Approximate Solution Methods for Partially Observable Markov and Semi-Markov Decision Processes

by

Huizhen Yu

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2006

Author ....
Department of Electrical Engineering and Computer Science
January 12, 2006

Certified by ....
Dimitri P. Bertsekas
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Approximate Solution Methods for Partially Observable Markov and Semi-Markov Decision Processes

by
Huizhen Yu

Submitted to the Department of Electrical Engineering and Computer Science
on January 12, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

We consider approximation methods for discrete-time infinite-horizon partially observable Markov and semi-Markov decision processes (POMDP and POSMDP). One of the main contributions of this thesis is a lower cost approximation method for finite-space POMDPs with the average cost criterion, and its extensions to semi-Markov partially observable problems and constrained POMDP problems, as well as to problems with the undiscounted total cost criterion.

Our method is an extension of several lower cost approximation schemes, proposed individually by various authors, for discounted POMDP problems. We introduce a unified framework for viewing all of these schemes together with some new ones. In particular, we establish that due to the special structure of hidden states in a POMDP, there is a class of approximating processes, which are either POMDPs or belief MDPs, that provide lower bounds to the optimal cost function of the original POMDP problem.

Theoretically, POMDPs with the long-run average cost criterion are still not fully understood. The major difficulties relate to the structure of the optimal solutions, such as conditions for a constant optimal cost function, the existence of solutions to the optimality equations, and the existence of optimal policies that are stationary and deterministic. Thus, our lower bound result is useful not only in providing a computational method, but also in characterizing the optimal solution. We show that regardless of these theoretical difficulties, lower bounds of the optimal liminf average cost function can be computed efficiently by solving modified problems using multichain MDP algorithms, and the approximating cost functions can be also used to obtain suboptimal stationary control policies. We prove the asymptotic convergence of the lower bounds under certain assumptions.

For semi-Markov problems and total cost problems, we show that the same method can be applied for computing lower bounds of the optimal cost function. For constrained average cost POMDPs, we show that lower bounds of the constrained optimal cost function can be computed by solving finite-dimensional LPs.

We also consider reinforcement learning methods for POMDPs and MDPs. We propose an actor-critic type policy gradient algorithm that uses a structured policy known as a finite-state controller. We thus provide an alternative to the earlier actor-only algorithm GPOMDP. Our work also clarifies the relationship between the reinforcement learning methods for POMDPs and those for MDPs. For average cost MDPs, we provide a convergence and convergence rate analysis for a least squares temporal difference (TD) algorithm, called LSPE, and previously proposed for discounted problems. We use this algorithm in the critic portion of the policy gradient algorithm for POMDPs with finite-state controllers.

Finally, we investigate the properties of the limsup and liminf average cost functions of various types of policies. We show various convexity and concavity properties of these cost functions, and we give a new necessary condition for the optimal liminf average cost to be constant. Based on this condition, we prove the near-optimality of the class of finite-state controllers under the assumption of a constant optimal liminf average cost. This result provides a theoretical guarantee for the finite-state controller approach.

Thesis Supervisor: Dimitri P. Bertsekas
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I thank my advisor Prof. Dimitri Bertsekas for his trust, patience, guidance and mentoring. The past two and a half years of working with him at LIDS have been the most fruitful and exciting period during my entire graduate study at MIT. I shall always treasure this experience and this time. In every aspect, he has made me better than I was before.

I thank my thesis committee. Thanks to Prof. John Tsitsiklis for his warm encouragement and helpful comments on my thesis. I deeply appreciate his time spent reading and discussing my thesis draft. Thanks to Prof. Daniela Pucci De Farias for her helpful comments and discussions before and during my thesis research, and also for her wonderful teaching on the subject of MDPs, from which my work has benefitted. Thanks to Prof. Leslie Kaelbling for all the helpful discussions. My thesis work indeed originated from my area exam on the POMDP problem supervised by Prof. Kaelbling. My research path subsequently changed from computer vision. So I would like to make special thanks to her.

I thank Prof. Sanjoy Mitter and Prof. Vivek Borkar for their interests in my research and their inspiring discussions, which led me to new investigations and discoveries. I appreciate all the knowledge they shared with me and all the helpful advice.

Before I started my current thesis, I had worked at the AI Lab, now part of the CSAIL. I would like to thank my supervisor there, Prof. Eric Grimson. Eric provided me with the research environment where I could explore and learn. I am very thankful for his support. I also thank Prof. Tommi Jaakkola at the AI Lab. I had greatly benefitted from Tommi's classes and his reading group meetings on statistics and machine learning, where he was always so patient tutoring, answering every question and puzzle, as well as sharing his deep insights. These experiences trained me and helped me become mature in research. I thank Tommi for his mentorship, and my friends from the reading group for the joys we had in sharing thoughts and the continuing friendships beyond the lab.

I thank several other faculties from LIDS, in particular, Prof. Robert Gallager for helpful discussions on information theory. I thank the staff members at MIT: Doris Inslee, Petr Swedock, Marilyn Pierce and the International Student Office for their constant support. This work was partially supported by NSF Grant ECS-0218328 and it is much appreciated.

Finally, I would like to thank all my friends from LIDS and CSAIL, from the Eastgate bible study group, and from my dormitory Green Hall for their warm friendships and constant encouragements, as well as the inspiring spiritual and intellectual experiences that they shared with me. Last but not least, I thank my parents and my sister for their caring, understanding, and unconditional support at all times. I dedicate this thesis to them.

Contents

Chapter 1

Overview

1.1 The Scope of the Thesis and Related Works

We consider discrete-time infinite-horizon partially observable Markov decision processes (POMDPs) with bounded per-stage costs. Such models have a broad range of applications, e.g., robot navigation, system control, resource allocation, experiment design and sequential coding. Our work addresses several topics relating to the exact and approximate solutions of POMDPs. In this section we give a preview of our results, and the background and earlier works to which they are related. We address lower cost approximations of the optimal cost function (Section 1.1.1), properties of the optimal solution to the average cost problem (Section 1.1.2), and approximation algorithms for POMDPs and MDPs based on reinforcement learning (Section 1.1.3).

1.1.1 Lower Cost Approximations for POMDPs with Undiscounted Cost Criteria

The main contribution of this thesis is a lower cost approximation method for finite-space POMDPs with the average cost criterion, and its extensions to semi-Markov partially observable problems (POSMDPs) and constrained POMDP problems, as well as to problems with the undiscounted total cost criterion. The part on the average cost POMDP has a preliminary version previously published in [YB04].

Relation to Theory on Optimal Solutions

Our lower cost approximation result for POMDPs with undiscounted cost criteria is a consequence of the special structure of hidden states – a property that is unique to POMDPs and not shared by general MDPs. We will show that by constructing fictitious processes that in some sense exploit the information of the hidden states, one can derive lower bounds of the optimal cost functions. Some of the lower bounds, which we call discretized lower approximations, can then be computed exactly when the spaces are finite. The property of hidden states that we use is also related to the concave preserving property of the DP mappings of POMDPs [Åst69]. Elsewhere, this concavity has been shown useful not only in deriving efficient computational methods for discounted POMDPs (e.g., [Son78, Lov91, CLZ97]), but also recently in analyzing the existence of optimal policies that are stationary and deterministic for average cost POMDPs ([HCA05]).

At present the structure of the optimal solution to POMDPs with the average cost criteria is still not well understood. The problem is significantly harder to analyze than its counterpart with the discounted cost criterion, because the average cost DP equations are only sufficient, and not necessary conditions for optimality. For the average cost criterion, the focus of research has been on the constant optimal average cost POMDP problem and on when the constant average cost DP equation has a bounded solution [Ros68, Pla80, FGAM91, ABFG$^+$93, RS94, Bor00, HCA05]. It is still hard to tell, however, whether most of the problems in practice will or will not have this nice property, from the sufficient conditions established in the literature so far. As we will show, there are very simple examples of POMDPs in which the optimal average cost function is not constant.

In light of the preceding discussions, the usefulness of our results is thus not only in suggesting computationally an approximation method, but also in characterizing the optimal solutions and in providing performance measures for other computational approaches for the average cost problems.

Relation to Discretized Approximations and Model Approximations

Our lower cost approximation approach for average cost POMDPs in fact grows out from the same approach for discounted POMDPs. There, several discretized or continuous lower approximation schemes are known. The first one was proposed by Lovejoy [Lov91] as a measure of convergence for the subgradient based cost approximation proposed in the same paper. Lovejoy's lower bound was later improved by Zhou and Hansen [ZH01], and also proposed by them as the approximate cost-to-go function for suboptimal controls. These lower bounds are based on the concavity of the optimal discounted cost functions. In reducing the computational complexity of incremental pruning – an LP based algorithm of value iteration for discounted POMDPs ([LCK96, Cas98, ZL97, CLZ97]), Zhang and Liu [ZL97] proposed a continuous approximation scheme, the "region-observable" POMDP. In the "region-observable" POMDP, to derive approximation schemes one assumes that a subset of the state space containing the true state would be revealed to the controller by a fictitious "information oracle." A different design of the partition of the state space, called "region systems", gives a different approximating POMDP, and the class of approximating processes can range from the completely observable MDP to the POMDP itself. Prior to Zhang and Liu's work, the approximation based on the value of the completely observable MDP had also been proposed by Littman, Cassandra, and Kaelbling [LCK95] to tackle large problems.

Our work can be viewed as a generalization and extension of the lower approximation approach for discounted POMDPs. We introduce a framework to characterize lower approximation schemes as a whole, including not only the previously known discretized and continuous schemes that we mentioned above, but also new lower approximation schemes that we have proposed. Conceptually, the idea in our development is close to that of the information oracle in the "region-observable" POMDP. We present two lines of analysis for establishing lower bounds. The first one combines the information oracle argument and the monotonicity property of the DP mappings. It is an intuitive and constructive approach for designing approximation schemes, and the use of the property of DP mappings is also a common technique in analyzing general MDPs. Mathematically, this line of analysis, however, leads to a weaker lower bound result, due to the dependence of the argument on the DP mappings and consequently the convergence of value iteration. Our second line of analysis subsequently alleviates this dependence, and proves a mathematically stronger lower bound result, which also applies to problems where the convergence of value iteration is not guaranteed. This again uses the hidden state structure of a POMDP.

Another key feature in our characterization based on the information oracle idea, is the model approximation view. We consider our approximation schemes as processes from modified models, and reason out the relation between the optimal solution of the modified problem and that of the original POMDP problem under various cost criteria and constraints. This model approximation view can be conceptually helpful in cases where the original problem is hard to analyze. For example, in the case of an average cost problem, it allows a line of analysis that does not rely on the existence of solution of the DP equations.

Relation to MDPs and SMDPs

The reason why lower bounds of the optimal cost functions are computable for finite space POMDPs or POSMDPs, is because the modified problem corresponding to a discretized lower approximation scheme can be viewed essentially as a finite state and action MDP or SMDP – with additional infinitely many transient states – on the belief space. The development of our lower cost approximation method thus has heavily relied on and benefited from the completeness of the theories and algorithms for finite space MDPs and SMDPs.

In particular, there are two works important to our analysis that we would like to mention. One of these works is the sensitive optimality theory, developed by Veinott, Miller and many others (see the bibliography remarks of Chapter 10 of Puterman [Put94]). It enables us to treat the average cost and total cost cases together, and it also has other uses in our analysis. Another related result is the multichain average cost algorithms for MDPs, which enable us to compute the lower bounds without imposing any conditions on either the modified problem or the original POMDP problem.

We now address the difference between the way we derive lower bounds for POMDPs and corresponding approaches in MDPs. As mentioned earlier, we use the unique feature of hidden states of a POMDP, which does not apply to an MDP. One technique of deriving lower or upper bounds in an MDP is based on the violations of the optimality equations, also called Bellman residues. To apply this technique, one still need methods of approximating the average and differential cost functions, which are not addressed by the technique itself. Furthermore, aside from the fact that the bounds based on Bellman residues tend to be loose and hard to compute in the case of POMDP, for average cost problems these bounds are constant and hence not useful in the case where the POMDP has a non-constant optimal average cost. Another result in an average cost MDP that is related to lower bounds, is based on the fact that the optimal cost is the maximal solution to some inequalities associated with the average cost optimality equations. This is however more useful in analysis than computation, because finding a solution satisfying the inequalities is very hard in the case of a POMDP.

Other Related Works

We also note several works somewhat related yet substantially different from ours. For discounted and finite-horizon undiscounted MDPs with application to POMDPs, the work by Lincoln and Rantzer [LR02] and Rantzer [Ran05] contains a lower cost approximation method which is based on modifying the per-stage cost function and is computed through value iteration. For discounted problems, discretized approximations that are not necessarily lower bounds of the optimal are standard, and some related analysis can be found in e.g., [Ber75, Bon02].

The book by Runggaldier and Stettner [RS94] considers approximation methods (including both belief approximation and cost approximation) for general space POMDPs under discounted and average cost criteria, and contains comprehensive analysis. For the average cost criterion Runggaldier and Stettner assumed a strong condition to ensure the existence of a bounded solution to the constant average cost DP equation and the convergence of the vanishing discount approach. They also suggested to use a near-optimal policy for a discounted problem with sufficiently large discount factor as the suboptimal control in the average cost problem, and proved the asymptotic convergence of the cost of these policies under the assumption of a bounded solution to the constant average cost DP equation. The convergence of costs of policies is an important issue. Since our analysis only established convergence of the cost approximation but not that of the cost of the policies obtained by solving an average cost modified problem, their results suggest, at least theoretically, that for suboptimal control it may be more reliable to follow their approach under certain assumptions.

For average cost MDPs with a continuous state space, the work by Ormoneit and Glynn [OG02] is the closest to ours, and indeed some of our POMDP discretization schemes can be viewed as special cases of their general discretized approximation schemes. However, the work of Ormoneit and Glynn addresses a different context, one of approximating the expectation by a sample mean and approximating the DP mapping by a random mapping. Thus their schemes do not rely on the lower approximation property as ours do. Furthermore, their work has the usual recurrence assumptions for a general MDP, which are not satisfied by a POMDP, therefore the analysis on various optimality and convergence issues are also different in our work and theirs.

Computing lower bounds based on the information argument had appeared earlier in a different field, information theory. In analyzing the entropy rate of stationary hidden Markov sources, lower bounds of the entropy rate are established based on the fact that information reduces entropy (see Cover and Thomas [CT91]). That method corresponds to a special case of the information oracle. Since the entropy rate problem can be reduced to a control problem with average cost criterion, our work provides new methods of computing lower bounds for this problem as well as problems of the same kind, e.g., sequential coding problems.

