \gamma\{c(x_k, a_k) - \lambda t(x_k, a_k)\}}] < \infty,$$

then $J_x^{\Pi} = \lambda$.

Proof:

Assume that

$$\limsup_{N\to\infty} E_x^{\Pi}[e^{\sum_{k=0}^{N}\gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] = 0$$

is true. Label the time of the N^{th} decision epoch (in continuous time) t_N. We obtain

$$\limsup_{N\to\infty} E_x^{\Pi}[e^{-\gamma\lambda t_N}\cdot e^{\int_{t=0}^{t_N}\gamma g(t)dt}] = 0.$$

or equivalently

$$\limsup_{N\to\infty} E_x^{\Pi}[e^{-\gamma\lambda t_{N+1}}\cdot e^{\int_{t=0}^{t_{N+1}}\gamma g(t)dt}] = 0. \tag{8.2}$$

Assume that Assumption 4.2.1 is true. Therefore $E[e^{-\gamma\lambda t(x_{N+1},a_{N+1})}] > L > 0$. So we have

$$E_x^{\Pi}[e^{-\gamma\lambda t_{N+1}}\cdot e^{\int_{t=0}^{t_{N+1}}\gamma g(t)dt}$$

$$= E_x^{\Pi}[e^{-\gamma\lambda t_N}\cdot e^{\int_{t=0}^{t_{N+1}}\gamma g(t)dt}\cdot E[e^{-\gamma\lambda t(x_N,a_N)}|x_N]]$$

$$\geq E_x^{\Pi}[e^{-\gamma\lambda t_N}\cdot e^{\int_{t=0}^{t_{N+1}}\gamma g(t)dt}\cdot L].$$

Combining this with (8.2), we get

$$0 \geq \limsup_{N\to\infty} E_x^{\Pi}[e^{-\gamma\lambda t_N}\cdot e^{\int_{t=0}^{t_{N+1}}\gamma g(t)dt}\cdot L]$$

or, removing the L,

$$0 \geq \limsup_{N\to\infty} E_x^{\Pi}[e^{-\gamma\lambda t_N}\cdot e^{\int_{t=0}^{t_{N+1}}\gamma g(t)dt}].$$

Since $e^z \geq 0\ \forall z \in \Re$, this must hold with equality, so we obtain

$$\limsup_{N\to\infty} E_x^{\Pi}[e^{-\gamma\lambda t_N}\cdot e^{\int_{t=0}^{t_{N+1}}\gamma g(t)dt}] = 0. \tag{8.3}$$

Let $N(t)$ be the number of transitions that have occured prior to time t. We get

$$\limsup_{T\to\infty} E_x^{\Pi}[e^{-\gamma\lambda T}\cdot e^{\int_{t=0}^{T}\gamma g(t)dt}]$$

$$\leq \limsup_{T\to\infty} E_x^{\Pi}[e^{-\gamma\lambda T_{N(t)}}\cdot e^{\int_{t=0}^{T_{N(t)+1}}\gamma g(t)dt}] = 0,$$

where the equality follows from (8.3). Again, since $e^z \geq 0\ \forall z \in \Re$, we must have that

$$\limsup_{T\to\infty} E_x^{\Pi}[e^{-\gamma\lambda T}\cdot e^{\int_{t=0}^{T}\gamma g(t)dt}] = 0$$

or

$$\limsup_{T\to\infty} e^{-\gamma\lambda T} E_x^{\Pi}[e^{\int_{t=0}^{T}\gamma g(t)dt}] = 0. \tag{8.4}$$

Now suppose that

$$J_x^{\Pi} = \limsup_{T\to\infty} \frac{1}{\gamma T}\ln\{E_x^{\Pi}[e^{\int_{t=0}^{T}\gamma g(t)dt}]\} > \lambda.$$

Then,

$$\limsup_{T\to\infty} (\frac{1}{\gamma T}\ln\{E_x^{\Pi}[e^{\int_{t=0}^{T}\gamma g(t)dt}]\} - \lambda) > 0.$$

Or,

$$\limsup_{T\to\infty} \frac{1}{\gamma T}(\ln\{E_x^{\Pi}[e^{\int_{t=0}^{T}\gamma g(t)dt}]\} - \lambda\gamma T) > 0.$$

Or,

$$\limsup_{T\to\infty} \frac{1}{\gamma T}\ln\{E_x^{\Pi}[e^{\int_{t=0}^{T}\gamma g(t)dt}]\cdot e^{-\lambda\gamma T}\} > 0.$$

By (8.4), the term inside the natural log approaches 0 in the limit. Therefore the limit looks like $\frac{\ln(0)}{\infty} = \frac{-\infty}{\infty}$, so we know that it must be ≤ 0. This is a contradiction, and so implication **1** is proved.

Assume that

$$\limsup_{N\to\infty} E_x^{\Pi}[e^{\sum_{k=0}^{N}\gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] = \infty$$

is true. We obtain

$$\limsup_{N\to\infty} E_x^{\Pi}[e^{-\gamma\lambda t_N}\cdot e^{\int_{t=0}^{t_N}\gamma g(t)dt}] = \infty. \tag{8.5}$$

Assume that Assumption 3.4.1 is true. Therefore $E[e^{-\gamma\lambda t(x_{N+1},a_{N+1})}] < U < 0$. So we have

$$E_x^{\Pi}[e^{-\gamma\lambda t_{N+1}}\cdot e^{\int_{t=0}^{t_N}\gamma g(t)dt}]$$

$$= E_x^{\Pi}[e^{-\gamma\lambda t_N}\cdot e^{\int_{t=0}^{t_N}\gamma g(t)dt}\cdot E[e^{-\gamma\lambda t(x_N,a_N)}|x_N]]$$

$$< E_x^{\Pi}[e^{-\gamma\lambda t_N}\cdot e^{\int_{t=0}^{t_{N+1}}\gamma g(t)dt}\cdot U].$$

Combining this with (8.5), we get

$$\infty \le \limsup_{N\to\infty} E_x^{\Pi}[e^{-\gamma\lambda t_{N+1}}\cdot e^{\int_{t=0}^{t_N}\gamma g(t)dt}\cdot U]$$

or, removing the U,

$$\infty \le \limsup_{N\to\infty} E_x^{\Pi}[e^{-\gamma\lambda t_{N+1}}\cdot e^{\int_{t=0}^{t_N}\gamma g(t)dt}].$$

Clearly this holds with equality so we obtain

$$\limsup_{N\to\infty} E_x^{\Pi}[e^{-\gamma\lambda t_{N+1}}\cdot e^{\int_{t=0}^{t_N}\gamma g(t)dt}] = 0. \tag{8.6}$$

We get

$$\limsup_{T\to\infty} E_x^{\Pi}[e^{-\gamma\lambda T}\cdot e^{\int_{t=0}^{T}\gamma g(t)dt}]$$

$$\ge \limsup_{T\to\infty} E_x^{\Pi}[e^{-\gamma\lambda T_{N(t)+1}}\cdot e^{\int_{t=0}^{T_{N(t)}}\gamma g(t)dt}] = \infty,$$

where the equality follows from (8.6).

Since this equation must hold with equality, we get

$$\limsup_{T\to\infty} E_x^{\Pi}[e^{-\gamma\lambda T} \cdot e^{\int_{t=0}^{T} \gamma g(t)dt}] = \infty$$

or

$$\limsup_{T\to\infty} e^{-\gamma\lambda T} E_x^{\Pi}[e^{\int_{t=0}^{T} \gamma g(t)dt}] = \infty. \tag{8.7}$$

Now suppose that

$$J_x^{\Pi} = \limsup_{T\to\infty} \frac{1}{\gamma T} \ln\{E_x^{\Pi}[e^{\int_{t=0}^{T} \gamma g(t)dt}]\} < \lambda.$$

Then,

$$\limsup_{T\to\infty} (\frac{1}{\gamma T} \ln\{E_x^{\Pi}[e^{\int_{t=0}^{T} \gamma g(t)dt}]\} - \lambda) < 0.$$

Or,

$$\limsup_{T\to\infty} \frac{1}{\gamma T} (\ln\{E_x^{\Pi}[e^{\int_{t=0}^{T} \gamma g(t)dt}]\} - \lambda\gamma T) < 0.$$

Or,

$$\limsup_{T\to\infty} \frac{1}{\gamma T} \ln\{E_x^{\Pi}[e^{\int_{t=0}^{T} \gamma g(t)dt}] \cdot e^{-\lambda\gamma T}\} < 0.$$

By (8.7), the term inside the natural log approaches ∞ in the limit. Therefore the limit looks like $\frac{\ln(\infty)}{\infty} = \frac{\infty}{\infty}$, so we know that it must be ≥ 0. This is a contradiction, and so implication **2** is proved.

The proof of implication **3** is a simple extension of the proofs of the first two implications and is omitted for brevity.

□

Corollary 8.2.1 *If Assumption 3.4.1 and Assumption 4.3.1 hold,*

$$J_x^\Pi = \limsup_{T\to\infty} \frac{1}{\gamma T} \ln\{E_x^\Pi [e^{\int_{t=0}^{T} \gamma g(t)dt}] = \lambda,$$

and $\lambda_l, \lambda_u \in \Re$ are such that $\lambda_l < \lambda < \lambda_u$, then the following two equalities hold:

$$\limsup_{N\to\infty} E_x^\Pi [e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda_u t(x_k,a_k)\}}] = \limsup_{T\to\infty} e^{-\gamma\lambda T} E_x^\Pi [e^{\int_{t=0}^{T} g(t)dt}] = 0.$$

and

$$\limsup_{N\to\infty} E_x^\Pi [e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda_l t(x_k,a_k)\}}] = \limsup_{T\to\infty} e^{-\gamma\lambda T} E_x^\Pi [e^{\int_{t=0}^{T} g(t)dt}] = \infty.$$

Proof:

The corollary follows from Lemma 8.2.4 **3** and the fact that

$$\limsup_{N\to\infty} E_x^\Pi [e^{\sum_{k=0}^{N} \gamma\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]$$

is decreasing in λ.

□

Lemma 8.2.5 *If $x, y \in S$, $\Pi^{'} \in \Pi^{HR}$, and*

$$P_x^{\Pi^{'}}[\tau_y < \infty] > 0,$$

then

$$J_x^{\Pi^{'}} \geq \inf_{\Pi\in\Pi^{HR}} J_y^\Pi.$$

Proof:

$$J_x^{\Pi^{'}} = \lim_{T\to\infty} \frac{1}{\gamma T} \ln E_x^{\Pi^{'}} [e^{\gamma \int_{t=0}^{T} g(t)dt}]$$

$$= \lim_{T\to\infty} \frac{1}{\gamma T} \ln\{E_x^{\Pi^{'}} [e^{\gamma \int_{t=0}^{T} g(t)dt} | T_y < \infty] P_x^{\Pi^{'}} [T_y < \infty] +$$

$$E_x^{\Pi'}[e^{\gamma\int_{t=0}^{T} g(t)dt}|T_y=\infty]P_x^{\Pi'}[T_y=\infty]\}$$

$$\geq \lim_{T\to\infty}\frac{1}{\gamma T}\ln E_x^{\Pi'}[e^{\gamma\int_{t=0}^{T} g(t)dt}|T_y<\infty]$$

$$\geq \lim_{T\to\infty}\frac{1}{\gamma T}\ln E_x^{\Pi'}[e^{\gamma\int_{t=T_y}^{T} g(t)dt}|T_y<\infty]. \tag{8.8}$$

Denote $\lambda_y^{**} \doteq \inf_{\Pi\in\Pi^{HR}} J_y^{\Pi}$. By optimality of λ_y^{**} and Corollary Equiv-converse, the following holds $\forall\lambda<\lambda_y^{**}$

$$\lim_{T\to\infty} E_x^{\Pi'}[e^{-\gamma\lambda(T-T_y)}e^{\gamma\int_{t=T_y}^{T} g(t)dt}|T_y=K<\infty]=\infty.$$

Therefore,

$$\lim_{T\to\infty} E_x^{\Pi'}[e^{-\gamma\lambda T}e^{\gamma\int_{t=0}^{T} g(t)dt}|T_y<\infty]$$

$$=\lim_{T\to\infty} E_x^{\Pi'}[e^{-\gamma\lambda T_y}e^{\gamma\int_{t=0}^{T_y} g(t)dt}\cdot E_x^{\Pi'}[e^{-\gamma\lambda(T-T_y)}e^{\gamma\int_{t=T_y}^{T} g(t)dt}|T_y]|T_y<\infty]$$

$$=\lim_{T\to\infty} E_x^{\Pi'}[e^{-\gamma\lambda T_y}e^{\gamma\int_{t=0}^{T_y} g(t)dt}\cdot\infty|T_y<\infty],$$

which $=\infty$ unless

$$E_x^{\Pi'}[e^{-\gamma\lambda T_y}e^{\gamma\int_{t=0}^{T_y} g(t)dt}|T_y<\infty]=0.$$

So we have two cases:

Case 1:

$$\lim_{T\to\infty} E_x^{\Pi'}[e^{-\gamma\lambda T}e^{\gamma\int_{t=0}^{T} g(t)dt}|T_y<\infty]=\infty.$$

In this case, we have by Lemma 8.2.4 (**2**) that

$$\lim_{T\to\infty}\frac{1}{\gamma T}\ln E_x^{\Pi'}[e^{\gamma\int_{t=T_y}^{T} g(t)dt}|T_y<\infty]\geq\lambda_y^{**}.$$

And by (8.8), the lemma is satisfied in Case 1.

Case 2:

$$E_x^{\Pi'}[e^{-\gamma\lambda T_y}e^{\gamma\int_{t=0}^{T_y}g(t)dt}|T_y<\infty]=0.$$

In this case, we have $\forall\lambda<\lambda_y^{**}$ that

$$0=E_x^{\Pi'}[e^{-\gamma\lambda T_y}e^{\gamma\int_{t=0}^{T_y}g(t)dt}|T_y<\infty]$$

$$\geq E_x^{\Pi'}[e^{-\gamma\lambda T_y}|T_y<\infty].$$

If $T_y<\infty$ w.p.1, then $\exists Z<\infty$ such that $P[T_y<Z]>\frac{1}{2}$. But then no matter what Z is, we get

$$E_x^{\Pi'}[e^{-\gamma\lambda T_y}|T_y<\infty]>0.$$

This is a contradiction, and so the lemma is satisfied in case 2.

□

Corollary 8.2.2 *If $x,y\in S$, $\Pi'\in\Pi^{HR}$, and*

$$P_x^{\Pi'}[\tau_y<\infty]>0,$$

then

$$J_x^{\Pi'}\geq\lim_{T\to\infty}\frac{1}{\gamma T}\ln E_x^{\Pi'}[e^{\gamma\int_{t=0}^{T}g(t)dt}|T_y<\infty].$$

Proof:

$$E_x^{\Pi'}[e^{\gamma\int_{t=0}^{T}g(t)dt}]$$

$$=E_x^{\Pi'}[e^{\gamma\int_{t=0}^{T}g(t)dt}|T_y<\infty]\cdot P[T_y<\infty]+E_x^{\Pi'}[e^{\gamma\int_{t=0}^{T}g(t)dt}|T_y=\infty]\cdot P[T_y=\infty].$$

Therefore,

$$J_x^{\Pi'} = \lim_{T\to\infty} \frac{1}{\gamma T} E_x^{\Pi'}[e^{\gamma \int_{t=0}^{T} g(t)dt}] =$$

$$= \max\{\lim_{T\to\infty} \frac{1}{\gamma T} E_x^{\Pi'}[e^{\gamma \int_{t=0}^{T} g(t)dt} | T_y < \infty], \ \lim_{T\to\infty} \frac{1}{\gamma T} E_x^{\Pi'}[e^{\gamma \int_{t=0}^{T} g(t)dt} | T_y = \infty]\}$$

$$\geq \lim_{T\to\infty} \frac{1}{\gamma T} \ln E_x^{\Pi'}[e^{\gamma \int_{t=0}^{T} g(t)dt} | T_y < \infty].$$

□

Lemma 8.2.6 *Let $D \subset SR$, and let $R'(x, D)$ and Assumption 8.2.1(γ) be true. Then,*

$$\inf_{\Pi \in \Pi^R(x,D)} J_x^{\Pi} = \sup_{y \in D} \inf_{\Pi \in \Pi^{HR}} J_y^{\Pi}.$$

Proof:

By Lemma 8.2.3, left hand side $\leq$ right hand side.

By Lemma 8.2.5 and the fact that $P_x^{\Pi}[\tau_y < \infty] > 0 \ \forall y \in D$ (i.e., the fact that $R'(x, D)$ is true), left hand side $\geq$ right hand side.

□

Corollary 8.2.3 *Let $D \subset SR$. If $R(x, D)$ and Assumption 8.2.1(γ) are true and*

$$\inf_{\Pi \in \Pi^R(x,D)} J_x^{\Pi} < \sup_{y \in D} \inf_{\Pi \in \Pi^{HR}} J_y^{\Pi},$$

then not $R'(x, D)$, i.e., $\exists C \subset D$, $C \neq D$ such that $R(x, C)$.