1.1.2 On the Optimal Average Cost Functions

Some of our work also relates to the exact solution of the average cost POMDP problem with finite space models. These include:

(i) Examples of POMDPs with a non-constant optimal average cost function while the associated completely observable MDPs are recurrent and aperiodic.

(ii) A necessary condition for a constant optimal liminf average cost function, (which is also the necessary condition for the constant average cost DP equation to have a bounded solution), based on the concavity of the optimal liminf cost.

Platzman [Pla80] gave an interesting example in which the optimal average cost of the POMDP is constant, but there is no solution to the constant average cost DP equation. The type of non-constant optimal average cost examples in (i) has not been addressed before, and in fact one of our examples serves also as a counter-example to an earlier result on the average cost problem in the literature [Bor00].

The concavity of the optimal liminf average cost function and the necessary condition in (ii) are new results also. The necessary condition is somewhat peculiar. It states that if the optimal liminf average cost function is constant, then for any $\epsilon > 0$ there exists a history dependent randomized policy that does not depend on the initial distribution of the state (i.e., only depend on the past observations and controls), but has an ϵ-optimal liminf average cost for all initial distributions. Based on the existence of such policies, we can further conclude that if the optimal liminf average cost is constant, then the optimal limsup average cost equals the same constant.

Results (i) and (ii) are complementary to existing results (e.g., [Pla80, FGAM91, HCA05]) on when the optimal average cost is constant and when the constant average cost DP equation has a bounded solution. Furthermore, (i) and (ii) suggest in some sense that it might be relatively strong to assume for a POMDP that the constant average cost DP equation admits a bounded solution.

1.1.3 Reinforcement Learning Algorithms

In the last part of this thesis (Chapters 10 and 11) we consider approximation algorithms for finite space POMDPs and MDPs under the reinforcement learning setting. This setting differs from the previous problem setting – often referred as the planning setting – in that a model of the problem is no longer required, and neither is an exact inference mechanism. The motivation as well as inspiration behind these methods comes from animal learning and artificial intelligence, and is to build an autonomous agent capable of learning good policies through trial and error while it is operating in the environment (thus the name “reinforce”). The reinforcement learning setting does require, however, that the per-stage costs or rewards are physically present in the environment, which is not true for many planning problems. So the reinforcement learning and the planning settings are complementary to each other. Via simulations of the model, reinforcement learning algorithms can be applied in the planning setting to obtain approximate solutions. For problems whose models are too large to handle exactly and explicitly, such methods are especially useful.

The reinforcement learning methods for POMDPs are so far built upon the same methodology for MDPs, for which there has been a rich literature developed in recent years. For value iteration and policy iteration methods, most algorithms and their analyses can be found in the books by Bertsekas and Tsitsiklis [BT96], Sutton and Barto [SB98], and other works, e.g., [TV97, TV99, Kon02]. For policy gradient methods, the idea and detailed analysis can be found in e.g., [GL95, CW98, SMSM99, KT99, BB01, MT01, Kon02].

Computationally, our contributions in this field are on analysis and extensions of two algorithms. One is the policy gradient estimation algorithm for POMDPs and POSMDPs with a subset of policies called finite-state controllers, and the other is a least squares temporal difference algorithm called LSPE. The part on policy gradient in POMDPs was published previously in [Yu05].

Theoretically, our contribution to the finite-state controller approach is a proof of near-optimality of the class of finite-state controllers in the average cost case under a constant optimal average cost assumption. The result however is established through our analysis on the optimal average cost functions, and is not based on the theory of reinforcement learning algorithms.

Gradient Estimation for POMDPs and POSMDPs with Finite-State Controllers

The finite-state controller approach is a generalization of the finite memory (i.e., finite length of history window) approach for POMDPs. Like a probabilistic automaton, the internal state of the finite-state controller evolves probabilistically depending on the recent history, and the controller outputs a randomized control depending on its internal state.

An important property is that under a finite-state controller, the state and observation of the POMDP, and the internal state of the controller jointly form a Markov chain. Hence the asymptotic behavior of the POMDP is well understood based on the MDP theory, (even though the states are not observable), and consequently algorithms for MDPs are applicable, when we take additional care of the hidden states. This property, when compared to the difficulties in understanding the optimal policies for the average cost POMDP, makes the finite-state controller a highly appealing approach.

There are also other attractive features in the finite-state controller approach. The framework can incorporate naturally approximate inference; and it can separate the true mathematical model from the subjectiveness, such as model uncertainty and subjective prior distributions, by treating the latter as parameters of the controller.

Theoretically, the question of near-optimality of the class of finite-state controllers in the average cost case has not been resolved, however. An interesting new result of this thesis is that there exists an ϵ-optimal finite-state controller for any $\epsilon > 0$, under the condition that the optimal liminf average cost is constant.

The main computational methods to optimize over the set of finite-state controllers have been policy gradient type methods (Baxter and Bartlett [BB01], Aberdeen and Baxter [AB02]). There have been different opinions towards policy gradient methods in general as to their effectiveness and efficiency compared to model-based methods, largely due to issues on local minima, stochastic noises and convergence speed. Several recent works focus on variance reduction techniques for policy gradient algorithms (Henderson and Glynn [HG01], Greensmith, Bartlett and Baxter [GBB04]). On the other hand, whether a policy is obtained from a model-based approximation method or any other method, the policy gradient approach provides a generic way of further policy improvement.

Our contribution to POMDPs with finite-state controllers is to propose and analyze a gradient estimation algorithm that uses a value function approximator. Policy gradient algorithms proposed prior to our work for finite-state controllers belong to the so called actor-only type of methods ([BB01, AB02] for POMDPs and Singh, Tadic and Doucet [STD02] for POSMDPs). Overcoming a slight technical difficulty of hidden states that has perhaps been neglected previously, our result finally shows that the Actor-Critic framework developed for MDPs [SMSM99, KT99, Kon02] carries through to the case of POMDPs with finite-state controllers, and algorithmically, the latter can be viewed as a special case.

On LSPE Algorithm

Our second contribution in the area of approximation algorithms is on LSPE, a least squares temporal difference algorithm for policy evaluation in MDPs. The LSPE algorithm was first proposed by Bertsekas and Ioffe [BI96] without convergence proofs, and was analyzed for a diminishing stepsize by Nedić and Bertsekas [NB03], and for a constant stepsize by Bertsekas, Borkar and Nedić [BBN03]. Extending the convergence proof in [BBN03] for the discounted case, we prove the convergence of LSPE with a constant stepsize for the average cost criterion. Furthermore, we prove that for both discounted and average cost criteria, the asymptotic convergence rate of LSPE is the same as that of the LSTD algorithm, proposed initially by Bradtke and Barto [BB96] and extended by Boyan [Boy99]. The LSTD algorithm is another least squares type algorithm differing from LSPE. It is proved by Konda [Kon02] that LSTD has the optimal asymptotic convergence rate compared to other TD algorithms. The convergence rate of LSPE has not been addressed before, and its expression is not obvious due to the non-linearities involved in the LSPE updates. Thus the result of Konda [Kon02] on LSTD has provided us with a shortcut in proving the optimal asymptotic convergence rate of LSPE.

1.2 Contents and Organization

This thesis consists of two parts: Chapter 2 to Chapter 9 on POMDPs under the planning framework in which the model is assumed given, and Chapter 10 to 11 on reinforcement learning algorithms.

Chapter 2: Introduction to POMDP

In Chapter 2 we first give a brief introduction of POMDPs with general space models, including definition of models, induced stochastic processes, optimality criteria and the equivalent belief MDP as well as measurability issues. We then introduce the average cost POMDP problem for finite space models, and discuss the associated difficulties and recent progress. Along with the introduction we also show a few related miscellaneous results, including two examples of POMDPs with a non-constant optimal average cost function, and a somewhat peculiar necessary condition for a constant optimal liminf average cost. The necessary condition further leads to a stronger claim on optimal average cost functions and a proof of near-optimality of the class of finite-state controllers under the constant optimal liminf cost assumption – we show these results in Appendix D.

Chapter 3-5: Lower Cost Approximations for Discounted and Average Cost POMDPs

Chapter 3 considers general space POMDPs and establishes one of the main results of this thesis: due to the special structure of hidden states in POMDPs, there is a class of processes – which are either POMDPs themselves or belief MDPs – that provide lower bounds to the optimal cost functions of the original POMDP problem for either the discounted, finite-stage undiscounted, or average cost criteria. This chapter presents two lines of analysis, which lead first to a weaker and then to a strengthened lower bound result. This lays the foundation for the subsequent chapters up to Chapter 9.

Chapter 4 and Chapter 5 specialize the results of Chapter 3 to POMDPs with finite space models for which certain lower bounds can be computed exactly, and their focus is on issues in computing lower approximations and in analyzing approximation error and suboptimal controls. Chapter 4 considers the discounted case and summarizes the discretized lower approximation schemes and asymptotic convergence issues there. Chapter 5 considers the average cost case. It proposes to solve the modified problems under sensitive optimality criteria such as n-discount optimality for obtaining lower bounds and suboptimal controls for the original problem. It presents in details the algorithms, the approximation error and asymptotic convergence analyses. It shows that if the constant average cost DP equation of the original POMDP admits a bounded solution with the differential cost function being continuous, then the average cost approximations from the discretized schemes asymptotically converge to the optimal average cost.

Chapter 6: Extension to Semi-Markov Problems

Chapter 6 considers the more general partially observable semi-Markov decision processes (POSMDPs). Its developments parallel those from Chapter 3 to Chapter 5, presenting analogous results on lower cost approximation schemes and on algorithms solving modified problems for finite space models to obtain lower bounds of the optimal cost functions under the discounted or average cost criteria.

This chapter also addresses an application of the approximation results of POSMDPs to POMDP problems with certain hierarchical controllers, in which one can obtain for the POMDP problem lower bounds on the optimal cost over the subset of policies by transforming the POMDP into a POSMDP.

Chapter 7: Extension to Total Cost Problems

Chapter 7 considers the implication of the lower bound results of Chapter 3 on the expected undiscounted total cost POMDP problems for finite space models. It analyzes three cases: non-negative, non-positive and general per-stage cost models. The main results are:

(i) For non-negative and non-positive per-stage cost models, the differential cost of a 0-discount optimal policy of the modified problem with a zero optimal average cost is a lower bound of the optimal total cost of the original POMDP problem.

(ii) For non-negative per-stage cost models, under the assumption that the optimal total cost of the original POMDP is finite, the discretized lower cost approximations asymptotically converge to the optimal.

(iii) For general per-stage cost models, the differential cost of a 0-discount optimal policy of the modified problem with a zero optimal average cost is a lower bound of the optimal limsup total cost of the original POMDP problem.

Chapter 8: Applications of Lower Bounds

Through example applications, Chapter 8 illustrates that the lower bound result for average and total cost POMDPs can be applied to several problems, for which the use of the discretized lower cost approximations is not in providing suboptimal control policies, but in providing lower bounds and approximations to quantities of interest.

First, it demonstrates that problems such as reaching, avoidance and model identification can be cast as POMDP problems, and lower bounds of the optimal average cost or total cost can then be used to bound the probabilities or expected values of interest.

Secondly, it considers a classic problem on hidden Markov sources in information theory, and demonstrates that the discretized lower approximation approach provides a new method of computing lower bounds of the entropy rate of hidden Markov sources. Asymptotic convergence issues are also addressed and proved under certain conditions. Applications of the same type include sequential coding.

Chapter 9: Lower Bounds for Constrained Average Cost POMDPs

Chapter 9 investigates the applicability of the discretized lower approximation schemes, so far established for unconstrained problems, to constrained average cost problems with finite spaces. The constrained problems under consideration are POMDPs with multiple per-stage cost functions and with the objective being to minimize the limsup average cost with respect to one per-stage cost function, subject to prescribed constant bounds on the limsup average costs with respect to the other per-stage cost functions.

Chapter 9 describes and analyzes algorithms for solving the constrained modified problems (associated with discretized lower approximation schemes). The main results are

(i) A constant lower bound of the constrained optimal cost function can be computed by solving a single finite-dimensional LP, which is the unichain LP of the modified problem (regardless of its chain structure).

(ii) If the modified problem is multichain, a non-constant lower bound of the constrained optimal cost function can be computed for each initial distribution, by solving a finite-dimensional LP associated with that distribution.

(iii) If any one of the LPs is infeasible, then the original constrained problem is infeasible.

Chapter 10-11: Algorithms for Reinforcement Learning

Chapter 10 considers the estimation of the policy gradient in POMDPs with a special class of structured policies called finite-state controllers. It extends the approach of an earlier method GPOMDP, an actor-only method, and shows using ergodicity that policy gradient descent for POMDPs can be done in the Actor-Critic framework, by making the critic compute a “value” function that does not depend on the states of a POMDP. This function is the conditional mean of the true value function that depends on the states. The critic can be implemented using temporal difference (TD) methods with linear function approximations, and the analytical results on TD and Actor-Critic can be transfered to this case. Furthermore, it is shown that the same idea applies to semi-Markov problems with a subset of finite-state controllers.

Chapter 11 considers finite space MDPs and proves some convergence results for the LSPE algorithm, a least squares policy evaluation algorithm. In particular, it proves the convergence of the average cost LSPE with a constant stepsize, and the asymptotically optimal convergence rate of LSPE for both discounted and average cost cases.

Chapter 2

Introduction to the Problem of POMDPs with General Space Models

A Markov decision process (MDP), also called a controlled Markov chain, is a sequential decision problem in which the state evolves in a Markov way given the past states and controls, and the goal is to minimize certain long term costs by controlling the evolution of the states. The partially observable problem, POMDP, considers the case where the states are no longer observable, but are partially "observable" through the observations they generate. Though the dynamics of state evolution at every single stage is the same in a POMDP as in an MDP, the fact that the controller in a POMDP does not have perfect information about the states causes both analytical and computational difficulties to the POMDP problem.

In this chapter, we give a brief introduction of POMDPs with general spaces. (General space models will be used in Chapter 3 and Chapter 6, while discrete space models will be used in the rest of the thesis.) We will start with reviewing the definition of the POMDP problem, the induced stochastic processes, notions of optimality under different cost criteria, the equivalent belief MDP formulation, and optimality equations as well as measurability issues. We will then briefly introduce the average cost POMDP problem and its difficulties in Section 2.4.