We know that we can decompose SR as follows:

$SR = \cup_{i=1}^{Q} \aleph(s_i)$, where $s_i \notin \aleph(s_j)$ for $i \neq j$, and $0 \leq Q \leq \infty$. If $Q = \infty$, then there are countably many strongly communicating classes in SR. If $1 \leq Q < \infty$, then there are finitely many strongly communicating classes. If $Q = 0$, then all states are transient under any policy.

Each s_i is a representative of the strongly communicating class that contains it. Define the function $\upsilon : SR \to \cup_{i=1}^{Q}\{s_i\}$ as follows: If $y \in \aleph(s_i)$, then $\upsilon(y) = s_i$. The function $\upsilon(\cdot)$ is well-defined.

We do not choose each s_i arbitrarily from its strongly communicating class. We choose each s_i from its strongly communicating class in such a way that the following is true: $\forall x \in \aleph(s_i)$, $p_x^{\Pi^*_{\aleph(s_i)}}[\sum_{k=0}^{\infty} I(x_k = s_i) = \infty] = 1$. I.e., an optimal, Markov, deterministic, stationary policy exists for the restricted SMDP on $\aleph(s_i)$ such that s_i is in its positive recurrent class. The proof of Theorem 6.1.1 guarantees our ability to choose each s_i in such a way that this condition is satisfied. The reason we enforce this condition on s_i is to guarantee that a stationary, Markov, deterministic optimal policy on the restricted SMDP can hit s_i.

Lemma 8.2.7 *If $D \cap \aleph(y) = \emptyset$, then $\forall x \in S$ and any two nonempty subsets $U_1, U_2 \subset \aleph(y)$, $R(x, D \cup U_1)$ iff $R(x, D \cup U_2)$.*

Proof:

First we prove the more general transitivity result that if $D, B, C \subset S$, $R(x, D \cup B)$, and $R(B, C)$, then $R(x, D \cup C)$:

Let Π_1 be such that $P_x^{\Pi_1}[\tau_{D\cup B} < \infty] = 1$. Let Π_2 be such that $\forall y \in B$, $P_y^{\Pi_2}[\tau_C < \infty] = 1$. Define Π_3 to be the policy that follows Π_1 until B is reached. Then, the history is forgotten and policy Π_2 is followed. We have

$$P_x^{\Pi_3}[\tau_{D\cup C} < \infty] \geq P_x^{\Pi_1}[\tau_{D\cup B} < \infty] \cdot \inf_{y\in B} P_y^{\Pi_2}[\tau_C < \infty] = 1 \cdot 1.$$

To obtain the lemma, note that $R(U_1, U_2)$ and $R(U_2, U_1)$ are both true by definition of $\aleph(\cdot)$. And the lemma follows.

□

Define $A = \cup_{i=1}^{Q}\{s_i\}$, and let $2^A = \{D | D \subset A\}$. Define $2^A(x) \subset 2^A$ as follows:

$D \in 2^A(x)$ iff $R(x, D)$ and $D \in 2^A$.

N.B.: $2^A(x) \neq 2^{A_x}$ if we define $A_x \doteq \{y \in A | R^p(x, D)\}$ or even $A_x \doteq \{y \in A | R(x, D)\}$. In fact, $2^A(x)$ is not necessarily a power set of anything.

Lemma 8.2.8 *Given Assumption 8.2.2,* $2^A(x) \neq \emptyset \; \forall x \in S$

Proof:

Let $x \in S$ be given. By Assumption 8.2.2, $\exists D \subset SR$ such that $|D| < \infty$ and $R(x, D)$. Let $C = \cup_{x \in D} \upsilon(x)$. By Lemma 8.2.7, we have $R(x, C)$. By Lemma 7.0.6, $\exists F \subset C$ such that $R^{'}(x, F)$. By construction, $F \in 2^A(x)$. Therefore, the Lemma is true.

□

Lemma 8.2.9 *Let Assumption 4.3.2 and Assumption 4.3.1 be true and let* $\Pi \in \Pi^{HR}$ *be an arbitrary policy. If* $J_x^\Pi < \infty$*, then* $P_x^\Pi[\eta_{SR} = \infty] = 1$.

Proof:

Let $J_x^\Pi = F < \infty$. Define

$$CC_F = \{x \in S | \inf_{a \in \alpha(x)} E[e^{\gamma c(x,a) - 2Ft(x,a)}] \leq 1\}.$$

By Assumption 4.3.2 and Assumption 4.3.1, we know that CC_F contains a finite number of elements.

By Corollary 8.2.1, we know that

$$\limsup_{N \to \infty} E_x^\Pi[e^{\sum_{k=0}^N \gamma\{c(x_k,a_k) - 2Ft(x_k,a_k)\}}] = 0. \tag{8.9}$$

Claim:

$$P_x^\Pi[\eta_{CC_F} = \infty] = 1.$$

Proof of claim:

If not, then (8.9) is violated. (We have previously proved similar claims (e.g. claim in Lemma 4.2.3) in detail and the proof of this one is omitted.)

And the claim is proved.

Claim:

$$P_x^\Pi[\eta_{CC_F \cap SR} = \infty] = 1. \tag{8.10}$$

Proof of claim:

By the definition of SR it is clear that for any finite subset $G \subset SR^c$, $P_x^\Pi[\eta_G = \infty] = 0$.

So we have

$$1 = P_x^\Pi[\eta_{CC_F} = \infty] = P_x^\Pi[\eta_{CC_F \cap SR} = \infty] + P_x^\Pi[\eta_{CC_F \cap SR^c} = \infty]$$

$$= P_x^\Pi[\eta_{CC_F \cap SR} = \infty] + 0.$$

And the claim is proved.

The Lemma follows easily from this claim.

□

Corollary 8.2.4 *Let Assumption 4.3.2 and Assumption 4.3.1 be true and let $\Pi \in \Pi^{HR}$ be an arbitrary policy. If $J_x^\Pi < \infty$, then $\exists B \subset SR$, $|B| < \infty$ such that $R^\Pi(x, B)$.*

Proof:

By the second claim in the proof of Lemma 8.2.9, we know that there is a finite set $CC_F \subset S$ such that $P_x^\Pi[\eta_{CC_F \cap SR} = \infty] = 1$. Therefore, $R^\Pi(x, CC_F \cap SR)$ and the corollary holds.

□

Lemma 8.2.10 *Let Assumption 8.2.1(γ) be true.*

Then $\forall x \in S$,

$$\inf_{\Pi\in\Pi^{HR}} J_x^{\Pi} = \inf_{D\in 2^A(x)} \sup_{s\in D} \inf_{\Pi\in\Pi^{HR}} J_s^{\Pi}.$$

Proof:

By Lemma 8.2.3, it is clear that

$$\inf_{\Pi\in\Pi^R(x,\cup_{i=1}^Q\{s_i\})} J_x^{\Pi} = \inf_{\Pi\in\cup_{D\in 2^A(x)}\Pi^R(x,D)} J_x^{\Pi} \leq \inf_{D\in 2^A(x)} \sup_{s\in D} \inf_{\Pi\in\Pi^{HR}} J_s^{\Pi}.$$

since $\Pi^R(x,\cup_{i=1}^Q\{s_i\}) \subset \Pi^{HR}$, we get

$$\inf_{\Pi\in\Pi^{HR}} J_x^{\Pi} \leq \inf_{\Pi\in\Pi^R(x,\cup_{i=1}^Q\{s_i\})} J_x^{\Pi} \leq \inf_{D\in 2^A(x)} \sup_{s\in D} \inf_{\Pi\in\Pi^{HR}} J_s^{\Pi}.$$

Let $\Pi^{'} \in \Pi^{HR}$ such that $J_x^{\Pi^{'}} < \infty$ be given. By Lemma 8.2.9, we know that $P_x^{\Pi^{'}}[\eta_{SR} = \infty] = 1$. Clearly then, $R^{\Pi^{'}}(x, SR)$. And by Lemma 7, $\exists C \subset SR$ such that $R^{'\Pi^{'}}(x, C)$.

Then, by Lemma 8.2.5

$$J_x^{\Pi^{'}} \geq \sup_{y\in C} \inf_{\Pi\in\Pi^{HR}} J_y^{\Pi}, \tag{8.11}$$

Let $D = \cup_{y\in C}\upsilon(y)$. By Lemma 7.0.1, we know that $\forall y \in SR$,

$$\inf_{\Pi\in\Pi^{HR}} J_y^{\Pi} = \inf_{\Pi\in\Pi^{HR}} J_{\upsilon(y)}^{\Pi}.$$

Therefore,

$$\sup_{y\in C} \inf_{\Pi\in\Pi^{HR}} J_y^{\Pi} = \sup_{y\in C} \inf_{\Pi\in\Pi^{HR}} J_{\upsilon(y)}^{\Pi}$$

and by (8.11),

$$J_x^{\Pi^{'}} \geq \sup_{s\in D} \inf_{\Pi\in\Pi^{HR}} J_s^{\Pi},$$

By Lemma 8.2.7, $R(x, C)$ implies $R(x, D)$. Therefore, $D \in 2^A(x)$ and

$$\sup_{s\in D} \inf_{\Pi\in\Pi^{HR}} J_s^{\Pi} \geq \inf_{D\in 2^A(x)} \sup_{s\in D} \inf_{\Pi\in\Pi^{HR}} J_s^{\Pi},$$

and the Lemma is satisfied.

□

Lemma 8.2.10 is useful for finding the infimum of achievable cost for a state $x \notin SR$ in terms of the infimum of achievable cost for the states in SR. If $x \in SR$, the statement of the theorem can be simplified to

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \inf_{\Pi \in \Pi^{HR}} J_{v(x)}^{\Pi}.$$

Therefore for $x \in SR$, it is clear that

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} \leq \inf_{D \in 2^A(x) - \{v(x)\}} \sup_{s \in D} \inf_{\Pi \in \Pi^{HR}} J_s^{\Pi}.$$

Lemma 8.2.11 *Let Assumption 8.2.1(γ) be true and let $x \in SR$.*

If

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} < \inf_{D \in 2^A(x) - \{v(x)\}} \sup_{s \in D} \inf_{\Pi \in \Pi^{HR}} J_s^{\Pi},$$

then

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \lambda^*_{\aleph(x)}.$$

Proof:

First, we note that the assumption of the lemma implies $\exists \epsilon > 0$ such that

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} + \epsilon < \inf_{D \in 2^A(x) - \{v(x)\}} \sup_{s \in D} \inf_{\Pi \in \Pi^{HR}} J_s^{\Pi}. \tag{8.12}$$

We know that for the restricted SMDP (i.e., actions must be in $\beta(x)$), $\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \lambda^*_{\aleph(x)}$. Therefore, for the regular SMDP (any actions in $\alpha(x)$),

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} \leq \lambda^*_{\aleph(x)}. \tag{8.13}$$

All that remains to be shown is that $\forall \Pi \in \Pi^{HR}$,

$$J_x^{\Pi} \geq \lambda^*_{\aleph(x)}. \tag{8.14}$$

Let $\Pi^{'} \in \Pi^{HR}$ be given.

If $P_x^{\Pi^{'}}[a_k \in \beta(x_k)] = 1 \ \forall k < \infty$, then $\Pi^{'}$ is admissible under the restricted SMDP and therefore (8.14) holds.

Suppose that $\exists k < \infty$ such that $P_x^{\Pi^{'}}[a_k \in \beta(x_k)] < 1$, and $\forall i < k$, $P_x^{\Pi^{'}}[a_i \in \beta(x_i)] = 1$.

Denote $X_k \doteq \{y \in S | P_x^{\Pi^{'}}[x_k = y] > 0\}$.

Claim:

$X_k \subset \aleph(x)$.

Proof of claim:

Since policy $\Pi^{'}$ by definition takes only actions that are admissible under the restricted SMDP prior to time t_k, the claim is true.

So there must be $y \in \aleph(x)$ and a $p > 0$ such that

$$P_x^{\Pi^{'}}[x_k = y, a_k \notin \beta(y)] = p.$$

Therefore,

$$p = P_x^{\Pi^{'}}[x_k = y] \cdot P_x^{\Pi^{'}}[a_k \notin \beta(y) | x_k = y],$$

And we get the following two inequalities:

$$P_x^{\Pi^{'}}[x_k = y] > 0,$$

and

$$P_x^{\Pi^{'}}[a_k \notin \beta(y) | x_k = y] \geq p.$$

By Corollary 8.2.2, we get

$$J_x^{\Pi^{'}} \geq \inf_{\Pi^{''} \in \Pi^{HR} | P_y^{\Pi^{''}}[a_0 \notin \beta(y)] \geq p} J_y^{\Pi^{''}}, \tag{8.15}$$

since $\Pi^{'}$ can do no better than the infimum.

Because for any such policy Π'', an action that is not in $\beta(y)$ is selected with probability at least $p > 0$, we know that

$$P_y^{\Pi''}[\tau_{\aleph(x)} < \infty] < 1.$$

Therefore, $R^{\Pi''}(y, \aleph(x))$ and $R'^{\Pi''}(y, \{\upsilon(x)\})$ are not true for any such policy Π''. It is clear that if $D \notin 2^A(y)$, then $R'^{\Pi''}(y, D)$ is not true for any such policy Π''. Let $2^A(y)''$ denote the set of all sets $D \in 2^A$ such that $\exists \Pi''$ such that $P_y^{\Pi''}[a_0 \notin \beta(y)] \geq p$ with $R^{\Pi''}(y, D)$ true.

Therefore we have

$$\{\upsilon(x)\} \notin 2^A(y)'' \tag{8.16}$$

and

$$2^A(y)'' \subset 2^A(y) = 2^A(x). \tag{8.17}$$

Then by an argument analogous to the argument used to obtain a lower bound in Lemma 8.2.10, we get that

$$\inf_{\Pi'' \in \Pi^{HR} | P_y^{\Pi''}[a_0 \notin \beta(y)] \geq p} J_y^{\Pi''} = \inf_{D \in 2^A(y)''} \sup_{s \in D} \inf_{\Pi \in \Pi^{HR}} J_s^{\Pi}$$

$$\geq \inf_{D \in 2^A(x) - \{\upsilon(x)\}} \sup_{s \in D} \inf_{\Pi \in \Pi^{HR}} J_s^{\Pi}$$

$$> \inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} + \epsilon, \tag{8.18}$$

where the last inequality follows from (8.12); and the second to last inequality follows from (8.16) and (8.17).

From (8.15) and (8.18), we see that if a policy does not select actions within the restricted SMDP at all times w.p.1, then it cannot come to within ϵ of the optimal cost. Therefore, the optimal cost is no better than the best that can be achieved within the restricted SMDP, or

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} \geq \lambda^*_{\aleph(x)}.$$

The Lemma follows from the above inequality combined with (8.13)

□

Note: Lemma 8.2.11 makes intuitive sense: If you do better by not making the system leave the strongly communicating class it is in than you do by making the system leave the strongly communicating class it is in, then the best you can do is the optimal cost for the restricted SMDP, which is the SMDP that ensures that the system does not leave the strongly communicating class it is in.

Corollary 8.2.5 *If* $\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} < \lambda^*_{\aleph(x)}$, *then*

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \inf_{D \in 2^A(x) - \{v(x)\}} \sup_{s \in D} \inf_{\Pi \in \Pi^{HR}} J_s^{\Pi}.$$

But there is another possibility. It is possible to have

$$\lambda^*_{\aleph(x)} = \inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \inf_{D \in 2^A(x) - \{v(x)\}} \sup_{s \in D} \inf_{\Pi \in \Pi^{HR}} J_s^{\Pi}.$$

This occurs when the best cost you can attain by making the system leave $\aleph(x)$ is the same as the best cost you can attain by making the system stay in $\aleph(x)$.

It follows that the only way in which you will benefit (in terms of long term average cost) by having the system leave the strongly communicating class it currently occupies is if the condition of Corollary 8.2.5 holds. In the other case, i.e. when $\lambda^*_{\aleph(x)} = \inf_{\Pi \in \Pi^{HR}} J_x^{\Pi}$, the best cost within the restricted SMDP is the best possible cost.

Lemma 8.2.12 *Let* $D_1, D_2 \in 2^A$. *If* $R(x, D_1)$, $s \in D_1$, $s \notin D_2$, *and* $R(s, D_2)$, *then* $\exists D_3 \subset D_1 \cup D_2 - \{s\}$ *such that* $R(x, D_3)$.

Proof:

If $R(x, D_1 - \{s\})$, then set $D_3 = D_1 - \{s\}$ and we are done.

Suppose that not $R(x, D_1 - \{s\})$. Therefore, $\exists \Pi \in \Pi^{HR}$ such that $R^{\Pi}(x, D_1)$ and $P_x^{\Pi}[\tau_s = \tau_{D_1}] > 0$. Because $R(s, D_2)$, we know that $\exists \Pi' \in \Pi^{HR}$ such that $R^{\Pi'}(s, D_2)$. Let Π'' be the policy that follows Π until τ_{D_1}. Then, if $\tau_{D_1} = \tau_s$, the history is erased and policy Π' is followed. It is clear that $R^{\Pi''}(x, D_1 \cup D_2 - \{s\})$, which implies that $R(x, D_1 \cup D_2 - \{s\})$.

□

Define $B \subset \cup_{i=1}^{Q}\{s_i\}$ as follows:

$s_i \in B$ if

$$\lambda^*_{\aleph(s_i)} = \inf_{\Pi \in \Pi^{HR}} J_{s_i}^{\Pi}.$$

Define

$$2^B(x) = \{D' \in 2^A(x) | D' \subset B\}.$$

So $2^B(x)$ is the largest subset of $2^A(x)$ for which the optimal policy can be achieved within the restricted SMDP for each s in each D.

Given $0 < F < \infty$, define $C^F \subset \cup_{i=1}^{Q}\{s_i\}$ as follows:

$s_i \in C^F$ if $\exists y \in \aleph(s_i)$ such that

$$\inf_{a \in \alpha(y)} E[e^{\gamma\{c(y,a) - Ft(y,a)\}}] \leq 1.$$

Clearly, if Assumption 4.3.1 and Assumption 4.3.2 are true then $|C^F| < \infty$.

Define

$$2^{C^F}(x) = \{D' \in 2^A(x) | D' \subset C^F\}.$$

Lemma 8.2.13 *If $\exists \Pi \in \Pi^{HR}$ and $x \in S$ such that $J_x^{\Pi} < \infty$, then $\exists F > 0$ such that $2^{C^F}(x) \neq \emptyset$.*

Proof:

Let $\epsilon > 0$ be given.

Define $ZQ = \{z| \inf_{a \in \alpha(z)} E[e^{\gamma\{c(z,a)-(J_x^\Pi+\epsilon)t(z,a)\}}] \leq 1$.

Suppose $2^{C^{J_x^\Pi+\epsilon}}(x) = \emptyset$. Then if y is a self-reachable state that is probabilistically reachable from x, we have

$$\inf_{a \in \alpha(y)} E[e^{\gamma\{c(y,a)-(J_x^\Pi+\epsilon)t(y,a)\}}] > 1.$$

Therfore, $ZQ \subset SR^c$.

By the definition of J_x^Π and by Corollary 8.2.1, we know that

$$\lim_{T\to\infty} E_x^\Pi[e^{\gamma \sum_{k=0}^{T}\{c(x_k,a_k)-(J_x^\Pi+\epsilon)t(x_k,a_k)\}}] = 0.$$

Therefore, we must have that

$$P_x^\Pi[\eta_{ZQ} = \infty] = 1.$$

By Lemma 2.2.2, this means that $ZQ \cap SR \neq \emptyset$.

This is a contradiction, so the Lemma is proved.

□

Lemma 8.2.14 *If Assumption 4.3.1 and Assumption 4.3.2 are true, $x \in S$, and $\exists \Pi \in \Pi^{HR}$ such that $J_x^\Pi < \infty$, then the infimum over $2^A(x)$ in*

$$\inf_{\Pi \in \Pi^{HR}} J_x^\Pi = \inf_{D \in 2^A(x)} \sup_{s \in D} \inf_{\Pi \in \Pi^{HR}} J_s^\Pi \tag{8.19}$$

is achieved. Furthermore, $\exists\ D \in 2^B(x)$ that achieves the minimum.

Proof:

Let $\Pi^{'}$ be a policy such that $J_x^{\Pi^{'}} < \infty$. From the proof of Lemma 8.2.13, we know not only that for any $F > J_x^{\Pi^{'}}$, $2^{C^F}(x) \neq 0$, but that $R^\Pi(x, \cup_{s_i \in C^F} \aleph(s_i))$ for any policy Π such that $J_x^\Pi \leq J_x^{\Pi^{'}}$.

Since Assumption 4.3.1 and Assumption 4.3.2 are true, we know that

$\{x | \inf_{a\in\alpha(x)} E[e^{\gamma\{c(x,a)-Ft(x,a)\}}] \leq 1\}$ is finite for any F. Therefore, there are only finitely many $D \in 2^A(x)$ such that $D \in 2^{C^F}(x)$, so the minimum in (8.19) must be achieved.

Claim:

$2^B(x) \cap 2^{C^F}(x) \neq \emptyset$. Furthermore, the minimum in (8.19) is achieved by a $D \in 2^B(x) \cap 2^{C^F}(x)$

Proof of claim:

Suppose that $2^B(x) \cap 2^{C^F}(x) = \emptyset$. and that $D^{'} \in 2^{C^F}(x)$ achieves the minimum in (8.19). Let $D_0 \doteq D^{'}$.

Now, we will generate an infinite sequence $\{D_i\}|_{i=0}^{\infty}$, with $D_i \neq D_j$ for $i \neq j$, $D_i \in 2^{C^F}(x)$, and D_i achieves the minimum in (8.19) $\forall i$. Since $2^{C^F}(x)$ is finite, that is a contradiction.

Let D_i be given. Define D_{i+1} as follows:

Because $D_i \notin 2^B(x)$, $\exists y \in D_i$ such that $\lambda^*_{\aleph(y)} > \inf_{\Pi\in\Pi^{HR}} J^{\Pi}_y$. Therefore (since we've shown the infimum over $D \in 2^A(y)$ is achieved in $2^{C^F}(y)$, $\exists \bar{D} \in 2^A(y)$ such that $R(y, \bar{D})$ and

$$\inf_{\Pi\in\Pi^{HR}} J^{\Pi}_y = \sup_{s\in\bar{D}} \inf_{\Pi\in\Pi^{HR}} J^{\Pi}_s. \tag{8.20}$$

By Lemma 8.2.12, $\exists D_{i+1} \subset D_i \cup \bar{D} - \{y\}$ such that $R(x, D_{i+1})$. Furthermore, by (8.20) and the inductive hypothesis, we know that D_{i+1} achieves the minimum in (8.19).

Now we need to show that $D_i \neq D_j$ for $i \neq j$. By construction, we know that $R(D_i, D_{i+1})$ and therefore $R(D_i, D_j)$ if $i < j$. We also know that $D_i \neq D_{i+1}$. Therefore, if $D_i = D_j$ for $i \neq j$, then $|i - j| > 1$. Therefore, if $i < j$, we have $R(D_i, D_{i+1})$ and $R(D_{i+1}, D_j)$. Since $D_i = D_j$, that means $R(D_{i+1}, D_i)$.

Sub-Claim:

If $A, B \in 2^A$, $A \neq \emptyset$, $B \neq \emptyset$, $R(A, B)$, and $R(B, A)$, then $A = B$.

Proof of sub-claim:

Suppose otherwise. Then $\exists s_i \in A - B$. By the assumption of the sub-claim, we have $R(s_i, B)$ and $R(B, s_i)$ By the definition of $\aleph$, this means that $B \subset \aleph(s_i)$. But then $B = \emptyset$, which is a contradiction.

And the sub-claim is proved.

By the sub-claim, we then have that $D_i = D_{i+1}$, which is a contradiction.

Therefore, we have constructed an infinite sequence of sets $\{D_i\}_{i=0}^{\infty}$, with each $D_i \in 2^{C^F}(x)$, with $D_i \neq D_j$ for $i \neq j$. But this is impossible because $2^{C^F}(x)$ is a finite set! Therefore, there must be an $i < \infty$ such that $D_i \imath 2^B(x)_x$. By construction, this D_i achieves the minimum in (8.19).

And the claim is proved.

The D from the statement of the claim is the $D \in 2^B(x)$ stated in the conclusion of the lemma, and the lemma is proved.

□

Lemma 8.2.15 *Let Assumption 8.2.1(γ) be true.*

Then $\forall x \in S$,

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \inf_{D \in 2^A(x)} \sup_{s \in D} \lambda^*_{\aleph(s)}.$$

Proof:

We know that $\inf_{\Pi \in \Pi^{HR}} J_s^{\Pi} \leq \lambda^*_{\aleph(s)}$ because $\beta(x) \subset \alpha(x)$. Therefore, we have by Lemma 8.2.10 that

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \inf_{D \in 2^A(x)} \sup_{s \in D} \inf_{\Pi \in \Pi^{HR}} J_s^{\Pi} \leq \inf_{D \in 2^A(x)} \sup_{s \in D} \lambda^*_{\aleph(s)}.$$

Lemma 8.2.14 tells us that $\exists D^{'} \in 2^B(x)$ such that

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \sup_{s \in D^{'}} \inf_{\Pi \in \Pi^{HR}} J_s^{\Pi}.$$

Since $D^{'} \in 2^B(x)$, we see that

$$\inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \sup_{s \in D} \lambda^*_{\aleph(s)}.$$

Therefore $\inf_{\Pi \in \Pi^{HR}} J_s^{\Pi} \geq \lambda^*_{\aleph(s)}$ and the Lemma is proved.

□

This lemma makes intuitive sense. The best you can do is the best you can do by moving the system to a reachable set and then following the restricted SMDP. This is true because eventually a control must stay within an strongly communicating class; otherwise the trajectory would go to infinity and the cost would be infinite.

For $D \in 2^A$, define $\mu(D) = \sup_{s \in D} \lambda^*_{\aleph(s)}$.

If Assumption 4.3.2 holds, then for any $K < \infty$, there are only a finite number of s_i such that $\lambda^*_{\aleph(s)} < K$. Therefore, if Assumption 4.3.2 holds and $|D| = \infty$, then $\mu(D) = \infty$.

Theorem 8.2.1 *Suppose that Assumptions 3.4.1, 4.3.1, 4.3.2, and 4.3.3 hold and that Assumption 8.2.1(γ) holds for all $\gamma < \bar{\gamma}$.*

Then, for any $\gamma < \bar{\gamma}$ and any $x \in S$, there exists a stationary, Markov, deterministic policy Π_x^* such that

$$\inf_{D \in 2^A(x)} \mu(D) = \lim_{T \to \infty} \frac{1}{\gamma T} \ln E_x^{\Pi_x^*}[e^{\int_{t=0}^T g(t)dt}] \leq \lim_{T \to \infty} \frac{1}{\gamma T} \ln E_x^{\Pi}[e^{\int_{t=0}^T g(t)dt}]; \forall \Pi \in \Pi^{HR}.$$

Proof:

By Lemma 8.2.15, we know that

$$\inf_{D \in 2^A(x)} \mu(D) \leq \lim_{T \to \infty} \frac{1}{\gamma T} \ln E_x^{\Pi}[e^{\int_{t=0}^T g(t)dt}]; \forall \Pi \in \Pi^{HR}.$$

We will now construct a stationary policy $\Pi^* \in \Pi^{MD}$ that achieves the minimum cost, $\inf_{D \in 2^A(x)} \mu(D)$.

By Lemma 8.2.14, $\exists D^{'} \in 2^{A}(x)$ such that $\mu_{D'} = \inf_{D \in 2^{A}(x)} \mu(D)$. By Lemma 7.0.6, $\exists D^{*} \subset D^{'}$ such that $R^{'}(x, D^{*})$. Because $D^{*} \subset D^{'}$, we know that $\mu_{D^{*}} \leq \mu_{D'} = \inf_{D \in 2^{A}(x)} \mu(D)$. (Clearly this must hold with equality.) Lemma 7.0.7 tells us that $\forall s \in D^{*}$, not $R(s, D^{*} - \{s\})$.

Claim:

$\exists$ a stationary policy $\Pi_{x}^{D^{*}} \in \Pi^{MD}$ such that $E_{x}^{\Pi_{x}^{D^{*}}}[e^{\sum_{k=0}^{\tau_{D}} \gamma c(x_k, a_k)}] < \infty$.

Proof

By Assumption 8.2.1(γ), $\exists$ a policy $\Pi^{'}$ such that $E_{x}^{\Pi^{'}}[e^{\sum_{k=0}^{\tau_{D}} \gamma c(x_k, a_k)}] < \infty$.

Let us create a 'modified' SMDP by altering the state space: eliminate every state $y \in D$ and replace them with state d. Denote the transition probabilities for the modified SMDP by giving them a ' superscript. $\forall x, a$, define $P^{'}[x_{k+1} = d | x_k = x, a_k = a] = \sum_{y \in D} P[x_{k+1} = y | x_k = x, a_k = a]$. The transition probabilities out of state d are irrelevant, and set $P^{'} = P$ otherwise.

Essentially, all we have done to 'modify' the SMDP is aggregate all of the states in D into one state, called d.

By Lemma 6.1.1, $\exists$ a Markov, stationary policy Π_{γ}^{d} such that $E_{x}^{\Pi_{\gamma}^{d}}[e^{\sum_{k=0}^{\tau_{d}} \gamma c(x_k, a_k)}] < \infty$ in the modified SMDP.

This same policy in the original SMDP satisfies the claim, so set $\Pi_{x}^{D^{*}} = \Pi_{\gamma}^{d}$ and the claim is proved.

For each $s \in \cup_{i=1}^{Q} \{s_i\}$, by Theorem 6.1.1, there is a stationary optimal policy Π_{s}^{*} for the restricted SMDP over $\aleph(s)$, i.e. the SMDP that has admissible actions only in $\beta(x)$. Define the stationary policy Π_{*}^{*} over $x \in SR$ as follows:

$$\Pi_{*}^{*}(x) = \Pi_{\nu(x)}^{**}(x),$$

where $\Pi_{\nu(x)}^{**}$ denotes the stationary, Markov, deterministic optimal policy for the restricted SMDP on $\aleph(\nu(x))$. (This policy is derived in Theorem 6.1.1.)

Clearly, for $x \in SR$, $J_{x}^{\Pi_{*}^{*}} = \lambda_{\aleph(x)}^{*}$. (For $x \notin SR$, $J_{x}^{\Pi_{*}^{*}}$ is not defined because Π_{*}^{*} is not defined.)

Define $\aleph^{'}(D) \doteq \cup_{s\in D}\aleph(s)$, and define

$$\Pi_x^* = \begin{cases} \Pi_*^*;\text{ if } x \in \aleph^{'}(D^*) \\ \Pi_x^{D^*};\text{ otherwise} \end{cases}$$

Denote $\lambda^* = \mu(D^*)$.

Claim:

$$E_x^{\Pi_x^*}[e^{\gamma\sum_{k=0}^{\tau_D-1}\{c(x_k,a_k)-\lambda^* t(x_k,a_k)\}}] < \infty.$$

Proof of claim:

We get

$$E_x^{\Pi_x^*}[e^{\gamma\sum_{k=0}^{\tau_D-1}\{c(x_k,a_k)-\lambda^* t(x_k,a_k)\}}]$$

$$= E_x^{\Pi_x^*}[e^{\sum_{k=0}^{\tau_{\aleph'(D)}-1}\gamma\{c(x,a)-\lambda^* t(x,a)\}} \cdot e^{\sum_{k=\tau_{\aleph'(D)}}^{\tau_D-1}\gamma\{c(x,a)-\lambda^* t(x,a)\}}]$$

$$= E_x^{\Pi_*^*}[e^{\sum_{k=0}^{\tau_{\aleph'(D)}-1}\gamma\{c(x,a)-\lambda^* t(x,a)\}}.$$

$$E_{x_{\tau_{\aleph'(D)}}}^{\Pi_x^{D^*}}[e^{\sum_{k=\tau_{\aleph'(D)}}^{\tau_D-1}\gamma\{c(x,a)-\lambda^* t(x,a)\}}|\{x_0,x_1,...,x_{\tau_{\aleph'(D)}}\}]]$$

$$\leq E_x^{\Pi_*^*}[e^{\sum_{k=0}^{\tau_{\aleph'(D)}-1}\gamma\{c(x,a)-\lambda^*_{\aleph(x_{\tau_{\aleph'(D)}}}t(x,a)\}}.$$

$$E_{x_{\tau_{\aleph'(D)}}}^{\Pi_x^{D^*}}[e^{\sum_{k=\tau_{\aleph'(D)}}^{\tau_D-1}\gamma\{c(x,a)-\lambda^*_{\aleph(x_{\tau_{\aleph'(D)}})}t(x,a)\}}|\{x_0,x_1,...,x_{\tau_{\aleph'(D)}}\}]]$$

$$\leq E_x^{\Pi_*^*}[e^{\sum_{k=0}^{\tau_{\aleph'(D)}-1}\gamma\{c(x,a)-\lambda^*_{\aleph(x_{\tau_{\aleph'(D)}}}t(x,a)\}}.$$

$$E_{x_{\tau_{\aleph'(D)}}}^{\Pi_*^*}[e^{\sum_{k=\tau_{\aleph'(D)}}^{\tau_D-1}\gamma\{c(x,a)-\lambda^*_{\aleph(x_{\tau_{\aleph'(D)}})}t(x,a)\}}|\{x_0,x_1,...,x_{\tau_{\aleph'(D)}}\}]]$$

$$= E_x^{\Pi_*^*}[e^{\sum_{k=0}^{\tau_D-1}\gamma\{c(x,a)-\lambda^*_{\aleph(x_{\tau_{\aleph'(D)}})}t(x,a)\}}]$$

$$\leq E_x^{\Pi^*_*}[e^{\sum_{k=0}^{\tau_D -1} \gamma c(x,a)}] < \infty.$$

And the claim is proved.