Along with the introduction, we will also give a few miscellaneous results that do not seem to be known, or completely known. They are:

- concavity of the optimal liminf average cost function for general space models, and Lipschitz continuity of the optimal liminf and limsup average cost functions for discrete (i.e., finite or countably infinite) state space models (Section 2.2.2);

- examples of a finite space POMDP with non-constant optimal average cost when the associated MDP is recurrent and aperiodic (Section 2.4.2); and

- for finite state space models, a somewhat peculiar necessary condition for the optimal liminf average cost function being constant (Section 2.4.3), which will lead to a stronger claim on the optimal average cost functions for a finite space POMDP, as well as a proof of near-optimality of the class of finite-state controllers (Appendix D).

2.1 POMDP Model and Induced Stochastic Processes

Let $\mathcal{S}$, $\mathcal{Y}$ and $\mathcal{U}$, called *state, observation and control* spaces, respectively, be Borel measurable sets of complete separable metric spaces.

For a metric space X, let $\mathcal{B}(X)$ denote the Borel σ-algebra of X, (i.e., σ-algebra generated by open balls); let $\mathcal{P}(X)$ denote the set of probability measures on $\mathcal{B}(X)$; and let the metric on $\mathcal{P}(X)$ be the Prohorov metric.[1] Define functions called transition probabilities[2] as follows.

Definition 2.1. Let X and X' be separable metric spaces. We say a function $P(\cdot, \cdot) : X \times \mathcal{B}(X') \to [0, 1]$ is a *transition probability* from $(X, \mathcal{B}(X))$ to $(X', \mathcal{B}(X'))$, (abbreviated "X to X'"), if (i) for each $x \in X$, $P(x, \cdot)$ is a measure on $\mathcal{B}(X')$, and (ii) for each $A \in \mathcal{B}(X')$, $P(x, A)$ as a function of x is Borel measurable. We say that P is a *continuous* transition probability, if $x \to y$ implies that $P(x, \cdot) \to P(y, \cdot)$.

We consider a discrete-time and time-homogenous POMDP. Its state, observation and control at time t are denoted by S_t, Y_t and U_t, respectively. Its model, including the dynamics of state evolution, the generation of observations and the per-stage cost function, can be specified by a six-tuple $< \mathcal{S}, \mathcal{Y}, \mathcal{U}, P_S, P_Y, g >$, where

- $P_S((s, u), \cdot)$, called *state transition probability*, is a transition probability from $\mathcal{S} \times \mathcal{U}$ to $\mathcal{S}$, and it specifies the evolution of the state S_t given the previous state and control (S_{t-1}, U_{t-1});

- $P_Y((s, u), \cdot)$, called *observation probability*, is a transition probability from $\mathcal{S} \times \mathcal{U}$ to $\mathcal{Y}$, and it specifies the generation of the observation Y_t given the current state and the previous control (S_t, U_{t-1}); and

- $g : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$, called the *per-stage cost function*, is a real-valued Borel measurable function, and it specifies the per-stage cost $g(s, u)$ when the current state is s and control u.

Throughout the thesis, we will assume boundedness of the per-stage cost function. We also assume that for every state all controls are admissible, which is common practice in the POMDP field.

The decision rules that specify how controls are applied at every time t, are called policies. To define policies, first, we define $\mathcal{H}_t, t \geq 0$, called *history sets*, recursively as

$$\mathcal{H}_0 = \emptyset, \qquad \mathcal{H}_t = \mathcal{H}_{t-1} \times \mathcal{U} \times \mathcal{Y}.$$

The set $\mathcal{H}_t$ is the space of observed histories, which consist of controls and observations up to time t, $(U_0, Y_1, \ldots, U_{t-1}, Y_t)$, prior to U_t being applied. The most general set of policies can be defined as follows.

[1] Let P and Q be two laws on a metric space X with metric d. The Prohorov metric $\rho(P, Q)$ is defined by

$$\rho(P, Q) = \inf\{\epsilon : P(A) \leq Q(A^\epsilon) + \epsilon, \forall A \in \mathcal{B}(X)\}, \quad \text{where } A^\delta \overset{def}{=} \{x \in X : d(x, A) < \delta\}.$$

The Prohorov metric metrizes the convergence of laws. (See Dudley [Dud89].)

[2] While they are called transition probabilities by us and some authors, they are also called by other authors stochastic transition kernels, or conditional probabilities.

- We call $\pi = (\mu_t)_{t\geq 0}$ a *policy*, where $(\mu_t)_{t\geq 0}$ is a collection of functions such that for each t, $\mu_t(\cdot,\cdot)$ is a transition probability from $\mathcal{H}_t$ to $\mathcal{U}$. We say that π is a *history dependent randomized policy.*

- If for every t and $h \in \mathcal{H}_t$ the probability measure $\mu_t(h,\cdot)$ has all of its mass on one single control, we say that the policy π is *deterministic*.

- We denote the set of all history dependent randomized policies by Π.

The set Π is indeed the common set of admissible policies for *every* initial distribution. (Later we will introduce policies that also functionally depend on the initial distribution; such a policy can be viewed as the set of pairs $\{(\xi, \pi_\xi) | \xi \in \mathcal{P}(\mathcal{S}), \pi_\xi \in \Pi\}$, i.e., for every initial distribution one policy from Π is selected.)

Induced Stochastic Processes

The state, observation and control sequence $\{S_0, U_0, (S_t, Y_t, U_t)_{t\geq 1}\}$ is to be defined formally as a stochastic process induced by an initial distribution and a given policy. Definitions of expected cost and notions of optimality will be given after this.

Let $(\Omega, \mathcal{F})$ be the canonical sample space for $\{S_0, U_0, (S_t, Y_t, U_t)_{t\geq 1}\}$ with the Borel σ-algebra $\mathcal{F} = \mathcal{B}(\Omega)$ and

$$\Omega = \mathcal{S} \times \mathcal{U} \times \prod_{t=1}^{\infty} (\mathcal{S} \times \mathcal{Y} \times \mathcal{U}) \, .$$

With a sample $\omega = (s_0, u_0, \ldots, s_t, y_t, u_t, \ldots) \in \Omega$, the random variables are defined as the projections of ω on their respective spaces:

$$S_t(\omega) = s_t, \quad t \geq 0; \qquad U_t(\omega) = u_t, \quad t \geq 0; \qquad Y_t(\omega) = y_t, \quad t \geq 1.$$

Let $\xi \in \mathcal{P}(\mathcal{S})$ be the initial distribution, and let $\pi = (\mu_t)_{t\geq 0} \in \Pi$ be a history dependent randomized policy. There exists a probability measure $\mathbb{P}^{\xi,\pi}$ on $(\Omega, \mathcal{F})$, induced by ξ and π, that is consistent with the transition probabilities, i.e.,

$$\begin{aligned}
&\qquad \mathbb{P}^{\xi,\pi}(S_0 \in \cdot) = \xi(\cdot), \qquad \mathbb{P}^{\xi,\pi}(U_0 \in \cdot) = \mu_0(\cdot), \\
&\forall k \geq 1 : \\
&\qquad \mathbb{P}^{\xi,\pi}(S_k \in \cdot \mid (S_t, U_t, Y_t)_{t<k}) = \mathbb{P}^{\xi,\pi}(S_k \in \cdot \mid S_{k-1}, U_{k-1}) = P_S\left((S_{k-1}, U_{k-1}), \cdot\right), \\
&\qquad \mathbb{P}^{\xi,\pi}(Y_k \in \cdot \mid (S_t, U_t, Y_t)_{t<k}, S_k) = \mathbb{P}^{\xi,\pi}(Y_k \in \cdot \mid S_k, U_{k-1}) = P_Y\left((S_k, U_{k-1}), \cdot\right), \\
&\qquad \mathbb{P}^{\xi,\pi}(U_k \in \cdot \mid (S_t, U_t, Y_t)_{t<k}, S_k, Y_k)(\omega) = \mathbb{P}^{\xi,\pi}(U_k \in \cdot \mid \mathcal{F}_k)(\omega) = \mu_k\left(h_k(\omega), \cdot\right),
\end{aligned}$$

where $\{\mathcal{F}_k\}$ is an increasing sequence of σ-algebras generated by the past controls and observations prior to U_k being applied:

$$\mathcal{F}_0 = \{\emptyset, \Omega\}, \qquad \mathcal{F}_k = \sigma\left(U_0, Y_1, \ldots, U_{k-1}, Y_k\right), \quad k \geq 1,$$

and $h_k(\omega) = (u_0, y_1, \ldots, u_{k-1}, y_k) \in \mathcal{H}_k$ is the sample trajectory of the controls and observations up to time k and prior to U_k.

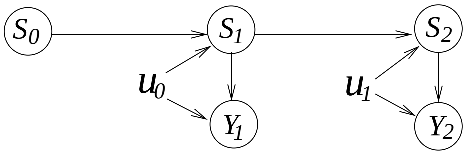

Figure 2-1: The graphical model of a partially observable Markov decision process.

Graphical Model and POMDP

It is succinct to represent a POMDP by a graphical model shown in Fig. 2-1. Since we will use this graph representation later in several places, we explain in some detail here the representation and the way we use it in a decision process.

The graphical model is a way of specifying the conditional independence structure of random variables by a graph. A reference to graphical models is [Lau96]. The type of graphs we will be using are directed and acyclic. In some cases the graph coincides with the actual physical mechanism that generates the random variables, so directed graphs are also called generative models.

A *directed acyclic* graph specifies the *form* of the joint distribution of a finite collection of random variables. Let $\mathcal{V} = \{V^1, \dots, V^n\}$ be the vertices of the graph. Each $V \in \mathcal{V}$ corresponds to a random variable in the stochastic process. Let V_{pa} denote the *parents* of V, that is, vertices adjacent to the incoming edges of V. We say that a joint distribution $P\left(V^1, \dots, V^n\right)$, or the stochastic process $(V^1, \dots, V^n)$, is *consistent* with the graph if there is a set of transition probabilities $\{P_V(v_{pa}, \cdot)\}$ from the space of V_{pa} to the space of V such that for any Borel set A in the product space of the spaces of V^i, $i \leq n$,

$$P\left((V^1, \dots, V^n) \in A\right) = \int \cdots \int \mathbf{1}_A(v^1, \dots, v^n) P_{V^1}(v^1_{pa}, dv^1) \dots P_{V^n}(v^n_{pa}, dv^n),$$

where $\mathbf{1}_A(x)$ denotes the indicator function of the event $\{x \in A\}$.

Hence, a stochastic process can be fully specified by a graph with which it is consistent and the collection of transition probabilities $\{P_V(v_{pa}, \cdot)\}$ associated with the graph. Let V_{an}, called *ancestors* of V, be the set of vertices that have downward paths to V. The graph representation is then equivalent to and more succinct than the statements "$\mathbb{P}(V \in \cdot | V_{an}) = \mathbb{P}(V \in \cdot | V_{pa})$" for each vertex V in specifying the stochastic process.

In the case of a POMDP, which is a decision process, the graph in Fig. 2-1 specifies the form of the joint distributions of those circled random variables, *conditioned on* the non-circled variables, (here the controls). The six-tuple model of a POMDP specifies the parameters P_S and P_Y associated with the directed edges. When a policy is chosen, it specifies the edges and parameters associated with U_t, and we then have a full graph representation of the stochastic process $\{(S_t, U_t, Y_t)\}$ corresponding to a POMDP under that policy. In the induced stochastic process, certain conditional independence statements, such as "$\mathbb{P}(Y_1 \in \cdot | S_0, U_0, S_1) = \mathbb{P}(Y_1 \in \cdot | U_0, S_1)$," can then be "read" off easily from the graphical model.

2.2 Cost Criteria

We consider primarily two expected cost criteria, discounted cost and average cost, for infinite-horizon problems. Throughout the thesis, we assume a bounded per-stage cost function.

Assumption 2.1. *There is a constant $L > 0$ such that $|g(s,u)| \leq L$, for all s, u.*

2.2.1 Definitions

Discounted cost

Let $\beta \in [0, 1)$ be called a *discount factor.* In discounted infinite-horizon problems, the expected total discounted cost J_β^π of a policy $\pi \in \Pi$ for an initial distribution ξ is defined by

$$J_\beta^\pi(\xi) = E^{\mathbb{P}^{\xi,\pi}} \Big\{ \sum_{t=0}^{\infty} \beta^t g(S_t, U_t) \Big\},$$

and the optimal cost function $J_\beta^*(\xi)$ is correspondingly defined by

$$J_\beta^*(\xi) = \inf_{\pi \in \Pi} J_\beta^\pi(\xi). \tag{2.1}$$

The notation $E^{\mathbb{P}}$ is used to specify that the expectation is taken with respect to the probability measure $\mathbb{P}$. Since we will deal with different probability measures induced by different policies, we shall keep using this notation until it can be simplified without ambiguity.

Average Cost

The average cost of a policy is defined by the limits of the long-run average of its expected cost in the finite-horizon problem. Let the expected k-stage cost $J_k^\pi(\xi)$ of a policy $\pi \in \Pi$ for an initial distribution ξ be defined by

$$J_k^\pi(\xi) = E^{\mathbb{P}^{\xi,\pi}} \Big\{ \sum_{t=0}^{k-1} g(S_t, U_t) \Big\}.$$

The optimal k-stage cost function $J_k^*(\xi)$ is defined by

$$J_k^*(\xi) = \inf_{\pi \in \Pi} J_k^\pi(\xi). \tag{2.2}$$

The limit $\lim_{k \to \infty} \frac{1}{k} J_k^\pi(\xi)$ does not necessarily exist. So define the *liminf* and *limsup average costs* of a policy by:

$$J_-^\pi(\xi) = \liminf_{k \to \infty} \frac{1}{k} J_k^\pi(\xi), \qquad J_+^\pi(\xi) = \limsup_{k \to \infty} \frac{1}{k} J_k^\pi(\xi),$$

which are the asymptotically best and worst, respectively, long-run average cost under the policy π. Correspondingly, the *optimal liminf* and *limsup* average cost functions are defined by

$$J_-^*(\xi) = \inf_{\pi \in \Pi} J_-^\pi(\xi), \qquad J_+^*(\xi) = \inf_{\pi \in \Pi} J_+^\pi(\xi). \tag{2.3}$$

In the literature, by the optimal average cost function we mean the optimal limsup cost function, and by the optimal policy we mean the one whose limsup cost equals $J_+^*(\xi)$. We will follow this convention, and when both optimal limsup and liminf cost functions are of interest, we will address limsup and liminf explicitly.