Claim:

$\forall \lambda > \lambda^*$,

$$E_x^{\Pi^*_x}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] = 0.$$

Proof of claim:

$$E_x^{\Pi^*_x}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]$$

$$= E_x^{\Pi^*_x}[e^{\gamma \sum_{k=0}^{\tau_D -1}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}} \cdot e^{\gamma \sum_{k=\tau_D}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]$$

$$= \sum_{m=1}^{\infty} P_x^{\Pi^*_x}[\tau_D = m] E_x^{\Pi^*_x}[e^{\gamma \sum_{k=0}^{m-1}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}} \cdot e^{\gamma \sum_{k=m}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}} | \tau_D = m]$$

(since $m < \infty$, we can replace it with 0)

$$= \sum_{m=1}^{\infty} P_x^{\Pi^*_x}[\tau_D = m] E_x^{\Pi^*_x}[e^{\gamma \sum_{k=0}^{m-1}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}.$$

$$\sum_{s \in D} P_x^{\Pi^*_x}[x_m = s | m = \tau_D] E_s^{\Pi^*_x}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] | \tau_D = m]$$

(define $s^{'} = \arg\max_{s \in D} \lambda^*_{\aleph(s)}$)

$$\leq \sum_{m=1}^{\infty} P_x^{\Pi^*_x}[\tau_D = m] E_x^{\Pi^*_x}[e^{\gamma \sum_{k=0}^{m-1}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}.$$

$$E_{s^{'}}^{\Pi^*_x}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] | \tau_D = m]$$

$$= E_{s^{'}}^{\Pi^*_x}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}].$$

$$\sum_{m=1}^{\infty} P_x^{\Pi_x^*}[\tau_D = m] E_x^{\Pi_x^*}[e^{\gamma \sum_{k=0}^{m-1}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}} | \tau_D = m]$$

$$= E_{s'}^{\Pi_x^*}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] \cdot E_x^{\Pi_x^*}[e^{\gamma \sum_{k=0}^{\tau_D -1}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}]$$

$$\leq E_{s'}^{\Pi_x^*}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] \cdot E_x^{\Pi_x^*}[e^{\gamma \sum_{k=0}^{\tau_D -1}\{c(x_k,a_k)-\lambda^* t(x_k,a_k)\}}]$$

(from the definition of Π_*^*)

$$= E_{s'}^{\Pi^{**}_{s'}}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] \cdot E_x^{\Pi_x^*}[e^{\gamma \sum_{k=0}^{\tau_D -1}\{c(x_k,a_k)-\lambda^* t(x_k,a_k)\}}].$$

We know from the previous claim that

$$E_x^{\Pi_x^*}[e^{\gamma \sum_{k=0}^{\tau_D -1}\{c(x_k,a_k)-\lambda^* t(x_k,a_k)\}}] < \infty.$$

Also, we know from Corollary 8.2.1 that

$$E_{s'}^{\Pi^{**}_{s'}}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}}] = 0.$$

And the claim follows.

From the above claim and Lemma Equiv, we know that $J_x^{\Pi_x^*} \leq \lambda^*$, and the theorem is proved.

□

Theorem 8.2.1 is a powerful result. However, although the optimal policy is stationary, Markov, and deterministic; it still might depend on the initial state. This is a phenomenon that is not restricted to a risk sensitive objective function – in fact the optimal policy depends on the initial state for a risk neutral objective function whenever it does for a risk sensitive objective function (with sufficiently small γ) for the same SMDP. Figure 8.1 illustrates a Markov decision process for which the optimal policy depends on the initial state. (It will be explained in the next section.)

It is interesting to note that if we augment the state with the initial state, we can then get a single policy that is optimal from any initial condition. So the entire history dependence of an optimal policy can be summarized in one piece of informaition: the initial state, or equivalently, the optimal cost achievable from that initial state.

For a risk neutral objective function, not only might the optimal policy depend on the initial state, but there might not be an optimal policy at all. In fact, Theorem 8.2.1 does not hold for a risk neutral objective function. I.e., there are some SMDPs in which there is an optimal policy for a risk sensitive objective function but not for a risk neutral objective function.

To illustrate these points, we shall now provide some concrete examples. As a consolation, we note that it will be shown in the next chapter that in the finite state case, these problems go away for both risk neutral and risk sensitive objective functions.

Example 8.2.1 (Examples of problems that occur when $S = \infty$)

Figure 8.1 shows a complex MDP with an infinite state space. SR consists of 2 states, which we call z_9 and z_{10}, labeling each by its cost. (In the following examples, $c(x, a)$ does not depend on a, so we write it as $c(x)$.) There are also 3 separate columns of states. We label the states in the left-hand column the 'a' states: $\{a_1, a_2, a_3, ...\}$; where again each state is labeled according to its transition cost. The states in the middle column are labeled $\{b_1, b_2, b_3, ...\}$, again according to their transition costs. The states in the right hand colum are labeled $\{c_1, c_2, c_3, ...\}$. The cost of transitioning out of state c_n is $f(n)$, where $f(\cdot)$ is defined according to the following recursion:

$$f(1) = 1; f(n+1) = 2^{f(n)}.$$

The transition probabilities are as labeled, although $P[x_{k+1} = a_{n+1} | x_k = a_n] = 2^{-n}$ is not labeled in order to prevent crowding of the diagram. As can be

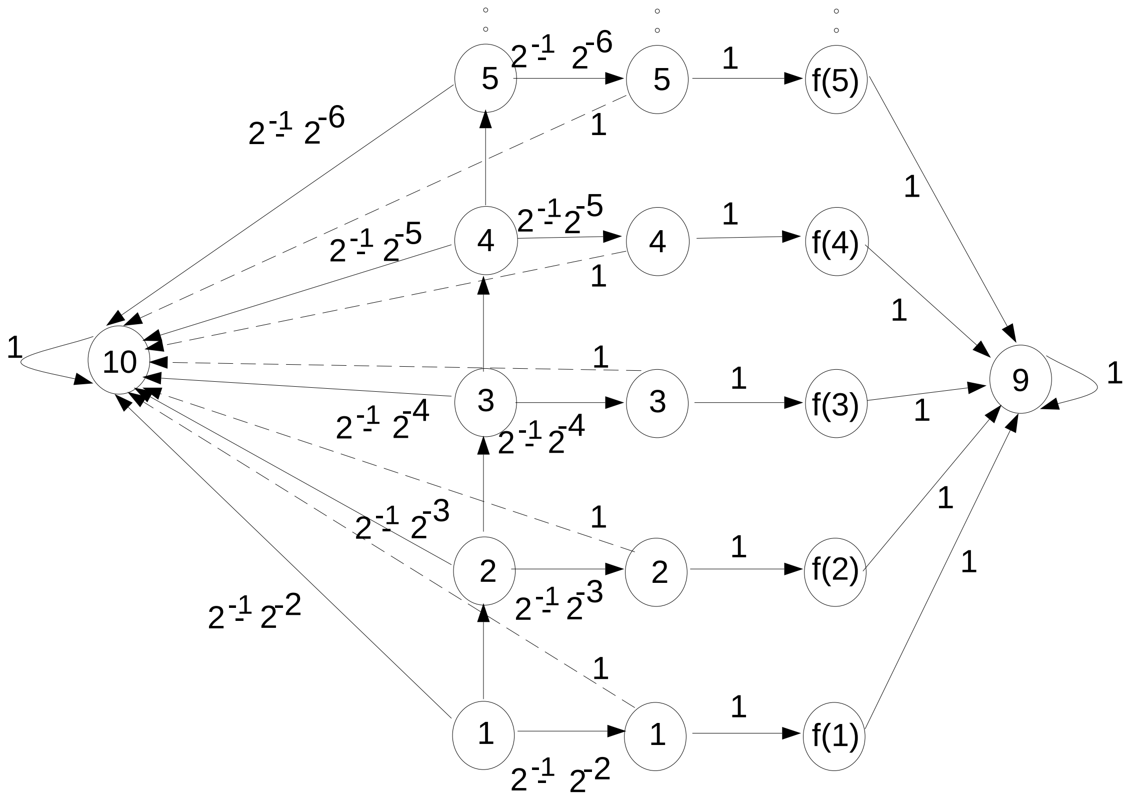

Figure 8.1: An MDP for which no single policy is optimal from every initial state.

seen from the diagram, transitions are deterministic except from the 'a' states, where they are random. $|\alpha(x)| = 1$ if $x \in \{z_9, z_{10}\} \cup \{a_1, a_2, a_3, ...\} \cup \{c_1, c_2, c_3, ...\}$. However, $|\alpha(x)| = 2$ if $x \in \{b_1, b_2, b_3, ...\}$. Let us call the two admissible actions in state b_n action A_9 and action A_{10}. $c(b_n, A_9) = c(b_n, A_{10}) = n$. The transition probabilities for action A_{10} are shown with the broken lines, while the transition probabilities for action A_9 are shown with solid lines, just like the uncontrollable transition probabilities from the other states.

Note: The MDP defined in figure 8.1 is well formulated (i.e., does not incur infinite costs for all policies from any initial state,) since the cost to reach the b states from any a state is finite and you can drive the system to the 10 state in 1 step with cost 1 from any b state. Since the 10 state is self-reachable, there is a policy to reach SR with finite cost.

8.3 Properties of optimal policies for the SMDP in Example 8.2.1

In this section, we explore an example that illustrates the limitations of Theorem 8.2.1. We also explore the behavior of the risk neutral objective function in the same example draw generalizations from it.

8.3.1 The risk sensitive case

Define policy Π_9 to be the policy that always chooses action A_9; and define policy Π_{10} to be the policy that always chooses action A_{10}. Clearly, both are stationary, Markov, deterministic policies.

The techniques of the proof of Theorem 8.2.1 can be used to show that $\forall x \in \{a_1, a_2, a_3, ...\}$,

$$J_x^{\Pi_{10}} = 10 \leq J_x^{\Pi}; \, \forall \Pi \in \Pi^{HR}.$$

Similarly, the same techniques can be used to show that $\forall x \in \{b_1, b_2, b_3, ...\}$,

$$J_x^{\Pi_9} = 9 \leq J_x^{\Pi}; \, \forall \Pi \in \Pi^{HR}.$$

However, it can be seen that if $\Pi^{'}$ is a policy that chooses action A_{10} in state b_n, then $J_{b_n}^{\Pi^{'}} = 10 > 9$. Therefore, policy Π_{10} is not optimal if the initial state is in $\{b_1, b_2, b_3, ...\}$.

Now, let us solve for the value of $J_{a_n}^{\Pi_9}$. (For simplicity, we assume $n = 1$, but the result is the same for any n.)

$$J_{a_1}^{\Pi_9} = \lim_{N \to \infty} \frac{1}{\gamma N} \ln E_{a_1}^{\Pi_9}[e^{\gamma \sum_{k=0}^{N} c(x_k, a_k)}].$$

We can see that

$$E_{a_1}^{\Pi_9}[e^{\gamma \sum_{k=0}^{N} c(x_k,a_k)}] \leq E_{a_1}^{\Pi_9}[e^{\gamma \sum_{k=0}^{N} c(x_k,a_k) \cdot I(x_k \in \{c_1,c_2,\ldots,c_n\})}]$$

$$= \sum_{n=1}^{N-1} P[x_{n+1} = c_n] \cdot e^{\gamma f(n)}$$

$$\geq \sum_{n=1}^{N-1} (2^{-n-1})^{n+1} e^{\gamma f(n)},$$

since the transition probabilities are bounded below by 2^{-n-1} and it takes $n+1$ transitions to reach state c_n.

A sum of positive terms is bounded above by its last term. Therefore,

$$J_{a_1}^{\Pi_9} \leq \lim_{N\to\infty} \frac{1}{\gamma N} \ln(2^{-n-1})^{n+1} e^{\gamma f(n)}$$

$$\leq \lim_{N\to\infty} \frac{1}{\gamma N} - (n+1)^2 + \gamma f(n) = \infty,$$

by definition of $f(n)$.

In fact, the above development also shows something stronger: that if Π'' chooses action A_9 for an infinite number of states $\in \{b_1, b_2, \ldots\}$, then $J_{a_n}^{\Pi''} = \infty$.

Therefore it can be seen that any policy that is optimal for all $x \in \{b_1, b_2, b_3, \ldots\}$ (and there is only one: Π_9) leads to infinite cost if used starting at any state in $\{a_1, a_2, a_3, \ldots\}$. Similarly, any policy that is optimal starting at any state in $\{a_1, a_2, a_3, \ldots\}$ must choose action A_9 for only a finite number of states, and therefore is not optimal for all initial states in $\{b_1, b_2, b_3, \ldots\}$.

So we see that Figure 8.1 shows an example of an MDP that meets the conditions of Theorem 8.2.1 for which there is no policy that is optimal starting from every state. This is true no matter what the value of $\gamma > 0$.

8.3.2 The risk neutral case

Let's see what happens when $\gamma = 0$, i.e., in the risk neutral case. Clearly, if the initial state is not in $\{a_1, a_2, a_3, ...\}$, then policy Π_9 is optimal and the optimal cost is 9. In order to examine the behavior of the cost if the initial state is in $\{a_1, a_2, a_3, ...\}$, we assume without loss of generality that the initial state is a_1. (The same thing happens for any other a_n, as will become evident.)

Define policy Π^n as the policy that chooses action A_9 if $x \in \{b_1, b_2, ..., b_{n-1}, b_n\}$; and action A_{10} otherwise. We therefore have $\lim_{n\to\infty} \Pi^n = \Pi_9$ and $\Pi^0 = \Pi_{10}$.

Because $P[x_{k+1} = a_{n+1} | x_k = a_n] = 2^{-n}$, we see that for any $\Pi \in \Pi^{HR}$, $P^n \doteq P^{\Pi}_{a_1}[\tau_{a_n} < \infty] = \Pi_{i=1}^{n-1} 2^{-i} > 0$, while $\lim_{n\to\infty} P^n = 0$.

It can be shown that

$$9 < \mathcal{J}^{\Pi^n}_{a_1} \leq 9 + P^n.$$

Therefore, we see that $\lim_{n\to\infty} \mathcal{J}^{\Pi^n}_{a_1} = 9$.

However, the same development that showed us that $J^{\Pi_9}_{a_1} = \infty$ also shows us that $\mathcal{J}^{\Pi_9}_{a_1} = \infty$. And as in the risk sensitive case, if $\Pi^{''}$ chooses action A_9 for an infinite number of states $\in \{b_1, b_2, ...\}$, then $\mathcal{J}^{\Pi^{''}}_{a_n} = \infty$. Also, we know that any policy $\Pi^{'''}$ that chooses action A_9 for only a finite number of states $\in \{b_1, b_2, ...\}$ gives cost $\mathcal{J}^{\Pi^{'''}}_{a_n} > 9$.

So we have that $\inf_{\Pi \in \Pi^{HR}} \mathcal{J}^{\Pi}_{a_n} = 9$, and that $\mathcal{J}^{\Pi}_{a_n} > 9 \; \forall \Pi \in \Pi^{HR}$. There is no optimal policy if the initial state is in $\{a_1, a_2, a_3, ...\}$!

Example 8.3.1

The following is a simpler example in which there is no optimal risk neutral policy.

In each state of the Markov chain in figure 8.2, there are two admissible actions. Action AA_0 gives the transition probabilities shown with the solid lines; action AA_1 gives the the transition probabilities shown with the dashed lines.

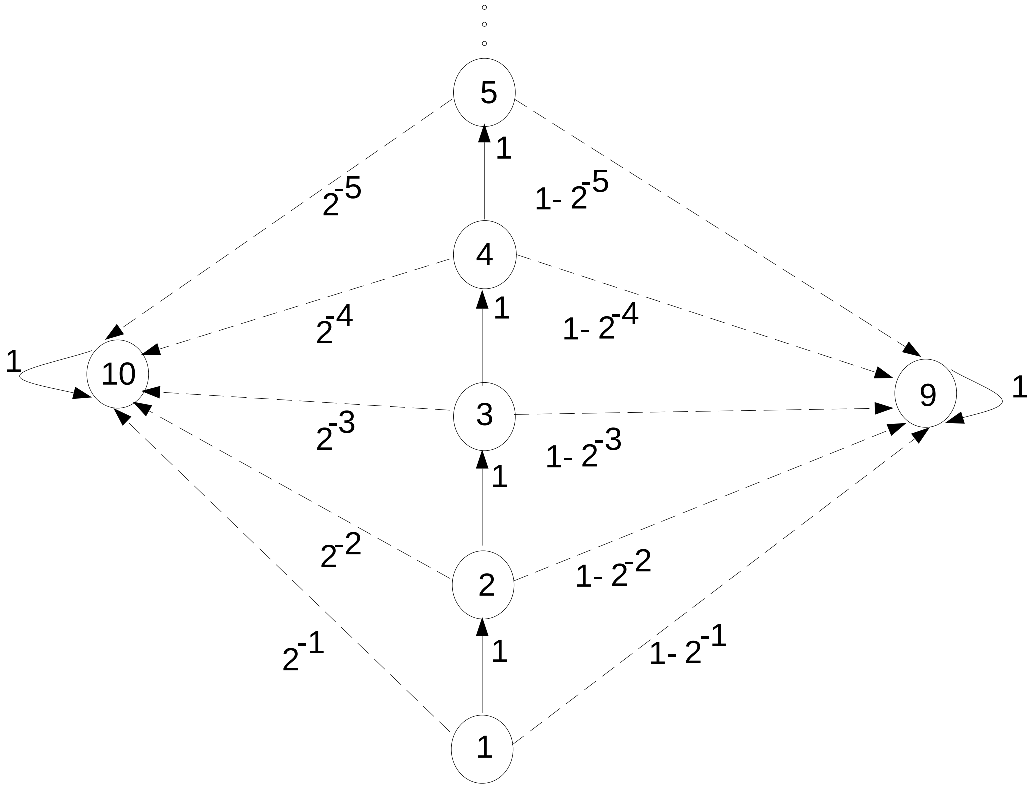

Figure 8.2: An MDP illustrating why Theorem 8.2.1 does not work for a risk neutral objective function.

It can be seen through a development similar to the one in Section 8.3.1 that a risk neutral policy can achieve cost as close to 9 as desired, but no policy achieves a 9 cost.

Every policy achieves a risk sensitive cost of 10 except those policies that have a nonzero probability of always choosing action AA_0. Such policies give infinite cost for both risk neutral and risk sensitive objective functions.

Therefore we see that it is possible for an optimal risk sensitive policy to exist when an optimal risk neutral policy does not. This is because of the maximization (over strongly communicating classes) property of the risk sensitive objective function.

Chapter 9

Optimal Policies and Optimality Equations for the Finite State Space Case

We now know that there is an optimal policy starting from any initial state. In this chapter we show that for the finite state space case, there is a single policy that is optimal from all initial states. Furthermore, there is a pair of optimality equations that hold. These optimality equations are important because they form the basis for computation of an optimal policy. (However, we do not address computation in this thesis.) Furthermore, now that we have examined optimality principles, we come full circle to extend the discussion of Chapter 5 and address the general behavior of a Markov chain.

9.1 Optimality equations and a policy optimal from any state

In Theorem 8.2.1, we learned that the optimal cost starting from state x is given by

$$J_x^* \doteq \inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \inf_{D \in 2^A(x)} \mu(D).$$

We use this property in Theorem 9.1.1 to find a policy that is optimal from any initial state for $|S| < \infty$. First we introduce some notation.

There is a natural ordering on the states $s \in \{s_1, s_2, ..., s_Q\}$, defined by $s_i < s_j$ if $i < j$. Similarly, we define a lexicographic ordering on the sets of states in 2^A: If $D_1, D_2 \in 2^A$, define $D_1 < D_2$ if $\exists s_i \in D_1 - D_2$ such that $\forall j < i$, $s_j \notin D_2 - D_1$.

Define $\tilde{Z}(x) \doteq \{D \in 2^A(x) | R^{'}(x, D)\} \cap \{D \in 2^B(x) | \mu(D) \leq \mu(D^{'}) \forall D^{'} \in 2^A(x)\}$.

We know from the proofs of Lemmas 8.2.14 and 8.2.15 that $\tilde{Z}(x) \neq \emptyset$.

Define $D^*(x) \in \tilde{Z}(x)$ to be the D in $\tilde{Z}(x)$ that comes first in the lexicographic ordering. It is important to point out that this could have been chosen as the D^* in the proof of Theorem 8.2.1, a fact that will be used in the proof of Theorem 9.1.1, which extends Theorem 8.2.1.

Theorem 9.1.1 *Suppose that assumptions 3.4.1, 4.3.1, 4.3.2, and 4.3.3 hold and that Assumption 8.2.1(γ) holds for all $\gamma < \bar{\gamma}$. Suppose also that $|S| < \infty$.*

Then, for any $\gamma < \bar{\gamma}$, there exists a stationary, Markov, deterministic policy Π^* such that $\forall x \in S$

$$\inf_{D \in 2^A(x)} \mu(D) = \lim_{T \to \infty} \frac{1}{\gamma T} \ln E_x^{\Pi^*}[e^{\int_{t=0}^T g(t) dt}] \leq \lim_{T \to \infty} \frac{1}{\gamma T} \ln E_x^{\Pi}[e^{\int_{t=0}^T g(t) dt}];\ \forall \Pi \in \Pi^{HR}.$$

Proof:

Theorem 8.2.1 tells us that for each x there is an optimal policy. Recall the definition of that optimal policy from the proof of Theorem 8.2.1:

$$\Pi_x^* = \begin{cases} \Pi_*^*; \text{ if } x \in \aleph^{'}(D^*(x)) \\ \Pi_x^{D^*(x)}; \text{ otherwise} \end{cases}$$

Recall that Π_*^* is the optimal policy under the restricted SMDP, and $\Pi_x^{D^*(x)}$ is the shortest path optimal policy to reach $D^*(x)$.

In order for a policy to be optimal starting from state x, Lemma 8.2.15 shows us that all it has to do is achieve cost $\mu(D^*(x))$. One way to do that (the way used by policy Π_x^*) is to drive the system to $\aleph^{'}(D^*(x))$ with finite cost and then follow policy Π_*^*. We will now define a policy that does at least that well starting from any intitial state. This policy opportunistically drives the system towards $D^*(x)$ when it is in state x. If it moves to another state y in which $D^*(y)$ comes before $D^*(x)$ in the lexicographical ordering, it will then drive the system to $D^*(y)$. Because there are only finitely many states, the system reaches $\aleph^{'}(B)$ with finite expected cost starting from any initial state.

Define

$$\Pi^* = \begin{cases} \Pi_*^*; \text{ if } x \in \aleph^{'}(B) \\ \Pi_{sp}^{D^*(x)}; \text{ otherwise} \end{cases} \tag{9.1}$$

Here we define Π_{sp}^D to be the stationary, Markov, deterministic policy that drives the system to D with minimum expected cost. This policy is shown to exist in the proof of Theorem 8.2.1.

Because $|S| < \infty$, we know that $Q < \infty$. Therefore $|B| < \infty$. As in previous proofs, it can be shown that Assumption 8.2.1(γ) implies that $\forall x$ $\forall y$ such that $R^{\Pi_{sp}^{D^*(x)}}(x, y)$,

$$E_x^{\Pi_{sp}^{D^*(x)}}[e^{\gamma \sum_{k=0}^{\tau_y} c(x_k, a_k)} I[\tau_y < \infty]] < \infty.$$

Define

$$MM = \max_{x \in S} \max_{y | R^{\Pi_{sp}^{D^*(x)}}(x,y)} E_x^{\Pi_{sp}^{D^*(x)}}[e^{\gamma \sum_{k=0}^{\tau_y} c(x_k, a_k)} I[\tau_y < \infty]].$$

Because $|S| < \infty$, we know that $MM < \infty$.

Claim:

$\forall x \in S$,

$$E_x^{\Pi^*}[e^{\gamma \sum_{k=0}^{\tau_{\aleph'(B)}} c(x_k, a_k)}] \leq MM^{(2^Q)}.$$

Proof of claim:

By (9.1), if the system is in state x, policy $\Pi_{sp}^{D^*(x)}$ is followed until either $\aleph'(B)$ is reached or a y is reached such that $D^*(y) \neq D^*(x)$. Clearly if such a y is reached, then $D^*(y)$ comes before $D^*(x)$ in the lexicographic ordering. Upon reaching such a y, policy $\Pi_{sp}^{D^*(x)}$ is followed until either $\aleph'(B)$ is reached or a z is reached such that $D^*(z) \neq D^*(y)$. And so on. Eventually $\aleph'(B)$ must be reached since 2^A has finitely many members. In fact, $|2^A| \leq 2^{|A|} = 2^{|Q|}$. Therefore, At most $2^{|Q|}$ policy changes take place before $\aleph'(B)$ is reached. Since the expected total cost accrued between each policy change is bounded above by MM, the expected value of the total cost accrued before B is reached is bounded above by $MM^{(2^Q)}$.

And the claim is proved.

It is evident by (9.1) that $P_x^{\Pi^*}[\lambda^*_{\aleph(\tau_B)} \leq \mu(D^*(x))] = \inf_{D \in 2^A(x)} \mu(D)$.

Therefore, policy Π^* takes the system with finite expected cost to a strongly communicating class that has optimal long term average cost less than or equal to the best possible cost that can be achieved starting at the initial state, so

$$J_x^{\Pi^*} = \inf_{D \in 2^A(x)} \mu(D),$$

and the theorem is proved.

□

Recall that for $x \in SR$, $\aleph(x) = \{y \in SR | R(x,y) \text{ and } R(y,x)\}$. For $x \notin SR$, define $\aleph(x) = \emptyset$.

Define $\mho(x) = \{y \in S | R(y,x)\}$, and define (for $A \subset S$, $x \in S$)

$$\Theta(x, A) = \begin{cases} \sup_{z \in A - \mho(x)} \lambda(z); \text{ if } A - \mho(x) \neq \emptyset \\ \lambda(x); \text{ otherwise} \end{cases} \tag{9.2}$$

Note that $\aleph(x) \subset \mho(x)$.

In order to understand the optimality equations of Theorem 9.1.2, (9.3) and (9.4), we must understand what $\Theta(x, A)$ is. In words, $\Theta(x, A)$ is the worst $\lambda(\cdot)$ one can get in the subset of A from which x is not reachable; or $\lambda(x)$ if that subset is the empty set. If A is replaced with $r(x, a)$, the set of states reachable in one transition from x under action a, then $\Theta(x, A)$ is the worst $\lambda(\cdot)$ for those one-transition reachable states (under a) from which x is not reachable. In other words, (9.3) says that once the system leaves an strongly communicating class, since it can't get back w.p.1, the maximum cost rule applies. (9.4) is just the standard dynamic programming equation adapted for a nonconstant λ. It is important to understand this in order to interpret the following results.

The following results (Theorem 9.1.2 and Lemma 9.1.1) are similar to the results in Puterman's Section 9.1 ([35]). In particular, there are dual optimality equations in both the risk sensitive and risk neutral cases. However, there is no equivalent to Theorem 9.1.2 in [35], and there is no equivalent to Proposition 9.1.1 ([35], P. 445) in the risk sensitive case due to the differing natures of the risk neutral and risk sensitive cases.

Also note: If $|S| = \infty$ then complications arise in Theorem 9.1.2. This is because the bias term ($W(x)$) cannot be reconciled consistently with the fact that λ depends on x. In the risk neutral case, this problem is avoided because instead of maximizing costs between different possible strongly communicating classes, costs are averaged in the risk neutral case.

Theorem 9.1.2 *Suppose that assumptions 3.4.1, 4.3.1, 4.3.2, and 4.3.3 hold and that Assumption 8.2.1(γ) holds for all $\gamma < \bar{\gamma}$. Suppose also that $|S| < \infty$.*

Then, for any $\gamma < \bar{\gamma}$, there exists two functions $\lambda : S \to \Re^+$ and $W : S \to \Re$, and four constants, $-\infty < K_1 < K_2 < \infty$ and $0 < K_3 < K_4 < \infty$ such that

$$K_1 < W(x) < K_2;\ \forall x \in S$$

and

$$K_3 < \lambda(x) < K_4;\ \forall x \in S.$$

Furthermore, the following two equations hold:

$$\lambda(x) = \inf_{a \in \alpha(x)} \Theta(x, r(x,a)) \tag{9.3}$$

$$e^{W(x)} = \inf_{a \in G_x} E[e^{\gamma\{c(x,a) - \lambda(x)t(x,a)\}}] \int e^{W(y)} P(dy|x,a), \tag{9.4}$$

where $G_x \subset \alpha(x)$ is defined as $G_x = \arg\min_{a \in \alpha(x)} \Theta(x, r(x,a))$. Moreover, the infimums in both equations are achieved.

Proof:

Let $\Pi^*, D^*(x)$ be as defined in the proof of Theorem 9.1.1.

Define $\lambda(x) \doteq \inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} = \mu_{D^*}(x) = J_x^{\Pi^*}$.

Claim:

$$\lambda(x) = \inf_{a \in \alpha(x)} \Theta(x, r(x,a)).$$

Proof of claim:

Given $x \in S$, let $a^* \doteq \Pi^*(x)$. It can be seen that for any stationary, Markov policy Π,

$$J_x^{\Pi} = \sup_{y \in r(x, \Pi(x))} J_y^{\Pi}.$$

Therefore we know that $\forall y \in r(x, a^*)$, $J_y^{\Pi^*} \leq J_x^{\Pi^*}$. By definition of $\Theta(\cdot,\cdot)$, this gives us

$$\lambda(x) = J_x^{\Pi^*} \geq \Theta(x, r(x, a^*)) \geq \inf_{a \in \alpha(x)} \Theta(x, r(x, a)).$$

Now we must show the reverse inequality to be true.

Suppose $\exists a \in \alpha(x)$ such that $\Theta(x, r(x, a)) < \lambda(x)$. Therefore $\exists a^{'} \in \alpha(x)$ such that $\sup_{z \in A - \mho(x)} \lambda(z) < \lambda(x)$. Define $AA \doteq A \cap \mho(x)$. $\forall y \in AA$, we know that $R(y, x)$. Therefore by Lemma 7.0.10 we have that $R(x, A - \mho(x))$. By Assumption 8.2.1(γ), $\exists$ a policy $\Pi^{x \to A - \mho(x)}$ such that $E_x^{\Pi^{x \to A - \mho(x)}}[e^{\gamma \sum_{k=0}^{\tau_{A - \mho(x)} - 1} c(x_k, a_k)}] \doteq C < \infty$.

Since the system can reach $A - \mho(x)$ from x with finite expected cost, we have that

$$\lambda(x) = \inf_{\Pi \in \Pi^{HR}} J_x^{\Pi} \leq \sup_{y \in A - \mho(x)} \inf_{\Pi \in \Pi^{HR}} J_y^{\Pi}$$

$$= \sup_{y \in A - \mho(x)} \lambda(y) < \lambda(x),$$

which is a contradiction. And the claim is proved.

Define $\Pi^{HR, G_x} = \{\Pi \in \Pi^{HR} |$ w.p.1, $a_k \in G_{x_k}\ \forall k\}$.

For $x \in S$, define

$$W(x) \doteq \ln\{\inf_{\Pi \in \Pi^{HR, G_x}} E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_B} \{c(x_k, a_k) - \lambda(x_k) t(x_k, a_k)\}}]\}.$$

Claim:

$W(\cdot)$ is bounded above and below over S.

Proof of claim:

We know that

$$e^{W(x)} = \inf_{\Pi \in \Pi^{HR}} E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_B} \{c(x_k, a_k) - \lambda(x_k) t(x_k, a_k)\}}] \leq \tag{9.5}$$

$$\inf_{\Pi \in \Pi^{HR}} E_x^{\Pi}[e^{\gamma \sum_{k=0}^{\tau_B} c(x_k, a_k)}] < \infty,$$

where the last inequality follows from the fact that policy Π^* reaches B with finite cost starting from any state $x \in S$.

Therefore, $W(\cdot)$ is bounded above.

We know from (9.5) that there is a dynamic program for $W(\cdot)$:

$$e^{W(x)} = \inf_{a \in G_x} E[e^{\gamma\{c(x,a)-\lambda(x)t(x,a)\}}] \sum_{y \in r(x,a)} P[y|x,a]e^{W(y)}.$$

Suppose that $\exists x \in S$ such that $W(x) = -\infty$.

We see from the dynamic program that $\exists a^*(x) \in G_x$ such that $W(y) = 0$ $\forall y \in r(x, a^*(x))$.

Define policy Π^0 to be the policy that chooses action $a^*(x)$ $\forall x$ such that $W(x) = 0$. It can be seen that Π^0 induces a recurrent class C^0 such that $C^0 \subset \aleph(s_i)$ for some $s_i \in A$.

If $s_i \notin B$, then $\lambda(x) < \lambda^*_{\aleph(x)}$ $\forall x \in C^0$. Furthermore $\tau_B = \infty$ w.p.1 since $B \cap C^0 = \emptyset$, so we can write

$$0 = e^{W(x)} = E_x^{\Pi^0}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda(x_k)t(x_k,a_k)\}}];\ x \in C^0,$$

by substituting $\tau_B = \infty$ into (9.5).