2.2.2 Concavity and Continuity Properties of Cost Functions

When the state space $\mathcal{S}$ is discrete (i.e., finite or countably infinite), it is easy to show that for all $\xi, \bar{\xi} \in \mathcal{P}(\mathcal{S})$, the β-discounted and k-stage cost functions of policy π satisfy, respectively,

$$|J_\beta^\pi(\xi) - J_\beta^\pi(\bar{\xi})| \leq \frac{L}{1-\beta}\rho(\xi, \bar{\xi}), \qquad |J_k^\pi(\xi) - J_k^\pi(\bar{\xi})| \leq Lk\rho(\xi, \bar{\xi}), \tag{2.4}$$

where $L = 2\max_{s,u}|g(s,u)|$, and $\rho(\cdot,\cdot)$ is the Prohorov metric on $\mathcal{P}(\mathcal{S})$ (where $\mathcal{S}$ is endowed with a discrete topology and a distance function, e.g., $d(s,s') = 0$ if and only if $s = s'$, and $d(s,s') = 1$ otherwise.).

Proposition 2.1. *The optimal β-discounted cost function $J_\beta^*(\xi)$ and the optimal k-stage cost function $J_k^*(\xi)$ are concave in ξ. If $\mathcal{S}$ is discrete, then they are also Lipschitz continuous.*

Proof: It is obvious that $J_\beta^\pi(\xi)$ and $J_k^\pi(\xi)$ are linear functions of ξ. Therefore, as the pointwise infimum of linear functions, $J_\beta^*(\xi) = \inf_{\pi\in\Pi} J_\beta^\pi(\xi)$ and $J_k^*(\xi) = \inf_{\pi\in\Pi} J_k^\pi(\xi)$ are concave in ξ.

When $\mathcal{S}$ is discrete, by Eq. (2.4) $\{J_\beta^\pi(\cdot) \mid \pi \in \Pi\}$ and $\{J_k^\pi(\cdot) \mid \pi \in \Pi\}$ are families of uniformly bounded Lipschitz continuous functions each with the same Lipschitz constant, hence $J_\beta^*(\cdot)$ and $J_k^*(\cdot)$ are Lipschitzian with the same Lipschitz constant $\frac{L}{1-\beta}$ and Lk, respectively.[3] □

Remark 2.1. There are two other alternative ways to prove the concavity of the optimal cost functions for general space models (under the bounded per-stage cost assumption), which we will mention in the next chapter. These proofs are based on the optimality equations (see the next section). The above proof, which is based on induced stochastic processes, seems the easiest.

We now study the continuity and concavity properties of the average cost functions. By the second equation of (2.4), when $\mathcal{S}$ is discrete, we have

$$\frac{1}{k}|J_k^\pi(\xi) - J_k^\pi(\bar{\xi})| \leq L\rho(\xi, \bar{\xi}), \tag{2.5}$$

i.e., for every policy $\pi \in \Pi$, all functions in the set $\{\frac{1}{k}J_k^\pi(\cdot) \mid k \geq 1\}$ are Lipschitz continuous with the same Lipschitz constant L.

[3] Let X be a subset of $\mathcal{R}^n$ and $\{f_i \mid f_i : X \to \mathcal{R}, i \in I\}$ be an arbitrary collection of uniformly bounded Lipschitz continuous functions each with Lipschitz constant C. The pointwise infimum function $f(x) = \inf_{i\in I} f_i(x)$ is Lipschitz continuous with the same Lipschitz constant: for any $i \in I$ and $x_1, x_2 \in X$,

$$f_i(x_1) \leq f_i(x_2) + Cd(x_1,x_2) \quad \Rightarrow \quad f(x_1) \leq f(x_2) + Cd(x_1,x_2);$$

similarly, one can show $f(x_2) \leq f(x_1) + Cd(x_1,x_2)$; and therefore $|f(x_1) - f(x_2)| \leq Cd(x_1,x_2)$.

Proposition 2.2. *The optimal liminf average cost function $J_-^*(\xi)$ is concave in ξ. If $\mathcal{S}$ is discrete, then the optimal liminf and limsup average cost functions $J_-^*(\xi)$ and $J_+^*(\xi)$ are Lipschitz continuous on $\mathcal{P}(\mathcal{S})$ with Lipschitz constant L.*

Proof: It is obvious that $\frac{1}{k}J_k^\pi(\xi)$ is a linear function of ξ, i.e.,

$$\frac{1}{k}J_k^\pi(\xi) = \sum_{i=1}^m \alpha_i \frac{1}{k} J_k^\pi(\xi_i)$$

for any convex combination $\xi = \sum_{i=1}^m \alpha_i \xi_i$. By the non-negativity of α_i and the inequality

$$\liminf_{n\to\infty} (a_n + b_n) \geq \liminf_{n\to\infty} a_n + \liminf_{n\to\infty} b_n,$$

$$J_-^\pi(\xi) = \liminf_{k\to\infty} \sum_{i=1}^m \alpha_i \frac{1}{k} J_k^\pi(\xi_i) \geq \sum_{i=1}^m \alpha_i \liminf_{k\to\infty} \frac{1}{k} J_k^\pi(\xi_i) = \sum_{i=1}^m \alpha_i J_-^\pi(\xi_i),$$

therefore $J_-^\pi(\xi)$ is a concave function of ξ. It follows that $J_-^*(\xi) = \inf_{\pi\in\Pi} J_-^\pi(\xi)$ as a pointwise infimum of concave functions, is concave.

When $\mathcal{S}$ is discrete, by the preceding discussions, for any $\pi \in \Pi$, $\{\frac{1}{k}J_k^\pi(\cdot) \mid k \geq 1\}$ is a family of uniformly bounded Lipschitz continuous functions with the same Lipschitz constant L. Hence it follows that $J_-^\pi(\cdot)$ and $J_+^\pi(\cdot)$ are Lipschitz continuous with the same Lipschitz constant L.[4] Since by definition $J_-^*(\xi) = \inf_{\pi\in\Pi} J_-^\pi(\xi)$ and $J_+^*(\xi) = \inf_{\pi\in\Pi} J_+^\pi(\xi)$, and the inf operation preserves Lipschitz continuity, it follows that J_-^* and J_+^* are Lipschitzian with the same constant. □

Remark 2.2. We do not know if the optimal limsup function is necessarily concave. Although one can show similarly that $J_+^\pi(\xi)$ is a convex function of ξ by the inequality

$$\limsup_{n\to\infty} (a_n + b_n) \leq \limsup_{n\to\infty} a_n + \limsup_{n\to\infty} b_n,$$

$J_+^*(\xi) = \inf_{\pi\in\Pi} J_+^\pi(\xi)$ as the pointwise infimum of convex functions, may be neither convex, nor concave.

2.3 Optimality and Measurability

The policies we have considered so far are structureless. Optimal or ϵ-optimal policies for one initial distribution ξ are not related to those for another ξ'. We now introduce the reduction of a POMDP to an MDP on $\mathcal{P}(\mathcal{S})$ with equal expected cost. The importance of this well-known reduction is that it allows us to transport from general space MDPs to POMDPs certain results such as optimality equations, structures of optimal policies, and

[4] Let X be a subset of $\mathcal{R}^n$ and $\{f_i \mid f_i : X \to \mathcal{R}, i \geq 1\}$ be a series of uniformly bounded equi-Lipschitzian functions with Lipschitz constant C. The pointwise liminf function $f(x) = \liminf_{i\to\infty} f_i(x)$ is Lipschitz continuous with the same Lipschitz constant: for any i and $x_1, x_2 \in X$,

$$f_i(x_1) \leq f_i(x_2) + Cd(x_1, x_2) \quad \Rightarrow \quad f(x_1) \leq f(x_2) + Cd(x_1, x_2);$$

similarly, one can show $f(x_2) \leq f(x_1) + Cd(x_1, x_2)$; and therefore $|f(x_1) - f(x_2)| \leq Cd(x_1, x_2)$. Similarly, the pointwise limsup function $f(x) = \limsup_{i\to\infty} f_i(x)$ is Lipschitz continuous with the same Lipschitz constant.

measurability conditions. These will be outlined in what follows. For the details, one can see e.g., the books by Bertsekas and Shreve [BS78], and Dynkin and Yushkevich [DY79].

2.3.1 The Equivalent Belief MDP

Consider the induced stochastic process $\{S_0, U_0, (S_t, Y_t, U_t)_{t\geq 1}\}$ and the probability measure $\mathbb{P}^{\xi,\pi}$. Define $\mathcal{P}(\mathcal{S})$-valued random variables $\{\xi_t\}$ with $\xi_0 = \xi$ and

$$\xi_t(\omega)(\cdot) = \mathbb{P}^{\xi,\pi}\left(S_t \in \cdot \mid U_0, (Y_k, U_k)_{k<t}, Y_t\right)(\omega),$$

i.e., ξ_t is a version of the conditional distribution of the state S_t given the observed history $(U_0, (Y_k, U_k)_{k<t}, Y_t)$ prior to U_t being applied. So the random variable ξ_t is a function of the initial distribution ξ and the observed history up to time t. We refer to ξ_t as *beliefs*. By taking iterative conditional expectations, the expected n-stage cost (similarly the expected discounted cost) can be expressed as

$$\begin{aligned} E^{\mathbb{P}^{\xi,\pi}}\left\{\sum_{t=0}^{n-1} g(S_t, U_t)\right\} &= \sum_{t=0}^{n-1} E^{\mathbb{P}^{\xi,\pi}}\left\{E^{\mathbb{P}^{\xi,\pi}}\{g(S_t, U_t) \mid U_0, (Y_k, U_k)_{k<t}, Y_t\}\right\} \\ &= \sum_{t=0}^{n-1} E^{\mathbb{P}^{\xi,\pi}}\left\{\bar{g}(\xi_t, U_t)\right\}, \end{aligned} \tag{2.6}$$

where the function $\bar{g}$ in the second equality is defined as $\bar{g} : \mathcal{P}(\mathcal{S}) \times \mathcal{U} \to \mathcal{R}$ and

$$\bar{g}(\hat{\xi}, u) = \int g(s, u)\hat{\xi}(ds), \quad \forall \hat{\xi} \in \mathcal{P}(\mathcal{S}),\ u \in \mathcal{U}, \tag{2.7}$$

and the second equality of Eq. (2.6) follows from the conditional independence of S_t and U_t given the history $(U_0, (Y_k, U_k)_{k<t}, Y_t)$. Equation (2.6) implies that for all n, the expected n-stage cost of π with respect to the per-stage cost function $g(s, u)$ and the state-control process $\{(S_t, U_t)\}$ can be equivalently viewed as the expected n-stage cost of π with respect to a different per-stage cost function $\bar{g}(\xi, u)$ and the belief-control process $\{(\xi_t, U_t)\}$, (although for each sample path of the POMDP the costs with respect to the two per-stage cost functions are different.)

The belief process $\{\xi_t\}$ is "observable," since the beliefs are functions of the histories and the initial distribution. Furthermore, the beliefs also evolve in a Markov way due to the Markovian property in a POMDP. In particular, for every $u \in \mathcal{U}$, denote by $P_0^{\xi,u}$ the marginal distribution of (S_0, S_1, Y_1) when the initial distribution is ξ and initial control u, i.e.,

$$P_0^{\xi,u}((S_0, S_1, Y_1) \in A) = \int \cdots \int \mathbf{1}_A(s_0, s_1, y_1) P_Y((s_1, u), dy_1) P_S((s_0, u), ds_1)\xi(ds_0)$$

for all Borel measurable sets A of the space of (S_0, S_1, Y_1). Correspondingly, let $P_0^{\xi,u}(S_1 \in \cdot \mid Y_1)$ be a version of the conditional distribution of S_1 given the observation Y_1, and let $P_y^{\xi,u}$ be the marginal distribution of Y_1. Then given $\xi_{t-1} = \xi$ and $(U_{t-1}, Y_t) = (u, y)$, the succeeding belief ξ_t is equal to $\phi_u(\xi, y) \in \mathcal{P}(\mathcal{S})$, where the function $\phi : \mathcal{U} \times \mathcal{P}(\mathcal{S}) \times \mathcal{Y} \to \mathcal{P}(\mathcal{S})$ is defined by

$$\phi_u(\xi, y)(A) = P_0^{\xi,u}(S_1 \in A \mid Y_1)\big|_{Y_1 = y}, \quad \forall A \in \mathcal{B}(\mathcal{S}). \tag{2.8}$$

(Note that $\phi_u(\xi, y)$ is determined by the model of a POMDP, not by any particular policy π.) So ξ_t can be viewed as a function of $(\xi_{t-1}, U_{t-1}, Y_t)$, instead of ξ_0 and the entire observed history. We will refer to the function ϕ as *the function for the next belief* in the POMDP.

Furthermore, when our objective is to minimize a certain expected cost, it is well-known that the belief ξ_t is a *sufficient statistic for control*, in the following sense. For any policy $\pi \in \Pi$ and any initial distribution $\xi \in \mathcal{P}(\mathcal{S})$, there exists a policy $\hat{\pi}_\xi$ that functionally depends only on the sufficient statistic, instead of the entire history, and induces a stochastic process with the same expected cost as the one induced by π. (The proof for the case of discrete spaces can be found in [Åst65], and for the case of general spaces in [BS78, DY79].) This together with the preceding discussions implies that for any expected cost criterion (e.g., discounted or undiscounted, finite or infinite horizon), controlling a POMDP can be equivalently viewed as controlling the $\mathcal{P}(\mathcal{S})$-valued process of beliefs. Furthermore, it is sufficient to consider such policies $\hat{\pi}_\xi$, under the control of which, the process $\{(\xi_t, U_t)\}$ can be equivalently viewed as a completely observable MDP on the state space $\mathcal{P}(\mathcal{S})$.

This equivalent MDP is called *belief MDP*. We can view it as a process "embedded" in the POMDP. When viewed separately by itself, its model can be described as follows. The equivalent belief MDP has

- state space $\mathcal{P}(\mathcal{S})$ and control space $\mathcal{U}$,
- per-stage cost $\bar{g}(\xi, u)$ as defined by Eq. (2.7), and
- state transition probability, denoted by P_ξ and defined by

  $$P_\xi\left((\xi, u), A\right) = \int \mathbf{1}_A\left(\phi_u(\xi, y_1)\right) P_y^{\xi,u}(dy_1), \quad \forall A \in \mathcal{B}(\mathcal{P}(\mathcal{S})),$$

  where $P_y^{\xi,u}$ is the marginal distribution of Y_1 in the POMDP with initial distribution ξ and control u.