Since the optimal policy on C^0 can do no better than an average cost of $\lambda^*_{\aleph(x)}$ and Π^0 can do no better than the optimal policy, we get by Corollary 8.2.1 that

$$E_x^{\Pi^0}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda(C^0)t(x_k,a_k)\}}] = \infty;\ x \in C^0,$$

where we have substituted $\lambda(C^0) \doteq \lambda(x)$ since each $x \in C^0$ has an identical value of $\lambda(x)$ due to the fact that $C^0 \in \aleph(s_i)$ for some $s_i \in A$. But this contradicts $W(x) = 0$!

Therefore we must have that $s_i \in B$. If $s_i \in C^0$, then $P_x^{\Pi^0}[\tau_{s_i} < \infty] = 1$. But we have seen in previous proofs that for any $\lambda < \infty$,

$$E_x^{\Pi^0}[e^{\gamma \sum_{k=0}^{S}\{c(x_k,a_k)-\lambda t(x_k,a_k)\}} > 0,$$

where S is any stopping time for which $P[S < \infty] = 1$.

Therefore, we must have that $s_i \notin B$. This tells us that

$$E_x^{\Pi^0}[e^{\gamma \sum_{k=0}^{\infty}\{c(x_k,a_k)-\lambda(C^0)t(x_k,a_k)\}}] = 0.$$

Pick a state $y \in C^0$. The above eqation is true iff $C_\gamma^{y\to y}(\lambda(C^0)) < 1$. But since $C^0 < \infty$, we know that $C_\gamma^{y\to y}(\lambda(C^0)) = 1$, a contradiction. And the claim is proved.

The infimum in (9.3) is achieved because the action space is compact.

Claim:

The infimum in (9.4) is achieved.

Proof of claim:

G_x is compact because the transition probabilities are a continuous function of a.

And the claim is proved.

The theorem follows from the above claims.

□

Note that the value of $\lambda(\cdot)$ within a strongly communicating class is constant. That is, $\lambda(x) = \lambda(y)$ if $y \in \aleph(x)$. Recall the definition of $\Theta(x, A)$ from equation 9.2. If A is set equal to $r(x,a)$ for some action $a \in \alpha(x)$ as in equation 9.3, and if $A - \mho(x) = \emptyset$, then it is clear that $r(x,a) \subset \aleph(x)$. In equation 9.3, the reason that $\Theta(x, A)$ was set to $\lambda(x)$ in the case when $r(x,a) \subset \aleph(x)$ is because of the fact that $\lambda(\cdot)$ is constant within a strongly communicating class. The first optimality equation, equation 9.3, does not explicitly ensure that $\lambda(\cdot)$ is constant within a strongly communicating class. In order to cause the optimality equations to enforce that condition, equation 9.2 could be changed to the following:

$$\Theta(x,A) = \begin{cases} \sup_{z\in A-\mho(x)} \lambda(z); \text{ if } A - \mho(x) \neq \emptyset \\ \sup_{z\in A} \lambda(z); \text{ otherwise} \end{cases}$$

This definition is also correct (since the two reduce to each other because $\lambda(\cdot)$ is constant within a strongly communicating class), and can be substituted into equation 9.3, with the result that Theorem 9.1.2 will still hold true. The proof that Theorem 9.1.2 is true under this alternative definition of $\Theta(\cdot, \cdot)$ is omitted for the sake of brevity. In an actual dynamic programming situation using the optimality equations for either value or policy iteration, it might be easiest to do the following:

1. Solve for all of the strongly communicating classes.

2. Solve for the optimal cost within each strongly communicating class under the restricted SMDP.

3. Use equation 9.3 on a strongly communicating class merely to identify whether a 'better' set of strongly communicating classes can be reached from it.

Of course, step 3 would have to be applied repeatedly until the algorithm converged.

Lemma 9.1.1 shows that the optimality equations in Theorem 9.1.2 are worthy of their name. Note that it applies also to the countable state space case, $|S| = \infty$.

Also note: this lemma is the risk sensitive equivalent of Theorem 9.1.2 ([35], P. 446). It is a more difficult result because the optimality equations are more complex in the risk sensitive case.

Lemma 9.1.1 (Verification Lemma) *Suppose there exist two functions $\lambda : S \to \Re^+$ and $W : S \to \Re$, with $W(\cdot)$ bounded above and below and $\lambda(\cdot)$ bounded below away from zero and bounded above, such that (9.3) and (9.4) hold. Suppose furthermore that the infimums in both equations are achieved.*

Then the stationary policy $\Pi_*^{**} \in \Pi^{MD}$ that minimizes both (9.3) and (9.4) achieves the optimal cost starting from any initial state, and that cost is given by

$\lambda(\cdot)$. *Specifically,*

$$\lambda(x) = J_x^{\Pi_*^{**}} = \inf_{\Pi \in \Pi^{HR}} J_x^{\Pi}.$$

Proof:

If the initial state, $x \in \aleph^{'}(B)$, then $\lambda(x) = \lambda^*_{\aleph(x)}$ and the lemma reduces to the optimal policy shown in Theorem 6.1.1.

If $x \notin B$, then Π_*^{**} takes the system to B with finite cost and follws the optimal policy from Theorem 6.1.1 from there.

Now all that remains to be shown is that Π_*^{**} takes the system to an element of B that is as good as any other policy.

Since under policy Π_*^{**}, $\lambda(x_{k+1}) \leq \lambda(x_k)$ w.p.1, we know that $\lambda(x_{\tau_B}) \leq \lambda(x)$ w.p.1..

Assume $\exists \Pi^\dagger \in \Pi^{HR}$ such that under $\Pi^\dagger$, $\lambda(x_{\tau_B}) < \lambda(x)$ w.p.1..

Then there must be a $y \in S$ such that $\lambda(y) > \Theta(y, r(y, \Pi^\dagger(y)))$, but this contradicts the definition of $\lambda(\cdot)$. Therefore, under any policy $\lambda(x_{\tau_B}) \geq \lambda(x)$ w.p.1., and the lemma is proved.

□

9.2 Behavior for a fixed Markov, deterministic, stationary policy (i.e., a reducible Markov chain)

In Chapter 5, we saw that for a small risk parameter ($\gamma \downarrow 0$), the risk sensitive cost approaches the risk neutral cost of a stationary, Markov, deterministic policy within one of the policy's positive recurrent classes. However, if the semi-Markov process induced by the policy is not irreducible and the initial state is not in a positive recurrent class, then the relationship between risk neutral cost and risk sensitive cost for a small γ becomes more complex.

Denote a realization of the embedded Markov chain of the SMDP as $\tilde{o}$. $\tilde{o} = \{x_0, x_1, x_2, ...\}$. We say that $\tilde{o} \in \aleph_o^a(z)$ if $\exists N < \infty$ such that $x_j \in \aleph(z)\ \forall j > N$.

We say that $\tilde{o} \in \aleph_o^e(z)$ if $\eta_{\aleph(s_i)} = \infty$. Therefore $\aleph_o^a(z) \subset \aleph_o^e(z)$.

The following lemma shows that with probability one, a realization will eventually be confined to one strongly communicating class under any policy that induces a finite expected long term average risk sensitive cost.

Lemma 9.2.1 *If a stationary* $\Pi^{'} \in \Pi^{MD}$ *is such that* $J_x^{\Pi^{'}}(\gamma) < \infty$ *for some* $x \in S$, *then* $\exists D \in 2^A(x)$ *such that*

$$\sum_{s_i \in D} P_x^{\Pi^{'}}[\tilde{o} \in \aleph_o^a(s_i)] = 1.$$

Furthermore, if Assumption 4.2.3 holds, then $|D| < \infty$.

Proof:

By the proof of Lemma 8.2.13, we know that there is a finite set $ZQ \subset S$ such that $P_x^{\Pi^{'}}[\eta_Z = \infty] = 1$. By Lemma 2.2.2, we can say $ZQ \subset SR$. Therefore $\exists D \in 2^A$ such that $ZQ \subset \aleph^{'}(D)$. Because the initial state is x, we can say that $D \in 2^A(x)$. Because $|ZQ| < \infty$ by the norm-like costs assumption (Assumption 4.2.3), we know that $|D| < \infty$.

We have

$$P_x^{\Pi^{'}}[\eta_{\aleph^{'}(D)} = \infty] = 1.$$

Therefore we know that

$$P_x^{\Pi^{'}}[\tilde{o} \in \cup_{s \in D} \aleph_o^e(s)] = 1$$

because $\aleph^{'}(D)$ being hit infinitely many times implies that $\aleph(s)$ is hit infinitely many times for some $s \in D$.

It can be seen that if $\tilde{o} \in \aleph_o^e(s)$, then $\exists x \in \aleph(s)$ such that $\eta_x = \infty$. Therefore we have that

$$P_x^{\Pi'}[\sup_{x \in \aleph'(D)} I(\eta_x = \infty)] = 1. \tag{9.6}$$

Claim:

$$P_x^{\Pi'}[\tilde{o} \in \aleph_o^a(\nu(x)) | \eta_x = \infty] = 1.$$

Proof of claim:

Recall that Π' is stationary, Markov, and deterministic. Suppose that

$$P_x^{\Pi'}[\tilde{o} \in \aleph_o^a(\nu(x)) | \eta_x = \infty] < 1.$$

Then, $\exists y \notin \aleph(x)$ such that

$$P_x^{\Pi'}[\tau_y < \infty] > 0,$$

which means that

$$P_x^{\Pi'}[\tau_y < \infty | \eta_x = \infty] = 1.$$

Therefore,

$$P_x^{\Pi'}[\eta_y = \infty | \eta_x = \infty] = 1.$$

Therefore,

$$P_y^{\Pi'}[\tau_x < \infty] = 1$$

which implies that $y \in \aleph(x)$, a contradiction!

And the claim is proved.

The lemma follows from (9.6) and the above claim.

□

In this section we are examining cost performance of a fixed stationary, Markov, deterministic policy. For that reason, we need to do more than consider strongly communicating classes $\aleph(z)$ for $z \in SR$. Since we are considering a fixed stationary policy $\Pi \in \Pi^{MD}$, we need to consider the equivalence classes induced by Π. For $z \in SR^{\Pi}$, define $\aleph_{\Pi}(z) = \{x \in S | R^{\Pi}(x, z)$ and $R^{\Pi}(z, x)\}$. Clearly $\aleph_{\Pi}(z) \subset \aleph(z)$.

Similarly, for a realization $\tilde{o}$ of the embedded Markov chain of the SMDP, we say that $\tilde{o} \in \aleph^{a}_{\Pi}(z)$ if $\exists N < \infty$ such that $x_j \in \aleph_{\Pi}(z)\ \forall j > N$.

We also say that $\tilde{o} \in \aleph^{e}_{\Pi}(z)$ if $\eta_{\aleph_{\Pi}(s_i)} = \infty$. Therefore $\aleph^{a}_{\Pi}(z) \subset \aleph^{e}_{\Pi}(z)$.

The following corollary extends Lemma 9.2.1 to take into account the equivalence classes induced by policy $\Pi^{'}$:

Corollary 9.2.1 *If a stationary $\Pi^{'} \in \Pi^{MD}$ is such that $J^{\Pi^{'}}_{x}(\gamma) < \infty$ for some $x \in S$, then $\exists D \in 2^{A}(x)$ such that*

$$\sum_{s_i \in D} P^{\Pi^{'}}_{x}[\tilde{o} \in \aleph^{a}_{\Pi}(s_i)] = 1.$$

Furthermore, if Assupmtion 4.2.3 holds, then $|D| < \infty$.

This leads to a nice lemma that allows us to evaluate the performance of a given stationary, Markov, deterministic policy starting from a given initial state in terms of its performance on the irreducible subclasses $(\aleph_{\Pi}(s); s \in SR^{\Pi})$ that it induces.

Lemma 9.2.2 *If a stationary $\Pi \in \Pi^{MD}$ is such that $J^{\Pi}_{x}(\gamma) < \infty$ for some $x \in S$, then*

$$J^{\Pi}_{x} = \sup_{\aleph_{\Pi}(s) | R^{\Pi}(x, \aleph_{\Pi}(s))} J^{\Pi}_{s}.$$

Proof:

Recall Lemma 8.2.15:

$$\inf_{\Pi \in \Pi^{HR}} J^{\Pi}_{x} = \inf_{D \in 2^{A}(x)} \sup_{s \in D} \lambda^{*}_{\aleph(s)}.$$

If we are dealing with a fixed stationary $\Pi \in \Pi^{MD}$, we can see by application of Corollary 9.2.1 that

$$J_x^\Pi = \sup_{\aleph_\Pi(s) | R^\Pi(x, \aleph_\Pi(s))} \lambda_{\aleph_\Pi(s)},$$

where $\lambda_{\aleph_\Pi(s)} \doteq J_s^\Pi$, and s is any member of the induced equivalence class $\aleph_\Pi(s)$.

□

The statement of Lemma 9.2.2 illustrates the cost maximization nature of the infinite horizon average risk sensitive costs objective function. By contrast, the risk neutral function averages costs, as stated in the follwing lemma:

Lemma 9.2.3 *If a stationary* $\Pi \in \Pi^{MD}$ *is such that* $\mathcal{J}_x^\Pi < \infty$ *for some* $x \in S$, *then*

$$\mathcal{J}_x^\Pi = \sum_{\aleph_\Pi(s) | R^\Pi(x, \aleph_\Pi(s))} P[\tilde{o} \in \aleph_\Pi^a(s)] \cdot \mathcal{J}_s^\Pi.$$

9.3 Behavior for large or small risk sensitive parameter

We saw in Chapter 5 that within a positive recurrent class induced by at stationary, Markov, deterministic policy, the limit of the risk sensitive cost as $\gamma \downarrow 0$ is the risk neutral cost; and the limit of the risk sensitive cost as $\gamma \uparrow \infty$ is the maximum cost. Lemma 9.2.2 shows that the maximum property holds when starting from a transient state, so that Lemma 5.4.3 still holds over the entire state space, not just within a positive recurrent class induced by policy Π.

However, as illustrated by the differences between Lemmas 9.2.2 and 9.2.3, we see that Lemma 5.3.4 does not hold starting from a transient state. In fact, we can generalize Lemma 5.3.4 as follows:

Lemma 9.3.1 *If a stationary* $\Pi \in \Pi^{MD}$ *is such that* $J_x^{\Pi}(\gamma) < \infty$ *for some* $x \in S$, *then*

$$\lim_{\gamma \downarrow 0} J_x^{\Pi} = \sup_{\aleph_{\Pi}(s) | R^{\Pi}(x, \aleph_{\Pi}(s))} \mathcal{J}_s^{\Pi}.$$

Proof:

This follows directly from Lemma 5.3.4 and Lemma 9.2.2.

□

Chapter 10

Some Other Objective Functions

In this chapter, we consider some new objective functions to lend context to the ones we have studied.

10.1 Sample path convergence

In this thesis, we have studied objective functions determined by the expected value of some measure of average performance on the infinite horizon. It is appropriate to ask: "when is that measure of performance achieved with probability 1?" In ([1], PP. 286-288), the *sample path average cost* is defined. Here, we change the notation slightly to conform to our pattern. We also retain the MDP formulation, holding transition times to be constant.

$$J(s)_x^{\Pi} \doteq \lim \sup_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} c(x_t, a_t).$$

A policy Π^* is defined to be *sample path risk neutral average cost optimal* or *almost surely risk neutral average cost* optimal if there is a constant ρ^* such that

$$J(s)_x^{\Pi^*} = \rho^* w.p.1,$$

and

$$J(s)_x^{\Pi} \geq \rho^* w.p.1 \forall \Pi \in \Pi^{HR}.$$

(In [1], they used an arbitrary initial distribution on the state instead of x as the initial state, but in an MDP that is irreducible under all policies (which is assumed in [1]), the two are equivalent.

The following lemma describes sufficient conditions for a policy to achieve the risk neutral cost with probability 1:

Lemma 10.1.1 *Suppose that*

$$E_\theta^{\Pi'}[e^{\sum_{k=0}^{\tau_\theta - 1} \gamma c(\theta_k, a_k)}] < \infty \tag{10.1}$$

for some state $\theta \in S$ under a stationary policy $\Pi' \in \Pi^{MD}$.

Suppose furthermore that $E[c(x, \Pi'(x))]$ is bounded above and below away from zero $\forall x \in \aleph_{\Pi'}(\theta)$.

Then $\forall y \in \aleph_{\Pi'}(\theta)$, we get

$$J(s)_y^{\Pi'} = \mathcal{J}_\theta^{\Pi'} w.p.1.$$

Note: Lemma 10.1.1 and its corollary are true so long as costs are non-negative. However, the proof flows more easily if costs are assumed to be bounded away from zero, so we proceed that way.

Before we prove Lemma 10.1.1, let us introduce a useful theorem ([32], P. 368):

Geometric drift towards C

There exists an extended real-valued function $V : S \to [1, \infty]$, a measurable set C, and constants $\beta > 0, b < \infty$,

$$\Delta V(x) \leq -\beta V(x) + b I_C(x), x \in S. \tag{10.2}$$

where

$$I_C(x) = \begin{cases} 1 \text{ if } x \in C \\ 0 \text{ if } x \notin C \end{cases}$$

Theorem 10.1.1 ([32]) *If (10.2) holds, then for any $r \in (1, (1-\beta)^{-1})$ there exists $\epsilon = \frac{1}{r} - 1 + \beta > 0$ such that*

$$V(x) \leq E_x[\sum_{k=0}^{\tau_C - 1} V(\Phi_k) r^k] \leq \epsilon^{-1} r^{-1} V(x) + \epsilon^{-1} b I_C(x). \tag{10.3}$$

Proof of Lemma 10.1.1:

By (10.1), it can be shown (through a process very similar to the proof of Theorem 5.6.1) that $\exists$ a solution $\{W_\theta(\cdot)\}$, finite for each $x \in \aleph_{\Pi'}(\theta)$ and bounded below, to the following functional equation:

$$e^{W_\theta(x)} = E[e^{\gamma c(x,\Pi(x))}] \int \{e^{W_\theta(y)} \cdot [1 - I(y=\theta)] + I(y=\theta)\} P(dy|x, \Pi(x)); \; \forall x \in \aleph_{\Pi'}(\theta), \tag{10.4}$$

with $W_\theta(\theta) = \ln[C^{\theta \to \theta}(0)] \geq 1$.

Because costs are bounded below away from zero, we know that $\exists$ a constant $c_{\min} > 0$ such that $E[e^{\gamma c(x,\Pi(x))}] \geq e^{\gamma c_{\min}}$ $\forall x \in \aleph_{\Pi'}(\theta)$.

Therefore by (10.4), we get that

$$e^{W_\theta(x)} \geq e^{\gamma c_{\min}} \int \{e^{W_\theta(y)} \cdot [1 - I(y=\theta)] + I(y=\theta)\} P(dy|x, \Pi(x)); \; \forall x \in \aleph_{\Pi'}(\theta).$$

Define $V(x) \doteq e^{W_\theta(x)} \cdot [1 - I(x=\theta)] + I(x=\theta)$.

Substituting, we get

$$V(x_k) \geq e^{\gamma c_{\min}} E[V(x_{k+1})] - C^{\theta \to \theta}(0) I(x_k = \theta).$$

or taking differentials by defining $\Delta V(x_k) \doteq V(x_{k+1}) - V(x_k)$, we get

$$\Delta V(x) \leq \frac{1 - e^{\gamma c_{\min}}}{e^{\gamma c_{\min}}} V(x) + \frac{C^{\theta \to \theta}(0)}{e^{\gamma c_{\min}}} I(x=\theta).$$

This is the geometric drift condition (10.2) with $1 > \beta \doteq \frac{e^{\gamma c_{\min}-1}}{e^{\gamma c_{\min}}} > 0$ since $c_{\min} > 0$. Theorem 10.1.1 tells us that (10.3) holds under $\Pi^{'}$ with $C \doteq \{\theta\}$.

Clearly this implies that

$$E_x^{\Pi^{'}}[V(x_{\tau_\theta - 1})r^{\tau_\theta - 1}] \leq \epsilon^{-1}r^{-1}V(x) + \epsilon^{-1}bI(x = \theta).$$

Also, since $0 < \beta < 1$, we know that $(1-\beta)^{-1} > 1$, so we can select $r > 1$.

By definition of $c_{\min}$, we know that $V(x_{\tau_\theta - 1}) \geq e^{\gamma c_{\min}}$, so we obtain

$$E_x^{\Pi^{'}}[e^{\gamma c_{\min}} r^{\tau_\theta - 1}] \leq \epsilon^{-1}r^{-1}V(x) + \epsilon^{-1}bI(x = \theta).$$

Substituting $x = \theta$ and setting $K \doteq r \cdot e^{-\gamma c_{\min}}\{\epsilon^{-1}r^{-1}V(\theta) + \epsilon^{-1}b\}$, we get

$$E_\theta^{\Pi^{'}}[r^{\tau_\theta}] \leq K.$$

This provides us with a simple geometric bound on τ_θ:

$$P_\theta^{\Pi^{'}}[\tau_\theta \geq n] \leq \frac{K}{r^n} \tag{10.5}$$

$\mathcal{J}_\theta^{\Pi^{'}}$ must exist because (10.1) holds. Let us define, as usual, $\lambda_{\aleph_{\Pi^{'}(\theta)}}^{\Pi^{'}}(0) = \mathcal{J}_\theta^{\Pi^{'}}$.

It is evident from (10.1) that $E_\theta^{\Pi^{'}}[\sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)] < \infty$. Let's now determine a finite bound on $E_\theta^{\Pi^{'}}[\{\sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)\}^2]$.

We know that costs are bounded above, say by $c_{\max}$. Therefore $\sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k) \leq \tau_\theta \cdot c_{\max}$, and we get

$$E_\theta^{\Pi^{'}}[\{\sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)\}^2] \leq E_\theta^{\Pi^{'}}[\{\tau_\theta \cdot c_{\max}\}^2].$$

By (10.5), we get

$$E_\theta^{\Pi^{'}}[\{\sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)\}^2] \leq \sum_{n=0}^{\infty}(nc^{\max})^2 \frac{K}{r^n} < \infty \tag{10.6}$$

because $r > 1$ and the exponential dominates the quadratic.

We now know that $\sum_{k=0}^{\tau_\theta - 1} c(x_k, a_k)$ has finite expected value and finite variance. Therefore the strong law of large numbers applies to $\sum_{k=0}^{\infty} c(x_k, a_k)$. The strong law of large numbers is stated in, e.g., ([27], P. 280). [27] also covers long-term time averages such as the long term average risk neutral cost. (See [27], P. 299.) It can easily be seen from that discussion that the lemma is true.

□

The result Lemma 10.1.1 holds true under less strict conditions. For example, costs do not have to be bounded above. A simple growth condition on the cost function will suffice.

Here are 4 such conditions:

Assumption 10.1.1 (basic growth condition) *$\exists B < \infty$ such that $\Delta c(x) \leq B$ $\forall x \in \aleph_{\Pi'}(\theta)$.*

Assumption 10.1.2 (basic shrinkage condition) *$\exists B < \infty$ such that $\Delta c(x) \geq -B$ $\forall x \in \aleph_{\Pi'}(\theta)$.*

Assumption 10.1.3 (advanced growth condition) *$\exists B < \infty$ and $1 < d < \infty$ such that $\Delta[c(x)]^{\frac{1}{n}} \leq B$ $\forall x \in \aleph_{\Pi'}(\theta)$.*

Assumption 10.1.4 (advanced shrinkage condition) *$\exists B < \infty$ and $1 < d < \infty$ such that $\Delta[c(x)]^{\frac{1}{n}} \geq -B$ $\forall x \in \aleph_{\Pi'}(\theta)$.*

Corollary 10.1.1 *Suppose that*

$$E_{\theta}^{\Pi'}[e^{\sum_{k=0}^{\tau_\theta - 1} \gamma c(\theta_k, a_k)}] < \infty$$

for some state $\theta \in S$ under a stationary policy $\Pi' \in \Pi^{MD}$.

Suppose furthermore that $E[c(x, \Pi'(x))]$ is bounded below away from zero $\forall x \in \aleph_{\Pi'}(\theta)$ and that one of the above four assumptions holds.

Then $\forall y \in \aleph_{\Pi'}(\theta)$, we get

$$J(s)_y^{\Pi'} = \mathcal{J}_{\theta}^{\Pi'} w.p.1.$$

Proof:

The only modification to the proof of Lemma 10.1.1 occurs in equation (10.6). For the basic growth or shrinkage conditions, we end up with the exponential dominating the cubic instead of the exponential dominating the quadratic. For the advanced conditions, it is the exponential dominating the $n + 1^{\text{th}}$ power.

□

10.1.1 Ramifications of sample path convergence – optimality

Lemma 10.1.1 and its corollary can be seen to hold under very general circumstances. The growth conditions listed above are certainly not unreasonable, and our foundational Assumption 6.1.1 leads to the fulfilment of the assumption of finite round trip cost at $\lambda = 0$. So for most of the systems we have analyzed, we can now see that stationary, Markov, deterministic policies yield a fixed sample path average cost and furthermore the optimal risk neutral policy is optimal in the sample path optimality criterion stated at the beginning of this section as well.

This result is really not surprising. The existence of the risk sensitive cost ensured the geometric convergence of the embedded Markov chain, which in turn insured a finite variance in the risk neutral round trip cost. Then, sample path convergence followed by the strong law of large numbers.

Let us compare the result we have just obtained with comparable results from the literature. In [20], a similar method of proof (i.e., geometric convergence to find finite variance and then invoke the strong law of large numbers) is used. However, they assume the geometric convergence directly and add an assumption bounding the transition costs by a measurable function with certain properties.

In [26], a very powerful result is presented. (Recall the discussion of Section 6.2.) Lasserre builds on Borkar's convex analytic approach to prove that if the costs are norm-like and the transition probabilities are continuous in the action selected, then *there is an initial state x_0 and a stationary Markov policy Π_{sp} such that the optimal average risk neutral cost starting from any state under any policy is achieved w.p.1 by every sample path starting from x_0 under policy Π_{sp}*. This is a strong result and we are now able to interpret it. The optimal policy is simply any optimal (risk neutral average expected costs) policy, and the initial state is any state in the 'best' strongly communicating class induced by that policy. Lasserre points out not only the naturalness of the norm-like costs assumption (which applies equally to the assumptions in this thesis), but also the fact that his result is most useful when you can choose your starting state. Of course! If you can, you choose to start in the best strongly communicating class.

It is interesting to point out that because the risk neutral costs converge w.p.1 on every sample path, so do the risk sensitive *sample path* costs. This is because with the expectation operator removed, the exponential and the logarithm cancel out. "Why then is the risk sensitive average cost different than the risk neutral average cost?," one is compelled to ask. The answer is simple: large deviations. It is these deviations that the optimal risk sensitive controller strives to avoid.

For an explanation of this difference between sample path and expected risk sensitive costs, see, e.g., [40] and [33]. Laplace's Law is explained on pages 12-13 of [40]. We put that discussion into our framework as follows:

Suppose we take a large, fixed time T, and determine the probability density function $f(C_T)$ of the finite horizon sample path risk sensitive cost of an irreducible Semi-Markov chain accrued from time 0 to time T. The mode of this probability density determines the expected risk neutral cost, and the mean of this probability density determines the expected risk sensitive cost. Laplace's Law states additionally that the mode of $C_T \cdot f(C_T)$ determines the expected risk sensitive cost.

In [26], it is shown that a linear program can be used to solve for the optimal policy. This is true based only on the norm-like costs assumption and a simple continuity assumption on the transition kernel! Unfortunately in the risk sensitive case no such result yet exists. Solving for the average risk sensitive cost is a difficult task, as demonstrated in [7] in which Borkar and Meyn undertake value and policy iteration. Even under their strong irreducibility assumptions, the task is difficult.

10.2 A cost criterion without γ

A. Makowski has suggested ([29]) that the risk sensitivity parameter in the long term average risk sensitive costs objective function could be done away with, and he has proposed a new objective function:

$$J_x^\Pi(\text{no } \gamma) = \lim_{N\to\infty} \ln E_x^\Pi[e^{\frac{1}{n}\sum_{k=0}^{N-1} c(x_k,a_k)}].$$

Here, we have stated the discrete time version for convenience. All of our analysis carries over to the semi-Markov case with the usual justification.

Upon first examination, the $J_x^\Pi(\text{no } \gamma)$ objective function would appear to be no different from the risk neutral $\mathcal{J}_x^\Pi$ objective function. This is because the 'risk sensitivity parameter' $\frac{1}{N}$ approaches zero as $N \to \infty$. Lemma 5.3.4 would then apply, yielding the risk neutral objective function.

This first blush analysis turns out to be essentially correct in the irreducible case, but matters become more complicated in the not strongly communicating case. Let us take apart this objective function and examine the pieces. In order to do so, we define the cumulative density function of a random variable u to be $F_u(t) = P[u \leq t]$

$$J_x^\Pi(\text{no } \gamma) = \lim_{N\to\infty} \ln \int_{t=0}^{\infty} e^t dF_{\frac{1}{N}\sum_{k=0}^{N-1} c(x_k,a_k)}(t)$$

$$= \ln \int_{t=0}^{\infty} e^t dF_{\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1} c(x_k,a_k)}(t), \qquad (10.7)$$

where the last equality can be justified by convergence of $\mathcal{J}_x^\Pi$.

This is very interesting. Contrast it with Lemmas 9.2.2 and 9.2.3. Lemma 9.2.2 shows that the long run average risk sensitive cost starting from a transient state is given by the maximum cost of any recurrent subset reachable by the initial state. Lemma 9.2.3 shows that the long run average risk neutral cost starting from a transient state is given by the average of the costs of the reachable recurrent subsets weighted by the probability of reaching them. The 'no γ' objective function gives another different result:

Lemma 10.2.1 *If a stationary $\Pi \in \Pi^{MD}$ is such that $J_x^\Pi(\gamma) < \infty$ for some $x \in S$ and some $\gamma > 0$, then*

$$J_x^\Pi(no\ \gamma) = \ln \sum_{\aleph_\Pi(s) | R^\Pi(x, \aleph_\Pi(s))} P[\tilde{o} \in \aleph_\Pi^a(s)] \cdot e^{\mathcal{J}_s^\Pi}.$$

Proof:

See (10.7) and preceding arguments. Note that the fact that $J_x^\Pi(\gamma) < \infty$ for some $\gamma > 0$ implies that $\mathcal{J}_x^\Pi$ exists.

Chapter 11

Closing Remarks and Suggestions

11.1 Summary of major results

• In Chapter 3, we defined the 'deadline problem,' reduced it to an equivalent risk sensitive problem, and formulated a generalized solution technique for finite horizon optimization problems of an SMDP.

• In Chapter 4, we defined a dynamic program for an average cost risk sensitive SMDP. We then proved 2 verification theorems that define an optimal policy.

• In Chapter 5, we solved for the cost of a Markov chain within one of its equivalence classes.

• In Chapter 6, we found the optimal policy for a strongly communicating SMDP.

• In Chapter 8, we found an optimal policy starting from each initial state for an SMDP.

• In Chapter 9, we showed that the optimality equations hold in the finite state case. We also solved for the cost of a Markov chain with finite state space.

• In Chapter 10, we solved for the behavior of some other objective functions and related their behavior to the behavior of the risk sensitive average cost objective function

11.2 On the infinite and its reduction to the finite

In this thesis, we have addressed the problem of optimizing the risk sensitive average cost objective function when the state space is countable. We could also have addressed the more general problem of when the state space is locally compact, and we conjecture that the same results would hold with slight modifications, if any.

When the state space is infinite, the technique to solve the problem invariably becomes reducing it to the finite case. When the time horizon is infinite, the time aspect of the problem can be reduced by making one of the following assumptions:

1. using discounted costs.
2. using average costs.
3. considering a case in which total costs are bounded, e.g., if there is an absorbing state.

In this thesis, we used average costs. The *round trip cost* $C_\gamma^{\theta\to\theta}(\lambda)$ was used to reduce the problem of analyzing realizations with infinite durations to analyzing the finite problem of realizations that start and end at the same state. This is a standard technique used in also in the risk neutral case. (See, e.g., [35] or [4].)

In order to reduce an infinite state space to manageability, one may make one or more of the following assumptions:

1. There is a finite 'core' set of states that is returned to rapidly from any state.
2. The costs are *norm-like.*
3. The costs are bounded.
4. The entire state space is irreducible under all policies.
5. There is a policy that achieves a finite cost.

And in the risk sensitive case,

1rs. The cost to get from a state to a another set of states is finite under some policy.

Assumption (1) above is called the *simultaneous Doeblin* condition and is used, e.g., in [18] along with (3). Assumption (2) can be used with or without (3); see, e.g., [2] and [7]. (Note: In this thesis, we use (2) without (3), although we could just as easily use (3) if desired.) Assumption (4) above is the usual assumption. Assumption (5) is valuable in conjunction with assumption (2) as a starting point or 'bar' under which the optimal policy must fall. (The optimal policy can do no worse than this other policy, which allows us to focus on a finite set of strongly communicating classes.)

The importance of an assumption such as (1rs) is due to the problem of possibly infinite cost to get from one state to another as pointed out in [9]. In [9], the problem was circumvented by assuming that the risk sensitivity parameter γ was 'sufficiently small'. Here we avoid the problem with our assumptions 6.1.1 and 8.2.1.

An appropriate and interesting issue to bring up in this section is that, in the case of norm-like costs, the risk sensitive objective function forces a very disciplined behavior on the underlying Markov chain. From the dynamic programming equation (4.2), we can see that if the average cost λ is finite, then the probability of transitioning to a 'worse' (i.e., $W(\cdot)$ is the same or higher) state is bounded above by $\frac{1}{E[e^{\gamma\{c(x,a)-\lambda t(x,a)\}}]}$. This bound becomes very small for states with high transition costs by Assumptions 4.2.1 and 4.2.3. Therefore we see that under the norm-like costs assumption, there must be a way to drive the system towards 'better' (i.e., lower value of $W(\cdot)$) states with increasingly high probability. One kind of system that achieves this would be a queueing system in which admission control can be exercised. The system could be driven down, for example, by blocking all arrivals once the system is in a 'bad enough' state.