We can now define policies with respect to the equivalent belief MDP. Since beliefs are functions of the observed history, these policies are also admissible policies of the original POMDP. In particular, a history dependent randomized policy $\hat{\pi}$ of the belief MDP is a collection of transition probabilities: $\hat{\pi} = (\mu_t)_{t \geq 0}$, where μ_t is a transition probability from $\mathcal{H}'_t$ to $\mathcal{U}$, with

$$\mathcal{H}'_0 = \mathcal{P}(\mathcal{S})$$

being the space of the initial distribution, and

$$\mathcal{H}'_t = \mathcal{H}'_{t-1} \times \mathcal{U} \times \mathcal{P}(\mathcal{S})$$

being the space of the observed histories of beliefs and controls $(\xi_0, U_0, \ldots, U_{t-1}, \xi_t)$ up to time t and prior to U_t being applied. *Deterministic policies* are defined as policies with the probability measure $\mu_t(h'_t, \cdot)$ assigning mass one to one single control for all t. As known from the result of sufficient statistic for control in a POMDP, for a given initial distribution, it is sufficient to consider only the *Markov policies*, that is, policies $\hat{\pi} = (\mu_t)_{t \geq 0}$ with μ_t being a transition probability from $\mathcal{P}(\mathcal{S})$ to $\mathcal{U}$ and mapping the belief ξ_t to a randomized control law. A Markov policy is called *stationary* if μ_t are the same for all t.

As will be addressed in the next section, for discounted and finite-stage POMDP problems, it is sufficient to consider only the deterministic and stationary policies with respect to the belief MDP. (The average cost problems are more complex.)

2.3.2 Optimality Equations

The equivalent belief MDP formulation allows one to pass results from general space MDPs to POMDPs. Under measurability conditions to be specified later, the following results are known for the discounted infinite-horizon and the undiscounted finite-stage problems. (The average cost problem is not as well-understood and will be addressed separately in Section 2.4.)

1. The optimal β-discounted infinite-horizon cost function J^*_β and the optimal k-stage cost function J^*_k are measurable and bounded, and satisfy the following optimality equations, also called *dynamic programming (DP) equations*, or *Bellman equations*:

$$J^*_\beta(\xi) = \inf_{u\in\mathcal{U}} \left[\bar{g}(\xi,u) + \beta\, E^{P_0^{\xi,u}}\{J^*_\beta\left(\phi_u(\xi,Y_1)\right)\}\right], \tag{2.9}$$

$$J^*_0 = 0, \qquad J^*_k(\xi) = \inf_{u\in\mathcal{U}} \left[\bar{g}(\xi,u) + E^{P_0^{\xi,u}}\{J^*_{k-1}\left(\phi_u(\xi,Y_1)\right)\}\right], \quad k \geq 1. \tag{2.10}$$

These optimality equations can also be written as

$$J^*_\beta(\xi) = \left(\mathcal{T}J^*_\beta\right)(\xi), \quad \beta\in[0,1), \qquad J^*_k(\xi) = \left(\mathcal{T}J^*_{k-1}\right)(\xi), \quad \beta = 1,$$

where $\mathcal{T}$ is the mapping associated with the right-hand sides of Eq. (2.9) and (2.10), and defined by

$$\left(\mathcal{T}J\right)(\xi) = \inf_{u\in\mathcal{U}} \left[\bar{g}(\xi,u) + \beta\, E^{P_0^{\xi,u}}\{J\left(\phi_u(\xi,Y_1)\right)\}\right], \qquad \beta\in[0,1],$$

for any bounded measurable function J. We call $\mathcal{T}$ *the DP mapping* of the POMDP.

2. In the discounted infinite-horizon case, if for all ξ the minima of the optimality equation are attainable, then there exists an optimal policy that is deterministic and stationary (with respect to the belief MDP); otherwise, for every $\epsilon > 0$, there exist ϵ-optimal policies that are deterministic and stationary.

3. In the finite-stage case, if for all ξ the minima on the right-hand sides of the optimality equations are attainable, then there exist optimal policies that are deterministic (with respect to the belief MDP); otherwise, for every $\epsilon > 0$, there exist ϵ-optimal policies that are deterministic.

2.3.3 Measurability

We now describe the issues of measurability of cost functions and policies within our context. As these issues are quite complicated, we will not go into them in detail. Instead, we only briefly mention two approaches from the literature – the universal measurability and the semi-continuity approaches, (for references, see e.g., Bertsekas and Shreve [BS78], and Dynkin and Yushkevich [DY79]), and discuss what they mean in the context of a POMDP.

First note that we do not require the optimal cost functions to be measurable in the definition given earlier. These functions are well defined as pointwise infima over policies for each ξ. Nevertheless, this is not sufficient for the recursive optimality equations (2.9) and (2.10) to hold, which involve expectation of the optimal cost functions, and therefore require these functions to be measurable. The conditions so far given for general space

POMDPs do not guarantee that the optimal cost functions are Borel measurable.[5] Even when they are measurable, there are no guarantees for an optimal or ϵ-optimal policy that is measurable and stationary (with respect to the belief MDP).

To overcome these problems, the methodology of both the semi-continuity and universal measurability approaches is to consider a set of measurable functions such that, with certain conditions on the transition probabilities, the set is closed under the DP mapping. Furthermore, for every member function of the set, there always exists a measurable selection of the exact or approximate minimum (as a function of the state) in the DP equation. Thus, central to the methodology are the closedness of the set of measurable functions and the selection theorem.

First we describe a context where the semi-continuity approach applies. We assume (i) $\mathcal{U}$ is compact; (ii) $g(s, u)$ is continuous; (iii) $P_S((s, u), A)$ and $P_Y((s, u), A)$ are continuous transition probabilities; (iv) the next belief $\phi_u(\xi, y)$ as a function of (ξ, u, y) is continuous when restricted on a set satisfying certain conditions.[6] Under these assumptions, the results of Section 2.3.2 hold with Borel measurable policies and the optimal cost functions being continuous. The compactness assumption on $\mathcal{U}$ ensures that there exists an optimal policy that is deterministic. However, as a comment on the assumptions, we consider condition (iv) too restrictive for the POMDP problem, because it is not implied by (iii), which itself can already be a very strong condition for some problems. (We note, however, that when the state, observation and control spaces are discrete, all the continuity assumptions are satisfied.)

In the context of a POMDP, the universal measurability approach is as follows. We assume that (i) $g(s, u)$ is lower semi-analytic,[7] (ii) Π is defined to be the set of all history dependent universally measurable policies.[8] Under these assumptions, the results of Section 2.3.2 hold with universally measurable policies and the optimal cost functions being universally measuable, or more precisely lower semi-analytic. One thing to note in this case is that, in the absence of assumptions guaranteeing that J_β^* or J_k^* is lower semi-continuous, often one cannot ensure the existence of an optimal policy that is deterministic, even if $\mathcal{U}$ is compact.

Under these conditions there are no measurability issues for either the POMDP or its equivalent belief MDP. In this thesis whenever we talk about general space POMDPs, **we will assume that the measurability conditions are satisfied**, either Borel measurable or universally measurable, so that the results listed above hold. For the sake of notational simplicity, we will keep using Borel σ-algebra, and it should be understood that what we will state about general space POMDPs hold when the Borel σ-algebra and a measure on it is replaced by the universal σ-algebra and the completion of the measure. On the other hand for countable space POMDPs, the measurability issues are of no concern and can be ignored.

[5] Generally, the function $F(x) = \inf_y f(x, y)$ is not Borel measurable even if the function $f(x, y)$ is.

[6] Let this set be D and it is such that for any fixed (ξ, u), the set $\{y \mid (\xi, u, y) \in D^c\}$ has $P_y^{\xi,u}$-measure zero, where $P_y^{\xi,u}$ is the marginal distribution of Y_1 when the initial distribution is ξ and initial control u, and D^c denotes the complement of D.

[7] A set is called analytic, if it is the image of a Borel set under a Borel measurable function; a real-valued function f is called lower semi-analytic, if its lower level sets $\{x \mid f(x) \leq \mu\}$ are analytic.

[8] A function on X is called universally measurable, if it is measurable with respect to the universal σ-algebra of X – defined such that for any measure on $\mathcal{B}(X)$, a universally measurable function is measurable for the completion of that measure, hence the name "universal"; a policy is called universally measurable, if for every Borel set A of $\mathcal{U}$, $\mu_t(h, A)$ as a function of h is universally measurable.

2.4 The Average Cost Problem for Finite Space Models

2.4.1 Difficulties

Even for a POMDP with finite spaces, the average cost problem is still not well understood, in contrast to the simplicity of the average cost problem in a finite space MDP. This discrepancy can be intuitively explained as follows. While the asymptotic behavior of the state-control sequence under a history dependent randomized policy is in general very hard to analyze in a finite space MDP, fortunately, for an MDP there always exists an optimal policy that is stationary and deterministic, so it is sufficient to consider only stationary and deterministic policies under which the state process is simply a Markov chain. Admissible policies in a POMDP, however, are in general history dependent randomized policies of its associated completely observable MDP. Thus in a POMDP, even for stationary and deterministic policies, their asymptotic behaviors are not easy to characterize. This causes difficulties in analyzing the average cost POMDP problem.

A central question of the average cost POMDP is the existence of an optimal policy that is stationary and deterministic (with respect to the belief MDP) under the average cost criterion. By the theory of general MDPs, we have the following sufficient conditions. If either the two nested equations – referred as optimality equations,

$$J(\xi) = \min_{u \in \mathcal{U}} \; E^{P_0^{\xi,u}} \left\{ J\left(\phi_u(\xi, Y_1)\right) \right\}, \qquad U(\xi) \overset{def}{=} \arg\min_{u \in \mathcal{U}} \; E^{P_0^{\xi,u}} \left\{ J\left(\phi_u(\xi, Y_1)\right) \right\},$$
$$J(\xi) + h(\xi) = \min_{u \in U(\xi)} \; \left[\bar{g}(\xi, u) + E^{P_0^{\xi,u}} \left\{ h\left(\phi_u(\xi, Y_1)\right) \right\} \right], \tag{2.11}$$

or the two equations – referred as modified optimality equations,

$$J(\xi) = \min_{u \in \mathcal{U}} \; E^{P_0^{\xi,u}} \left\{ J\left(\phi_u(\xi, Y_1)\right) \right\},$$
$$J(\xi) + h(\xi) = \min_{u \in \mathcal{U}} \; \left[\bar{g}(\xi, u) + E^{P_0^{\xi,u}} \left\{ h\left(\phi_u(\xi, Y_1)\right) \right\} \right], \tag{2.12}$$

have a bounded solution $(J^*(\cdot), h^*(\cdot))$, then the optimal liminf and limsup average cost functions are equal and $J_-^*(\cdot) = J_+^*(\cdot) = J^*(\cdot)$,[9] and any stationary and deterministic policy that attains the minima of the right hand sides simultaneously is an average cost optimal policy.

If the optimal average cost function is constant, then the first equation of the optimality equations (2.11) or (2.12) is automatically satisfied, leaving us the second equation, which we will refer as the constant average cost DP equation:

$$\lambda + h(\xi) = \min_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + E^{P_0^{\xi,u}} \left\{ h\left(\phi_u(\xi, Y_1)\right) \right\} \right]. \tag{2.13}$$

If a bounded solution (λ^*, h^*) to Eq. (2.13) exists, then both the optimal liminf and limsup cost functions are equal to λ^*, and any stationary deterministic policy that attains the minima of the right hand side is an average cost optimal policy.

For discrete space MDPs, only under certain recurrence conditions is the optimal average cost guaranteed to be constant. Equations (2.11) and (2.12) are the optimality equations

[9] To see this, note first it is true for countable space MDPs (Theorem 9.1.3 and Theorem 9.1.2 (c) of [Put94]) and then note that for every initial belief a finite space POMDP can be viewed as a countable space MDP.

for the multichain case, and Eq. (2.13) is the optimality equation for the unichain case.[10]

A finite space POMDP is peculiar in that under any policy only countable beliefs are reachable from a given initial belief, so that naturally the belief space would be decomposable and the recurrence conditions would be in general violated. Nevertheless, even when this happens, in some of the cases the optimal average cost can still be proved to be constant, (see e.g., sufficient conditions for a constant optimal average cost in Hsu, Chuang and Arapostathis [HCA05]). Thus although these optimality equations can be taken as the starting point for the average cost POMDP problem, due to the special structure of hidden states of a POMDP, the POMDP theory is not subsumed by the general MDP theory.

The theoretical studies on the average cost POMDP problem so far have been centered on the existence of solution to the constant average cost DP equation (2.13), for which necessary and sufficient conditions based on the vanishing discount argument are given in [Ros68, Pla80, FGAM91, HCA05]. Ross's result on general MDPs [Ros68] showed that a sufficient condition in the context of a finite space POMDP is the equicontinuity of the discounted cost functions. Platzman [Pla80] essentially proved that a necessary and sufficient condition for a bounded solution to Eq. (2.13) in a finite space POMDP is the uniform boundedness of the optimal discounted relative cost functions $|J^*_\beta(\cdot) - J^*_\beta(\bar{\xi})|, \beta \in [0, 1)$ with $\bar{\xi}$ being an arbitrary fixed reference point; and Hsu, Chuang and Arapostathis in their recent work [HCA05] established the same necessary and sufficient condition in a POMDP with discrete state and observation spaces, and a compact control space. Fernández-Gaucherand, Arapostathis and Marcus [FGAM91] established sufficient conditions under which Eq. (2.13) has possibly unbounded solutions, using the countability property of the set of reachable beliefs in a discrete space POMDP.

Sufficient conditions relating to the transition structure of a POMDP that can be easier to verify for given problems are also proposed in [Ros68, Pla80, RS94, FGAM91, HCA05]. In particular, there are Ross's renewability condition [Ros68], Platzman's reachability and detectability condition [Pla80], Runggaldier and Stettner's positivity condition [RS94], and Hsu, Chuang and Arapostathis's interior and relative interior accessibility conditions [HCA05]. Hsu, Chuang and Arapostathis's conditions are the weakest among all. These conditions apply to certain classes of problems.

In summary, unlike for MDPs, we still do not fully understand the constant average cost POMDP problem. Details of recent progress are summarized in the survey paper on average cost MDPs [ABFG$^+$93] by Arapostathis et al. and the recent work on average cost POMDPs [HCA05] by Hsu, Chuang and Arapostathis. It is still unclear whether most practical problems will have a constant optimal average cost; and not much is understood for the average cost problem without the assumption of a constant optimal cost.

2.4.2 Examples of POMDP with Non-Constant Optimal Average Cost

It is natural to expect a non-constant optimal average cost of a POMDP, when its associated completely observable MDP is multichain, or is deterministic and periodic. We give examples to show that the optimal average cost is not necessarily constant, even when the completely observable MDP is recurrent and aperiodic (which means that the Markov chain

[10] Following [Put94], we classify an MDP as a unichain MDP, if under any stationary and deterministic policy the induced Markov chain has one single recurrent class with a possibly non-empty set of transient states, and we classify an MDP as a multichain MDP, if there exists a stationary and deterministic policy under which the induced Markov chain has multiple recurrent classes.

induced by any stationary deterministic policy of the MDP is recurrent and aperiodic). After the examples, we will discuss the construction and implications.

Example 2.1. First we describe the model of a small MDP that will be used subsequently to make a bigger MDP that we want. The MDP has 4 states $\{1, 2, 3, 4\}$ and 2 actions $\{a, b\}$. Fig. 2-2 shows the transition structures. We use the symbol '$-$' to mean either action a or b. For states except state 2, under either actions the transition probabilities are the same. In particular, $p(2 \mid 1, -) = p(3 \mid 1, -) = 1/2; p(4 \mid 3, -) = 1; p(1 \mid 4, -) = 1$. For state 2, the two actions differ: $p(2 \mid 2, a) = p(4 \mid 2, a) = 1/2; p(1 \mid 2, b) = 1$. For this MDP, there are essentially only two different deterministic and stationary policies, namely, taking action a or b at state 2. Both of them induce a recurrent and aperiodic Markov chain, as can be easily seen.

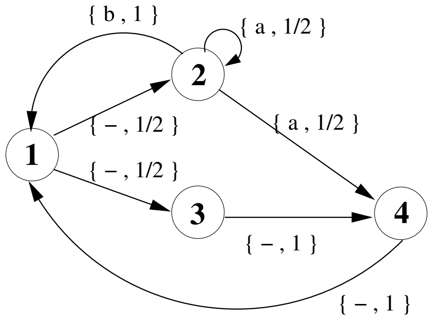

Figure 2-2: The symbolic representation of a MDP with 4 states $\{1, 2, 3, 4\}$ and 2 actions $\{a, b\}$. Possible state transitions are indicated by directed edges with {action, probability} values, and the symbol '-' stands for either action a or b.

In this MDP if we apply the non-stationary policy that repeats actions in this sequence:

$$a, a, b, a, a, b, \ldots$$

then starting from state 1, the set of states we possibly visit form a "cycle":

$$\{1\} \rightarrow \{2, 3\} \rightarrow \{2, 4\} \rightarrow \{1\}.$$

We choose the per-stage cost function so that the cost of this cycle is zero. In particular, we define $g(1, a) = 0, g(1, b) = 1; g(2, -) = 0; g(3, a) = 0, g(3, b) = 1; g(4, a) = 1, g(4, b) = 0$. Then the policy has zero cost from state 1. Since this policy does not depend on state information at all, for any partially observable problem associated with this MDP, the optimal average cost starting at state 1 is, with $\bar{\xi}(1) = 1$,

$$J^*(\bar{\xi}) = 0.$$

Now we combine two models of this MDP to have the MDP with 8 states as shown in Fig. 2-3. We then double the action set to define actions $\{a', b'\}$, under each of which the transition probabilities are the same as under a or b, respectively. This MDP of 8 states and 4 actions is recurrent and aperiodic (for every stationary and deterministic policy) by construction.

We define the per-stage costs as follows. On states 1-4 the per-stage cost of applying action a or b is as defined in the 4-state MDP, and the per-stage cost of applying a' or b'

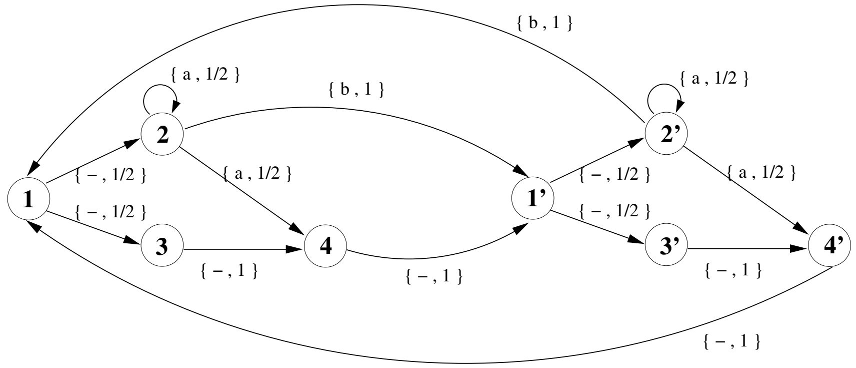

Figure 2-3: The symbolic representation of a MDP with 8 states, constructed from stitching up two 4-state MDPs shown in Fig. 2-2.

is 1; while on states $1'$-$4'$ the per-stage cost of applying a' or b' is as defined in the 4-state MDP, and the per-stage cost of applying a or b is 1.

Starting from $\bar{\xi}$ with $\bar{\xi}(1) = 1$, the optimal average cost is 0 for any partially observable problem associated with this MDP, and one of the optimal policies is the one that repeats controls in this sequence:

$$a, a, b, a', a', b', a, a, b, a', a', b', \ldots$$

Consider a partially observable problem in which no observations can distinguish states i from i' for $i = 1, \ldots, 4$. Consider the initial distribution ξ with $\xi(1) = \xi(1') = 1/2$. Then under any policy, because of the symmetry of the system, we always have the same probability of being in state i as in state i'. Hence the expected cost of each step is at least 1/2, and consequently the optimal average cost is at least 1/2. Hence $J^*_-(\xi) > J^*(\bar{\xi})$. The optimal average cost function cannot be constant. □

Remark 2.3. In the example above, one can see that we have still exploited the periodic structure of the chain, even though the completely observable MDP is aperiodic under stationary policies. As to the role of the periodic structure, we are inclined to believe that it is not fundamental to the problem, but a mere constructional convenience. Key to the construction in the example is the non-identifiability of the two subsets of the state space. The next example to some degree confirms this observation.

Remark 2.4. The 4-state POMDP in the first part of the example indeed has a constant optimal average cost which equals zero. This can be verified by showing that there is a policy which asymptotically drives the state process to follow the cycle.

Example 2.2. The MDP has 4 states $\{1, 2, 3, 4\}$ and 2 actions $\{a, b\}$, and its transition structures are as shown in Fig. 2-4, where we use the symbol '$-$' to represent any action. Under any policy the state process is a Markov chain and the transition probabilities are as follows: $p(1|1, -) = 1/2, p(2|1, -) = 1/2; p(3|2, -) = 1; p(3|3, -) = 1/2, p(4|3, -) = 1/2; p(1|4, -) = 1$. Clearly the Markov chain is recurrent and aperiodic.

Now we define the observations such that the states $\{1, 3\}$ are indistinguishable and $\{2, 4\}$ are indistinguishable. In particular, let the observation space be $\{c, d\}$ and let $p(c|1, -) = p(c|3, -) = 1; p(d|2, -) = p(d|4, -) = 1$. Thus if we know the initial state,

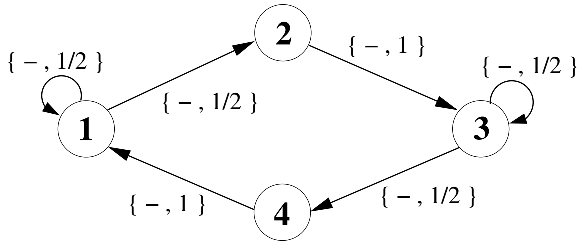

Figure 2-4: The symbolic representation of a MDP with 4 states and 2 actions. Possible state transitions are indicated by directed edges with {action, probability} values, and the symbol '-' stands for either action a or b.

then the state process can be inferred from the observations as if it were completely observable.

We then define the per-stage costs as $g(1, a) = g(3, b) = 1$ and all the other per-stage costs to be zero. It follows that if we start from an initial distribution $\xi(1) = 1$ or an initial distribution $\xi(3) = 1$, then the optimal average cost is zero, while if we start from an initial distribution ξ with $\xi(1) = \xi(3) = 1/2$, say, then the optimal average cost is strictly greater than zero. □

Remark 2.5. This example is also a counter example to a result of [Bor00] on the existence of a bounded solution to the constant average cost DP equation. Consider the state processes of two POMDPs governed by a common control process. It is stated in [Bor00] that if the expected coupling time of the two state processes is bounded by a constant K_0 under any control process, then the optimal average cost must be constant. The POMDP in this example satisfies the finite expected coupling time condition of [Bor00], since the Markov chain is uncontrolled and recurrent, yet the optimal average cost function is non-constant.

Remark 2.6. Of course our examples violate the sufficient conditions given in the literature for a constant optimal average cost. Our examples also violate a necessary condition that we are going to show next.

2.4.3 A Peculiar Necessary Condition

We now show for the finite state space model a necessary condition for J_-^* to be constant, which sounds somewhat peculiar.

Recall the set Π of history dependent randomized policies, defined in Section 2.1. Each $\pi \in \Pi$ is a collection of conditional control probabilities $\{\mu_t\}_{t \geq 0}$ where μ_t maps the history h_t consisting of past observations and controls up to time t, to a measure on the control space. (The initial history is by definition empty, so that μ_0 is a measure on the control space independent of the initial state distribution of the POMDP.)

Recall also that for any history dependent randomized policy $\pi \in \Pi$, $J_-^\pi(\cdot)$ is a concave function (see the proof of Prop. 2.2), and is bounded below by $J_-^*(\cdot)$, the pointwise infimum of all such functions $\{J_-^\pi(\cdot) | \pi \in \Pi\}$. Thus, if $J_-^*(\cdot)$ is constant, then there must be a function $J_-^\pi(\cdot)$ that is nearly "flat", implying that π is near-liminf optimal. This observation is stated and proved formally in the following proposition. Abusing notation, we use J_-^* to denote the constant value of the function $J_-^*(\cdot)$, when the latter is constant.

Proposition 2.3. *Assume a finite state space $\mathcal{S}$. If the optimal liminf average cost function J_-^* is constant, then for any $\epsilon > 0$, there exists a history dependent randomized policy π that does not functionally depend on the initial distribution ξ, such that π is ϵ-liminf optimal, i.e.,*

$$J_-^\pi(\xi) \leq J_-^* + \epsilon, \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

The proof starts with a lemma, which uses some basic notions of convex analysis.[11]

Lemma 2.1. *Let X be a compact and convex subset of $\mathcal{R}^n$, and D be the relative boundary of X. Let $f : X \to \mathcal{R}, f(x) \geq 0$ be a non-negative and concave function. Then for any relative interior point $\hat{x} \in ri(X) = X \setminus D$ and any $x \in X$,*

$$f(x) \leq C_{\hat{x}} f(\hat{x}), \qquad \text{where} \quad C_{\hat{x}} = \left(\frac{\max_{x \in X} \|x - \hat{x}\|}{\min_{x \in D} \|x - \hat{x}\|} + 1 \right).$$

Proof: For any $\hat{x} \in ri(X)$ and $x \in X$, let $r(x)$ be the intersection point of the relative boundary D and the ray that starts at x and passes through $\hat{x}$. By the concavity and non-negativity of the function f, we have

$$\begin{aligned} f(\hat{x}) &\geq \frac{\|\hat{x} - r(x)\|}{\|x - \hat{x}\| + \|\hat{x} - r(x)\|} f(x) + \frac{\|x - \hat{x}\|}{\|x - \hat{x}\| + \|\hat{x} - r(x)\|} f(r(x)) \\ &\geq \frac{\|\hat{x} - r(x)\|}{\|x - \hat{x}\| + \|\hat{x} - r(x)\|} f(x). \end{aligned}$$

Hence

$$f(x) \leq \frac{\|x - \hat{x}\| + \|\hat{x} - r(x)\|}{\|\hat{x} - r(x)\|} f(\hat{x}) \leq \left(\frac{\max_{x \in X} \|x - \hat{x}\|}{\min_{x \in D} \|x - \hat{x}\|} + 1 \right) f(\hat{x}),$$

and the claim follows. □

Proof of Prop. 2.3: For all $\pi \in \Pi$ (as defined in Section 2.1), $J_-^\pi(\cdot)$ is a concave function (Prop. 2.2) and $J_-^\pi(\cdot) \geq J_-^*$.

Since the state space $\mathcal{S}$ is finite, $\mathcal{P}(\mathcal{S})$ is a compact and convex set in $\mathcal{R}^n$ with $n = |\mathcal{S}|$. Let D be the relative boundary of $\mathcal{P}(\mathcal{S})$. Pick an arbitrary $\hat{\xi}$ in the relative interior of $\mathcal{P}(\mathcal{S})$, and let $C_{\hat{\xi}}$ be defined as in Lemma 2.1.

For any $\epsilon > 0$, let π be a policy such that $J_-^\pi(\hat{\xi}) \leq J_-^* + \frac{\epsilon}{C_{\hat{\xi}}}$. Applying Lemma 2.1 to the concave and non-negative function $J_-^\pi(\xi) - J_-^*$, we have,

$$J_-^\pi(\xi) - J_-^* \leq C_{\hat{\xi}} \left(J_-^\pi(\hat{\xi}) - J_-^* \right) \leq C_{\hat{\xi}} \frac{\epsilon}{C_{\hat{\xi}}} \leq \epsilon, \qquad \forall \xi \in \mathcal{P}(\mathcal{S}),$$

i.e., $J_-^\pi(\xi) \leq J_-^* + \epsilon$ and π is ϵ-liminf optimal. □

Remark 2.7. Note that the policy π in Prop. 2.3 is proved to be ϵ-liminf optimal, not ϵ-limsup optimal. By contrast, there is a policy stationary with respect to the conditional

[11] Let X be a subset of $\mathcal{R}^n$. The *affine hull* aff(X) of X is defined to be $S + \bar{x}$ where $\bar{x}$ is an arbitrary point in X and S is the subspace spanned by $X - \bar{x}$. A point $x \in X$ is called a *relative interior point* if there exists an open neighborhood $\mathcal{N}(x)$ of x such that $\mathcal{N}(x) \cap \text{aff}(X) \subset X$. The set of relative interior points of X is called the *relative interior* of X, and denoted by $ri(X)$. A convex set always has a non-empty relative interior. Let $cl(X)$ be the closure of X. The set $cl(X) \setminus ri(X)$ is called the *relative boundary* of X.

distributions of the states that is both liminf and limsup optimal, when the optimality equation (2.13) admits a bounded solution.

Remark 2.8. For a POMDP with finite state, observation and control spaces, Prop. 2.3 indeed leads to a stronger claim on the optimal average cost functions, as well as a proof of the near-optimality of the class of policies called finite-state controllers, to be discussed later. These results are reported in Appendix D.

Finally, we mention an example which is due to Platzman. The POMDP has a constant optimal average cost, but the constant average cost DP equation does not have a bounded solution, and there exists an optimal policy that is deterministic and non-stationary (with respect to the belief MDP). We show that for Platzman's example the policy in Prop. 2.3 can be easily constructed.