11.3 Future Research

This thesis addresses the properties of the risk sensitive average cost objective function over the infinite horizon and related problems. We have examined a large variety of issues that come up in semi-Markov decision problems in general, and in particular when the standard irreducibility assumption is removed. Although we do not concern ourselves with computational methods such as value and policy iteration, a few remarks will be helpful to the researcher who wishes to pursue this avenue of exploration.

Policy and value iteration are central to the computation of an optimal policy. Other methods include recursive computation, which I have used to solve some simple problems, linear programming (see e.g., [35], [4], [5], [19], [1], and references therein), which is applicable to the solution of the risk neutral objective function, even in the partially observed case (see, e.g., [42] and [28]). Policy and value iteration under the irreducibility assumption have been examined by [7], [4], [35], and others.

In the not strongly communicating case, we suggest that it would be inadvisable to begin value or policy iteration without first understanding the strongly communicating class structure of the embedded Markov chain. This means that each strongly communicating class must be identified and then $2^A(x)$ must be determined for each state x. After that, a framework exists to which one can apply the existing value and policy iteration results. The most relevant result to consult at that point would be Theorem 9.1.2, which shows the optimality equations in the not strongly communicating case. Two good starting points in the literature would be [35], which covers policy and value iteration in the risk neutral average costs case, and [7], which covers value and policy iteration in the risk sensitive average costs case under a strong irreducibility assumption combined with other assumptions.

11.4 Speculation on how to determine strongly communicating classes

In [35], Puterman describes algorithms to classify Markov chains as communicating, weakly communicating, or general. In Section 8.1, we discussed how Puterman classifies MDPs. Similarly, he calls a Markov chain communicating if the trivial MDP it forms is communicating and weakly communicating if the trivial MDP it forms is weakly communicating.

In order to classify Markov chains, Puterman makes use of the Fox-Landi Chain Decomposition algorithm ([35], P. 590) in conjunction with a Model Classification algorithm ([35], P. 351). The Fox-Landi algorithm is used to classify states of a Markov chain (not an MDP) in one of two categories: 'recurrent' or 'transient,' with the obvious definitions. However, by inspection it can be seen that the same algorithm can classify the different strongly communicating classes within the recurrent states. Similarly, transient states can be traced forward to see which subset of the recurrent strongly communicating classes they feed into by an obvious extension of the algorithm. Therefore, we can modify the Fox-Landi algorithm to give a complete picture of a Markov chain.

The Model Classification algorithm described by Puterman uses the Fox-Landi algorithm to classify an MDP. However, we are interested in whether the MDP is strongly communicating, and failing that, what its strongly communicating class structure is. These questions appear to be difficult to answer using an algorithm similar to Puterman's. Therefore, we propose another:

The Fox-Landi algorithm with the aforementioned modifications can be used to find $\aleph^{\Pi}(x)$ for all $x \in S$ and stationary $\Pi \in \Pi^{MD}$. We also want to know whether $\aleph^{\Pi}(x)$ is positive or null recurrent under Π. If $\aleph^{\Pi}(x)$ is finite, it must be positive recurrent. Otherwise, some technique needs to be used to classify it. Let us assume that we can classify it in this way.

We now define an algorithm that we **conjecture** will determine those stongly communicating classes for which a policy exists (on the restricted SMDP) that makes them positive recurrent. (Strongly communicating classes that are null recurrent under any policy are not of interest.):

1. Choose an arbitrary Π_1 in the set of stationary, Markov, deterministic policies.

2. For each $x \in S$, set $\aleph_1(x) = \aleph^{\Pi_1}(x)$ if $\aleph^{\Pi_1}(x)$ is a positive recurrent class, and set $\aleph_1(x) = \emptyset$ otherwise.

3. set $n = 2$

4. Choose a Π_n that hasn't been selected before. If they've all been selected, stop.

5. Combine strongly communicating classes $\aleph_{n-1}(\cdot)$ and $\aleph^{\Pi_n}$ to form the strongly communicating classes $\aleph_n(\cdot)$ as follows:

You can combine two strongly communicating classes if they have a nonempty intersection to form a larger strongly communicating class. You can continue doing this recursively until no strongly communicating classes intersect.

6. increment n and go to step 4. If, however, the whole state space is one big positive recurrent strongly communicating class, stop.

Notice that the above algorithm loops until all stationary, Markov, deterministic policies have been gone through or until the whole state space is one big positive recurrent strongly communicating class. If the latter happens, then the SMDP is strongly communicating. If the former, then we have characterized SR, and the transient states can be traced forward to determine which subsets of SR they can reach w.p.1 by looping through all stationary, Markov, deterministic policies using the modified Fox-Landi algorithm.

One obvious flaw with the above algorithm is that the set of stationary, Markov, deterministic policies might not be finite. One way to remedy that is to solve for all (possibly null recurrent) strongly communicating classes first by using the Fox-Landi algorithm, which cares only whether a transition probability is positive. If this yields a manageable set of strongly communicating classes, a second pass can be used to find those that are positive recurrent. Of course, if the state space is finite, things are much simpler and the above algorithm will converge.

11.5 Computational complexity

Computational complexity is another reason to use the long term average costs criterion over the discounted costs criterion for a risk sensitive objective function. In the risk neutral case, the discounted objective function can be solved efficiently and has nice properties, such as that the optimal policy is stationary, the rate of convergence can be calculated based on the discount factor, etc. In the risk sensitive case, as we saw in the introduction to Chapter 4, that [10] demonstrated that the optimal policy for the discounted risk sensitive objective function is not stationary. Furthermore, its computation is very complex, although for large times the optimal policy converges to the optimal risk neutral average costs policy since the risk sensitivity factor approaches zero, as pointed out in [43]. Furthermore, in [14] a chapter is devoted to risk sensitive queueing, in which the discounted criterion is used. It is shown that this leads to a requirement for a controller with infinite memory! In the average costs case, this does not occur. *Uniformization* is a technique developed by Serfozo in [38]. It is used to reduce a continuous time Markov process to a discrete time process, and works for both average and discounted risk neutral costs on the infinite horizon. However, the difficulty encountered in [14] illustrates that uniformization does not simplify the problem in the risk sensitive discounted costs case. However, in the average costs case, the dynamic program (4.2) reduces the problem to an equivalent discrete time problem.

Appendix A

Index of Notation

$\alpha(x)$, Page 12: admissible actions in state x

d_k, Page 13: deterministic decision rule mapping state to action

q_{d_k}, Page 13: randomized decision rule

Π, Page 13: Policy: a set of decision rules, one defined for each (continuous or discrete) time

$\Pi^L | L \in \{\text{HR,HD,MR,MD}\}$, Page 13: the set of all policies of a certain type

$t(x, a)$, Page 13: transition time from state x under action a

$c(x, a)$, Page 13: transition cost from state x under action a

$P_x^\Pi[\cdot]$, Page 14: probability of an event under policy Π starting at state x

$E_x^\Pi[\cdot]$, Page 14: expected value of a random variable under policy Π starting at state x

J, $\mathcal{J}$, etc., Page 19: notation for objective functions

τ_x, Page 19: first hitting time (greater than 0) of x

σ_x, Page 19: first hitting time of x

η_A^N; $\eta_A \doteq \eta_A^\infty$, Page 21: the number of times the state is in set A out of the first N transitions

$\Pi(R)$, Page 26: decision rule at (discrete or continuous) time R

$\lambda_C^\Pi(\gamma)$, Page 54: *Perron-Frobenius eigenvalue*: cost rate of policy Π in positive invariant suclass C for risk parameter γ

$\bar{\gamma}_C^\Pi$, Page 54: the smallest value of γ that results in $\lambda_C^\Pi(\gamma) = \infty$

$C^{\theta \to \theta}(\lambda)$, Page 54: *round trip cost* (I.e., cost to return) for state θ when discounted by λ

$\lambda_C^\Pi(0)$, Page 72: risk neutral (I.e., $\gamma = 0$) cost rate of policy Π

ss, Page 85: a sequence of states with finite length

$R(C, D)$, Page 101: logical relationship: true if $\exists$ policy to drive state from C to D w.p.1

SR, Page 102: the set of all 'self-reachable' states x such that $R(x, x)$

$\sim$, Page 102: an equivalence relation: $x \sim y$ if $R(x, y)$ and $R(y, x)$

$\aleph(x)$, Page 103: the equivalence class containing x

$R^{'}(x, C)$, Page 104: C is reachable but no proper subset of C is reachable

$\beta(x)$, Page 113: admissible actions in state x that will keep the system in $\aleph(x)$ w.p.1

$\lambda^*_{\aleph(x)}$, Page 114: the optimal cost attainable while staying in the same equivalence class forever

s_i, Page 126: a single representative of an equivalence class

A, Page 126: the set of all representatives of equivalence classes

2^A, Page 126: the set of all subsets of A

$2^A(x)$, Page 126: the set of all members of 2^A that are reachable from x

B, Page 134: $B \subset A$: $s_i \in B$ if the optimal cost from s_i can be achieved without leaving $\aleph(s_i)$

$2^B(x)$, Page 134: $D \in 2^B(x)$ if $D \in 2^A(x)$ and $D \subset B$

$\mu(D)$, Page 138: the optimal cost that can be achieved by staying within the worst equivalence class represented in D

$\mho(x)$, Page 152: the set of all states from which x can be reached

$\tilde{o}$, Page 160: a sequence of states with infinite length

$\aleph^a_o(x)$, Page 160: the set of all infinite sequences of states that enter $\aleph(x)$ and stay there

$\aleph^e_o(x)$, Page 160: the set of all infinite sequences of states that hit $\aleph(x)$ infinitely many times

$\aleph_\Pi(x)$, Page 162: the set of all states that are in the equivalence class induced by Π which contains x

$\aleph^a_\Pi(x)$, Page 162: the set of all infinite sequences that enter $\aleph_\Pi(x)$ and stay there

$\aleph^e_\Pi(x)$, Page 162: the set of all infinite sequences of states that hit $\aleph_\Pi(x)$ infinitely many times

$J(s)^\Pi_x$, Page 165: the sample mean of costs over the infinite horizon

BIBLIOGRAPHY

[1] Aristotle Arapostathis, Vivek S. Borkar, Emmanuel Fernandez-Gaucherand, Mrinal K. Ghosh, and Steven I. Marcus. Discrete-time controlled markov processes with average cost criterion: A survey. *SIAM Journal of Control and Optimization*, 31(2):282–344, March 1993.

[2] S. Balaji and S.P. Meyn. Multiplicative ergodicity for an irreducible markov chain. *Stochastic Processes and their Application*, April 2000.

[3] J.S. Baras and M.R. James. Robust and risk-sensitive output feedback control for finite state machines and hidden markov models. *J. Math Sys., Estimation and Control*, to appear.

[4] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, vol. 1.* Athena Scientific, 1995.

[5] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, vol. 2.* Athena Scientific, 1995.

[6] Tomasz Bielecki, Daniel Hernandez-Hernandez, and Stanley R. Pliska. Risk sensitive control of finite state markov chains in discrete time, with applications to portfolio management. *not yet published*, February 25 1999.

[7] V.S. Borkar and S.P. Meyn. Risk sensitive optimal control: Existence and synthesis for models with unbounded cost. *submitted for publication*, February 2 1999.

[8] M. Boue and P. Dupuis. Risk-sensitive and robust escape control for degenerate processes. *Technical Report 97-14, Lefschetz Center for Dynamical Systems, Brown University, Providence, RI*, October 1997.

[9] Rolando Cavazos-Cadena and Emmanuel Fernandez-Gaucherand. Controlled markov chains with risk-sensitive criteria: Average cost, optimality equations, and optimal solutions. *to be published*, 1999.

[10] Kun-Jen Chung and Matthew J. Sobel. Discounted mdp's: Distribution functions and exponential utility maximization. *SIAM Journal of Control and Optimization*, 25(1):49–62, January 1987.

[11] Earhan Cinlar. *Introduction to Stochastic Processes.* Prentice-Hall, 1975.

[12] S. Coraluppi and S. I. Marcus. Robust control of markov decision processes and connection to risk-sensitive control. *Proc. 34th Annual Allerton Conf. on Communication, Control, and Computing, Urbana, IL*, pages 353–362, October 1-4 1996.

[13] S. Coraluppi and S.I. Marcus. Risk-sensitive, minimax, and mixed risk-neutral/minimax control of markov decision processes. In et. al. W.M. McEneaney, editor, *Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W.H. Fleming*, pages 21–40. Boston: Birkhauser, 1999.

[14] Stefano P. Coraluppi. *Optimal Control of Markov Decision Processes for Performance and Robustness.* PhD thesis, University of Maryland, College Park, 1997.

[15] P. Dupuis, M. R. James, and I. Peterson. Robust properties of risk-sensitive control. *Technical Report 98-15, Lefschetz Center for Dynamical Systems, Brown University, Providence, RI*, August 1998.

[16] E. Fernandez-Gaucherand and S.I. Marcus. Risk-sensitive optimal control of hidden markov models: Structural results. *IEEE Trans. Automatic Control*, 42:1418–1422, October 1997.

[17] K. Glover and J. Doyle. State space formulae for all stabilizing controllers that satisfy an h_∞ norm bound and relations to risk-sensitivity. *Systems and Control Letters*, 11:167–172, 1988.

[18] Daniel Hernandez-Hernandez and Steven I. Marcus. Risk sensitive control of markov processes in countable state space. *Systems and Control Letters*, 29:147–155, 1996. Correction in Systems and Control Letters 34 (1998), 105-106.

[19] Onesimo Hernandez-Lerma and Jean Bernard Lasserre. *Discrete-Time Markov Control Processes; Basic Optimality Criteria.* Springer Verlag, 1996.

[20] Onesimo Hernandez-Lerma, Oscar Vega-Amaya, and Guadalupe Carrasco. Sample-path optimality and variance-minimization of average cost markov control processes. *SIAM Journal on Control and Optimization*, 38(1), November 1999.

[21] Ronald A Howard. *Dynamic Probabilistic Systems, Volume I: Markov Processes.* John Wiley & Sons, 1971.

[22] Ronald A Howard. *Dynamic Probabilistic Systems, Volume II: SemiMarkov and Decision Processes.* John Wiley & Sons, 1971. ISBN 0-471-41666-5.

[23] Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, March 1972.

[24] J. F. C. Kingman. The ergodic theory of subadditive stochastic processes. *Journal of the Royal Statistical Society*, B(30):499–510, 1968.

[25] J. F. C. Kingman. Subadditive ergodic theory. *The Annals of Probability*, 1(6):883–909, 1973.

[26] Jean B. Lasserre. Sample-path average optimality for markov control processes. *IEEE Transactions on Automatic Control*, 44(10), October 1999.

[27] Alberto Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. Addison-Wesley, 1989.

[28] Anthony R. Cassandra Leslie Pack Kaebling, Michael L. Littman. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101: 1-2:99–134, 1997.

[29] Armand Makowski. conversation, February 2000.

[30] Steven I. Marcus, Emmanuel Fernandez-Gaucherand, Daniel Hernandez-Hernandez, Stefano Coraluppi, and Pedram Fard. Risk sensitive markov decision processes. In et. al. C. I. Byrnes, editor, *Systems and Control in the Twenty-First Century*, pages 263–279. Boston: Birkhauser, 1997.

[31] G. B. Di Masi and L. Stettner. Risk-sensitive control of discrete-time markov processes with infinite horizon. *SIAM Journal on Control and Optimization*, December 1999.

[32] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Verlag, 1993.

[33] Neil O'Connell. Large deviations with applications to telecommunications. *lecture notes for a course given at Uppsala University*, November 1999.

[34] Stephen D. Patek. On terminating markov decision processes with a risk averse objective function. *submitted to Automatica*, February 2000.

[35] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley, 1994.

[36] H.L. Royden. *Real Analysis*. MacMillan, 1963.

[37] Linn I. Sennott. *Stochastic Dynamic Programming and the Control of Queueing Systems*. John Wiley & Sons, 1999.

[38] Richard F. Serfozo. An equivalence between continuous and discrete time markov decision processes. *Operations Research*, 27(3):617–621, May-June 1979.

[39] M. A. Shayman and E. Fernandez-Gaucherand. Risk-sensitive decision-theoretic diagnosis. *not yet published*, May 1 2000.

[40] Adam Shwartz and Alan Weiss. *Large Deviations For Performance Analysis; Queues, communications, and computing.* Chapman & Hall, 1995.

[41] R. D. Smallwood and E. J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

[42] Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21:1071–1088, 1973.

[43] Peter Whittle. *Risk-sensitive Optimal Control.* Wiley, 1990.