Example 2.3 (Platzman [Pla80]). The POMDP has 2 states $\{1, 2\}$, 3 observations $\{1, 2, 3\}$, and 3 actions $\{1, 2, 3\}$. Under any action, the state remains the same, i.e.,

$$p(S_1 = i | S_0 = i, -) = 1, \quad i = 1, 2.$$

By applying actions one can gain information or rewards, however the two goals are "mutually exclusive." Under action 1 or 2, the observation 3 is generated which bears no information of the state. Under action 3, the correct state is observed with probability p, i.e.,

$$p(Y = i | S = i, U = 3) = p, \quad i = 1, 2,$$

(without loss of generality, assuming $p > 1/2$.) The per-stage costs are

$$g(i, i) = -1, \quad g(i, 3) = 0, \quad i = 1, 2.$$

So, if the states are guessed correctly, the non-informative actions bring rewards.

By applying action 3 with a diminishing frequency, one can have an average cost of -1. Thus the optimal average cost function is constant and equals -1. The constant average cost DP equation does not have a bounded solution, however. To see this, suppose the contrary that there is a bounded solution to

$$-1 + h(\xi) = \min_{u \in \{1,2,3\}} \left[\bar{g}(\xi, u) + E^{P_0^{\xi,u}} \left\{ h\big(\phi_u(\xi, Y_1)\big) \right\} \right].$$

Then, for ξ with $1 > \xi(1) > 0$ and $1 > \xi(2) > 0$, actions 1 or 2 cannot attain the minimum of the right hand side. Otherwise, since action 1 or 2 does not bring out information about the state, the next belief remains the same, and by the DP equation,

$$-1 + h(\xi) = -\xi(1) + h(\xi), \quad \text{or} \quad -1 + h(\xi) = -\xi(2) + h(\xi),$$

a contradiction. So only action 3 can attain the minimum of the right hand side, and thus action 3 is optimal for all beliefs in the relative interior of the belief space, i.e., the set of ξ with $\xi(i), i = 1, 2$ strictly greater than zero. However, starting from such a belief ξ, the belief always remains in the relative interior, and the policy that applies action 3 all the time, incurs an average cost of zero, which cannot be optimal. The contradiction implies that the constant average cost DP equation does not admit a bounded solution. The preceding discussion also shows that neither does the DP equation admit an unbounded solution in the sense as defined and analyzed by [FGAM91]. □

Indeed for this example, the history dependent policy in Prop. 2.3 can be chosen to be the optimal. Independent of the initial distribution, fix a sequence t_k, which will be the time to apply action 3, such that action 3 is applied infinitely often with diminishing frequency, i.e.,

$$\lim_{k\to\infty} t_k = \infty, \qquad \lim_{t\to\infty} \frac{\max\{k \mid t_k \leq t\}}{t} = 0.$$

At time t, let $n_t(1)$ be the number of times that observation 1 is observed up to time t, and $n_t(2)$ the number of times that observation 2 is observed up to time t. At time $t \neq t_k$, apply action 1 if $n_t(1) \geq n_t(2)$, and apply action 2 otherwise. By the law of large number, it is easy to show that such a policy has average cost -1 and therefore optimal.

2.5 Summary

We have reviewed briefly POMDPs with general space models and the difficulties of the average cost POMDP problems for finite space models. We have also shown a few new results. Among them, we would like to acknowledge that the method of proving the concavity and Lipschitz continuity of optimal cost functions by considering the induced stochastic processes, instead of considering optimality equations, came to us from personal communications with Prof. S. K. Mitter.

Our results on non-constant optimal average cost examples and the necessary condition of Prop. 2.3 for a constant optimal liminf cost raise questions on whether for applied problems the constant average cost DP equation will be usually satisfied. Our necessary condition will further lead to a proof of near-optimality of the class of finite state controllers, which has not been proved before.

Chapter 3

Fictitious Processes, Inequalities for Optimal Cost Functions and Lower Bounds

3.1 Introduction

In this chapter we will establish that due to the special structure of hidden states in a POMDP, one can construct a class of processes (either POMDPs themselves or belief MDPs) that provide lower bounds of the optimal cost functions of the POMDP problem with either the discounted, finite-stage undiscounted, or average cost criteria and with general space models. This will lay the foundation for the subsequent chapters where the lower bound results will be further discussed or extended.

For discounted problems individual lower approximation schemes (eg., [Lov91, LCK95, ZL97, ZH01]) have been proposed as approximations to the optimal cost function. There, the role of the lower approximations has been more computational than analytical, since analytically the discounted problem is well understood.

Our main contribution is for the average cost criterion. The extension of these approximation schemes to the average cost problems and their role as lower bounds of the optimal average cost function have not been proposed and analyzed previously. Since the average cost POMDP problem is still not well understood (see Section 2.4), the role of the lower approximations is thus, in our opinion, as much analytical as computational.

The development of this chapter, consisting of two lines of analysis and summarized in Fig. 3-1, is as follows. We will analyze as a whole the class of processes that have the lower bound property, and characterize them in a way linking to the information of hidden states. To this end, we will first introduce in Section 3.2 processes that we will call fictitious processes. They resemble the original POMDP and differ from it only in the first few stages. Following their definitions and analyses, we give in Section 3.3 and 3.4 the lower bound property in its primitive form – it will be characterized by the inequalities satisfied by the optimal discounted and finite-stage cost functions J_β^* and J_k^* of the original POMDP. In particular, when fictitious processes are defined for all initial distributions, one can define, corresponding to their transition models, a belief MDP, which will be called the modified belief MDP. The inequalities will be of the following form:

$$J_\beta^* \geq \widetilde{\mathfrak{T}} J_\beta^*, \qquad J_k^* \geq \widetilde{\mathfrak{T}} J_{k-1}^*,$$

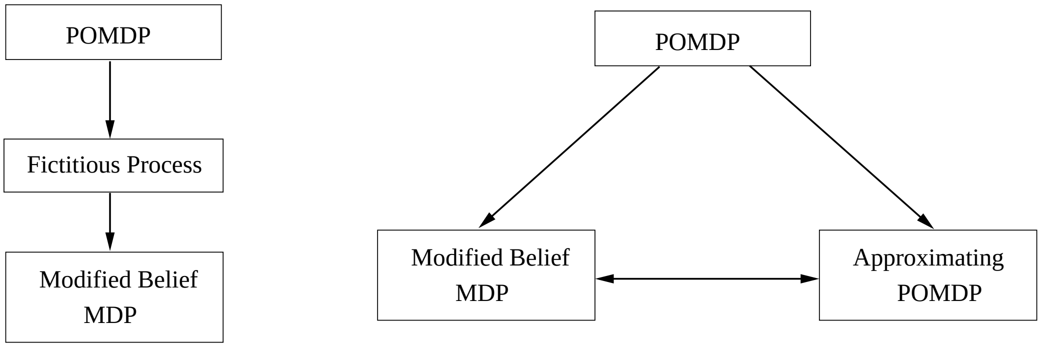

Figure 3-1: A summary of the development of this chapter. The left diagram corresponds to the first line of analysis, in which we first construct fictitious processes resembling the original POMDP by exploiting information of the hidden states, and we then derive the corresponding modified belief MDPs and lower approximation schemes. This line of analysis is more intuitive and constructive. However, it leads to a weaker lower bound result. The right diagram corresponds to the second line of analysis, in which we construct an approximating POMDP that is equivalent to the modified belief MDP and also relates to the original POMDP. This leads to a stronger lower bound result.

with $\widetilde{\mathfrak{T}}$ being the DP mapping of the modified belief MDP. The transition models of the belief MDP will be called a lower approximation scheme.

Since the average cost and the total cost are defined through limits of the finite-stage costs, the finite-stage cost inequalities then lead to lower bounds of the optimal cost functions for the average cost and total cost cases. Thus we obtain in Section 3.4 our first lower bound result (Prop. 3.3) for the average cost case, which will be strengthened later.

We then consider ways of designing fictitious processes that seem natural and can be useful in practice for designing cost approximation schemes. In Section 3.5, we will give examples which include approximation schemes previously proposed for discounted problems, as well as a new lower approximation scheme and various combinations of schemes, followed by a brief discussion on comparisons of schemes in Section 3.6. We will categorize the design methods into two types of approaches that seem to be different conceptually. One type involves an information oracle assumption, in which case information is revealed and the evolution model of the POMDP is however not altered. The other type, which is closely related to discretization-based approximation schemes, involves an alteration of the POMDP model in addition to the revelation of information.

Next, in Section 3.8, following the second line of analysis, we will strengthen our lower bound result of Section 3.3 and 3.4 to obtain the main theorem (Theorem 3.2), which is more powerful. It alleviates the dependence of the lower bound result on the DP mapping of the modified belief MDP, and enables one to claim directly that *the optimal expected cost of the modified belief MDP is a lower bound of the original POMDP under various cost criteria*, including constrained cases. Earlier, this property was known to be true only for those approximation schemes constructed using the information oracle assumption. For the general case it was proved only for finite space POMDPs and modified belief MDPs that have essentially finite spaces, ([YB04] for the average cost case and other works by the author for the total cost and constrained cases).

Our method of proof of Theorem 3.2 involves construction of an approximating POMDP that relates to both the original POMDP and the modified belief MDP. The approximating POMDP is such that while it has an equivalent belief MDP formulation identical to the modified belief MDP, it can be viewed at the same time as the original POMDP plus additional observations (generated in a certain non-stationary way). The approximating POMDP is a proof device here. For certain approximation schemes, it can be highly non-intuitive to interpret them as approximating POMDPs. Nevertheless, this interpretation and the main theorem show that at least mathematically, we need not distinguish between approximation schemes constructed using the information oracle method and those using the non-information oracle method.

In addition to the main line of analysis just outlined, we will also give in this chapter two alternative ways of proving the concavity of the optimal discounted and finite-horizon cost functions, besides the proof in Section 2.2.2. One is a by-product of the construction of the fictitious process, which shows the link of concavity to the structure of hidden states. The other is Åström's backwards induction argument [Åst69] of proving that the DP mapping of the POMDP preserves concavity. Åström's proof was for the case of finite state and observation spaces. For completeness, we will supply a proof for general space models in Section 3.7. Alternatively (even though less constructively as we argue), the inequalities we construct based on fictitious processes, can also be proved using the concave preserving property of the DP mapping. The latter leads to the immediate extension of the inequalities and hence the first lower bound property to the case where the per-stage cost is a concave function of the belief. It is worth to mention that the second, stronger lower bound result also holds for the concave per-stage cost case. Therefore there seems to be a technical limitation in that method of analysis that relies on the DP mapping.

As a last comment before we start the analysis, in this chapter, we will focus more on general analytical properties and be less concerned about computational issues. Chapters 4 and 5, as well as later chapters, will focus on the computation and application issues when specializing the results to discretized cost approximation schemes for finite space models with various optimality criteria.

3.2 Fictitious Processes

Notation

Fictitious processes are to be constructed on a sample space, which is *different* from the sample space of the original POMDP. In addition to states, observations and controls, there will be one more random variable $Q = (Q_1, \ldots, Q_m)$, which carries certain information of the hidden states that the controller can exploit. This will be the key idea in the construction.

Let Q be $\mathcal{M}$-valued, where $\mathcal{M} = \prod_{i=1}^m \mathcal{M}_i$ and $\mathcal{M}_i$ are Borel measurable sets of complete separable metric spaces. The sample space $(\widetilde{\Omega}, \widetilde{\mathcal{F}})$ of a fictitious process is defined as

$$\widetilde{\Omega} = \mathcal{M} \times \mathcal{S} \times \mathcal{U} \times \prod_{t=1}^{\infty} (\mathcal{S} \times \mathcal{Y} \times \mathcal{U}), \qquad \widetilde{\mathcal{F}} = \mathcal{B}\left(\widetilde{\Omega}\right).$$

With a sample $\tilde{\omega} = (q, s_0, u_0, \ldots, s_t, y_t, u_t, \ldots) \in \widetilde{\Omega}$, the random variables of a fictitious process,

$$\{Q, \tilde{S}_0, \tilde{U}_0, (\tilde{S}_t, \tilde{Y}_t, \tilde{U}_t)_{t>0}\},$$

are defined as the projections of $\widetilde{\omega}$ to their respective spaces:

$$Q(\tilde{\omega}) = q, \qquad \tilde{S}_t(\tilde{\omega}) = s_t, \quad t \geq 0, \qquad \tilde{Y}_t(\tilde{\omega}) = y_t, \quad t \geq 1, \qquad \tilde{U}_t(\tilde{\omega}) = u_t, \quad t \geq 0.$$

Define an increasing sequence of σ-algebras $\{\widetilde{\mathcal{F}}_t\}$ corresponding to the observable history consisting of controls and observations by

$$\widetilde{\mathcal{F}}_0 = \{\emptyset, \widetilde{\Omega}\}, \qquad \widetilde{\mathcal{F}}_t = \sigma\left(\tilde{U}_0, \tilde{Y}_1, \ldots, \tilde{U}_{t-1}, \tilde{Y}_t\right), \quad t \geq 1.$$

Definition

The fictitious process will be constructed from a directed graphical model. Roughly speaking, it is defined such that Q alters the POMDP model in the first stage, and from time 1 the process evolves like the original POMDP. The first stage model of the fictitious process, which will correspond to cost approximation schemes defined later, is subjected to our choice, provided that certain marginal distributions of the states, controls and observations (marginalized over Q) are preserved, so that the expected cost in the fictitious process and the expected cost in the original POMDP are equal for any common policy.

Formally, for a given initial distribution ξ, let there be an acyclic directed graph G_ξ on the vertex set $\mathcal{V} = \{Q_1, \ldots, Q_m, \tilde{S}_0, \tilde{U}_0, \tilde{S}_1, \tilde{Y}_1\}$ with a set of transition probabilities $\{P_V(V_{pa}, \cdot) \mid V \in \mathcal{V}\}$, where $V_{pa} \subset \mathcal{V}$ denotes the parent vertices of V. Furthermore, let there be no parent node to $\tilde{U}_0$, i.e., $V_{pa} = \emptyset$ for $V = \tilde{U}_0$. By choosing such a graph G_ξ, we specify the probabilistic independence structure of those random variables involved in the first stage of the fictitious process.

For *any* policy $\pi = (\mu_t)_{t \geq 0}$ defined as in Section 2.1, (i.e., a collections of conditional control probabilities), a probability measure $\widetilde{\mathbb{P}}^{\xi,\pi}$ is induced on $\left(\widetilde{\Omega}, \widetilde{\mathcal{F}}\right)$. The choice of G_ξ and $\{P_V(V_{pa}, \cdot)\}$ must be such that the fictitious process satisfies the set of conditions listed below.

Condition 3.1. *For a given ξ and all policies $\pi \in \Pi$, the following relations hold.*

1. $\widetilde{\mathbb{P}}^{\xi,\pi}(V \in \cdot \mid V_{pa}) = P_V(V_{pa}, \cdot)$ *for all* $V \in \mathcal{V}$, *i.e.,* $\widetilde{\mathbb{P}}^{\xi,\pi}$ *is consistent with the graph* G_ξ.

2. *Let* $\tilde{h}_k(\tilde{\omega}) = (\tilde{U}_0, \tilde{Y}_1, \ldots, \tilde{U}_{k-1}, \tilde{Y}_k)$, *i.e., the history of controls and observations. Then* $\widetilde{\mathbb{P}}^{\xi,\pi}(\tilde{U}_0 \in \cdot) = \mu_0(\cdot)$, *and for* $k \geq 1$,

$$\widetilde{\mathbb{P}}^{\xi,\pi}(\tilde{U}_k \in \cdot \mid Q, (\tilde{S}_t, \tilde{U}_t, \tilde{Y}_t)_{t<k}, \tilde{S}_k, \tilde{Y}_k)(\tilde{\omega}) = \widetilde{\mathbb{P}}^{\xi,\pi}(\tilde{U}_k \in \cdot \mid \widetilde{\mathcal{F}}_k)(\tilde{\omega}) = \mu_k\left(\tilde{h}_k(\tilde{\omega}), \cdot\right), \tag{3.1}$$

i.e., Q and the states are not observable to the controller. Furthermore, for $k \geq 2$,

$$\widetilde{\mathbb{P}}^{\xi,\pi}\left(\tilde{S}_k \in \cdot \mid Q, (\tilde{S}_t, \tilde{Y}_t, \tilde{U}_t)_{t<k}\right) = \widetilde{\mathbb{P}}^{\xi,\pi}\left(\tilde{S}_k \in \cdot \mid \tilde{S}_{k-1}, \tilde{U}_{k-1}\right) = P_S\left((\tilde{S}_{k-1}, \tilde{U}_{k-1}), \cdot\right), \tag{3.2}$$

$$\widetilde{\mathbb{P}}^{\xi,\pi}\left(\tilde{Y}_k \in \cdot \mid Q, (\tilde{S}_t, \tilde{Y}_t, \tilde{U}_t)_{t<k}, \tilde{S}_k\right) = \widetilde{\mathbb{P}}^{\xi,\pi}\left(\tilde{Y}_k \in \cdot \mid \tilde{U}_{k-1}, \tilde{S}_k\right) = P_Y\left((\tilde{S}_k, \tilde{U}_{k-1}), \cdot\right), \tag{3.3}$$

i.e., beginning from time 1 the process evolves like the original POMDP.

3. The transition probabilities $\{P_V\}$ are such that

$$\widetilde{\mathbb{P}}^{\xi,\pi}(\tilde{S}_0 \in A) = \xi(A), \ \forall A \in \mathcal{B}(\mathcal{S}), \tag{3.4}$$

$$\widetilde{\mathbb{P}}^{\xi,\pi}\left((\tilde{S}_1, \tilde{Y}_1) \in A \mid \tilde{U}_0 = u\right) = \mathbb{P}^{\xi,\pi}\left((S_1, Y_1) \in A \mid U_0 = u\right), \tag{3.5}$$
$$\forall A \in \mathcal{B}\left(\mathcal{S} \times \mathcal{Y}\right), \ \forall u \in \mathcal{U}.$$

In other words, the marginal distribution of $\tilde{S}_0$ is the same as that of S_0, and the conditional distribution of $(\tilde{S}_1, \tilde{Y}_1)$ given $\tilde{U}_0 = u$ is the same as that of (S_1, Y_1) given $U_0 = u$ in the original POMDP.

The first and second conditions in Condition 3.1 describe the structure of the fictitious process. Only the third condition places a constraint on the model parameters $\{P_V\}$. This condition ensures that by construction the fictitious process has the following property:

Lemma 3.1. *For any policy $\pi \in \Pi$, $(\tilde{S}_0, \tilde{U}_0)$ has the same marginal distribution as (S_0, U_0), and $(\tilde{S}_t, \tilde{Y}_t, \tilde{U}_t)_{t \geq 1}$ conditioned on $\tilde{U}_0 = u$ has the same distribution as $(S_t, Y_t, U_t)_{t \geq 1}$ conditioned on $U_0 = u$ in the original POMDP.*

Thus, for a given initial distribution ξ, it holds that for any $\pi \in \Pi$, the expected cost of the fictitious process and the expected cost of the POMDP are equal:

$$E^{\widetilde{\mathbb{P}}^{\xi,\pi}}\left\{\sum_{t=0}^{\infty} \beta^t g(\tilde{S}_t, \tilde{U}_t)\right\} = J_\beta^\pi(\xi), \qquad E^{\widetilde{\mathbb{P}}^{\xi,\pi}}\left\{\sum_{t=0}^{k-1} g(\tilde{S}_t, \tilde{U}_t)\right\} = J_k^\pi(\xi). \tag{3.6}$$

In particular, the above equations hold for the optimal or ϵ-optimal policies (of the POMDP) that are deterministic. We now use these facts to derive inequalities for the optimal cost functions.

3.3 Inequalities for Discounted Infinite-Horizon Case

Let $\widetilde{P}_0^{\xi,u}$ be the marginal distribution of $(Q, \tilde{S}_0, \tilde{S}_1, \tilde{Y}_1)$ in the fictitious process with the initial distribution ξ and initial control u. Define $\tilde{\phi}_u(\xi, (q, y_1))$ to be a version of the conditional distribution of $\tilde{S}_1$ given $(Q, \tilde{Y}_1)$:

$$\tilde{\phi}_u\left(\xi, (q, y_1)\right)(A) = \widetilde{P}_0^{\xi,u}(\tilde{S}_1 \in A \mid Q, \tilde{Y}_1)\Big|_{(Q,\tilde{Y}_1) = (q, y_1)}, \quad \forall A \in \mathcal{B}\left(\mathcal{S}\right),$$

which is the "belief" that would be if Q is revealed prior to control $\tilde{U}_1$ being applied. The inequality we show next says intuitively that if Q is revealed at time 1 then the cost will be smaller for a controller that exploits this information.

Proposition 3.1. *For a given ξ, let $\{Q, \tilde{S}_0, \tilde{U}_0, (\tilde{S}_t, \tilde{Y}_t, \tilde{U}_t)_{t>0}\}$ be a fictitious process satisfying Condition 3.1. Then*

$$J_\beta^*(\xi) \geq \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + \beta\, E^{\widetilde{P}_0^{\xi,u}}\left\{J_\beta^*\left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1))\right)\right\}\right]. \tag{3.7}$$

Proof: First assume a deterministic optimal policy exists, and denote it by π_ξ. Define $\mathcal{G}_1 = \sigma\left(Q, \tilde{U}_0, \tilde{Y}_1\right)$. By Lemma 3.1,

$$J_\beta^*(\xi) = E^{\widetilde{\mathbb{P}}^{\xi,\pi_\xi}}\left\{g(\tilde{S}_0, \tilde{U}_0)\right\} + E^{\widetilde{\mathbb{P}}^{\xi,\pi_\xi}}\left\{E^{\widetilde{\mathbb{P}}^{\xi,\pi_\xi}}\left\{\sum_{t=1}^{\infty} \beta^t g(\tilde{S}_t, \tilde{U}_t)\,\Big|\, \mathcal{G}_1\right\}\right\}. \tag{3.8}$$

Let

$$\nu_{\tilde{\omega}}(\cdot) = \widetilde{\mathbb{P}}^{\xi,\pi_\xi}\left(\tilde{S}_1 \in \cdot \mid \mathcal{G}_1\right)(\tilde{\omega}),$$

then by construction of the fictitious process (Condition 3.1), for any $A \in \mathcal{F}$,

$$\widetilde{\mathbb{P}}^{\xi,\pi_\xi}\left(\left(\tilde{S}_1, \tilde{U}_1, (\tilde{S}_t, \tilde{U}_t, \tilde{Y}_t)_{t>1}\right) \in A \mid \mathcal{G}_1\right)(\tilde{\omega}) = \mathbb{P}^{\nu_{\tilde{\omega}},\pi_{\tilde{\omega}}}\Big(\,(S_0, U_0, (S_t, U_t, Y_t)_{t>0}) \in A\Big),$$

where $\pi_{\tilde{\omega}} = (\mu'_t)_{t\geq 0}$ is a policy defined by, assuming $\pi_\xi = (\mu_t)_{t\geq 0}$,

$$\mu'_t(h_t, A) = \mu_{t+1}\left((\tilde{U}_0, \tilde{Y}_1, h_t), A\right), \quad \forall A \in \mathcal{B}(\mathcal{U}).$$

Hence,

$$E^{\widetilde{\mathbb{P}}^{\xi,\pi_\xi}}\left\{\sum_{t=1}^{\infty} \beta^t g(\tilde{S}_t, \tilde{U}_t)\,\Big|\, \mathcal{G}_1\right\}(\tilde{\omega}) = \beta\, E^{\mathbb{P}^{\nu_{\tilde{\omega}},\pi_{\tilde{\omega}}}}\left\{\sum_{t=0}^{\infty} \beta^t g(S_t, U_t)\right\} \geq \beta J_\beta^*(\nu_{\tilde{\omega}}). \tag{3.9}$$

It follows from Eq. (3.8) and (3.9) that, assuming $U_0 = \bar{u}$ for π_ξ,

$$\begin{aligned} J_\beta^*(\xi) &\geq E^{\widetilde{\mathbb{P}}^{\xi,\pi_\xi}}\left\{g(\tilde{S}_0, \tilde{U}_0)\right\} + \beta\, E^{\widetilde{\mathbb{P}}^{\xi,\pi_\xi}}\left\{J_\beta^*\left(\nu_{\tilde{\omega}}\right)\right\} \\ &= \bar{g}(\xi, \bar{u}) + \beta\, E^{\widetilde{P}_0^{\xi,\bar{u}}}\left\{J_\beta^*\left(\tilde{\phi}_{\bar{u}}(\xi, (Q, \tilde{Y}_1))\right)\right\} \\ &\geq \inf_{u\in\mathcal{U}}\left[\bar{g}(\xi, u) + \beta\, E^{\widetilde{P}_0^{\xi,u}}\left\{J_\beta^*\left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1))\right)\right\}\right], \end{aligned}$$

where the expectation in the right-hand side of the first equation is justified because J_β^* is measurable and bounded, and $\nu_{\tilde{\omega}}$ is $\mathcal{G}_1$-measurable, and hence $J_\beta^*(\nu_{\tilde{\omega}})$ is $\mathcal{G}_1$-measurable; and similarly the expectations in the second and third equations are justified.

When a deterministic optimal policy does not exist, we can take a sequence of deterministic ϵ_k-optimal policies with $\epsilon_k \downarrow 0$ and repeat the same argument. Thus the claim is proved. □

Remark 3.1. The proof above does not assume continuity or concavity property of J_β^*. In summary it used the following facts of the POMDP problem: (i) J_β^* is measurable and bounded; and (ii) there exist deterministic optimal or ϵ-optimal policies to the POMDP problem.

One example of the fictitious processes is the following simple construction based on replacing the initial distribution by a mixture of distributions. The corresponding inequality is later to be used as a discretized lower cost approximation scheme. The inequality also implies that J_β^* is concave.

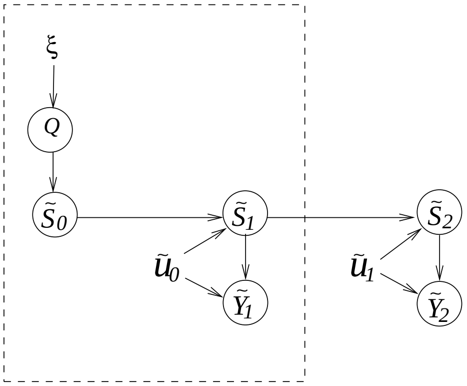

Figure 3-2: The graphical model of the fictitious process in Example 3.1.

Example 3.1. For any $\xi, \xi_i \in \mathcal{P}(\mathcal{S}), \gamma_i \in [0,1], i = 1, \ldots k$ such that $\xi = \sum_{i=1}^{k} \gamma_i \xi_i$, we generate $\tilde{S}_0$, whose marginal distribution is ξ, from a mixture of distributions $\{\xi_i\}$. The directed graphical model is shown in Fig. 3-2. Let $Q \in \mathcal{M} = \{1, 2, \ldots, k\}$, and define

$$P_Q(Q = i) = \gamma_i, \quad i = 1, \ldots, k; \qquad P_{\tilde{S}_0}(i, \tilde{S}_0 \in \cdot) = \xi_i(\cdot).$$

Condition 3.1 is clearly satisfied. Using the relation

$$\begin{aligned} E^{\widetilde{P}_0^{\xi,u}} \left\{ J_\beta^* \left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1)) \right) \right\} &= E^{\widetilde{P}_0^{\xi,u}} \left\{ E^{\widetilde{P}_0^{\xi,u}} \left\{ J_\beta^*\big(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1))\big) \mid Q \right\} \right\} \\ &= \sum_{i=1}^{k} \gamma_i \, E^{P_0^{\xi_i,u}} \left\{ J_\beta^*\big(\phi_u(\xi_i, Y_1)\big) \right\}, \end{aligned}$$

the inequality of Prop. 3.1 is now specialized to the following equation.

Corollary 3.1. *For $\xi = \sum_{i=1}^{k} \gamma_i \xi_i$, $\xi, \xi_i \in \mathcal{P}(\mathcal{S}), \gamma_i \in [0,1], i = 1, \ldots k$,*

$$J_\beta^*(\xi) \geq \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + \beta \sum_{i=1}^{k} \gamma_i \, E^{P_0^{\xi_i,u}} \left\{ J_\beta^*\big(\phi_u(\xi_i, Y_1)\big) \right\} \right]. \tag{3.10}$$

By exchanging the order of inf and summation, using the fact that $\bar{g}(\xi, u) = \sum_i \gamma_i \, \bar{g}(\xi_i, u)$, we have

$$J_\beta^*(\xi) \;\geq\; \sum_{i=1}^{k} \gamma_i \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi_i, u) + \beta \, E^{P_0^{\xi_i,u}} \left\{ J_\beta^*\big(\phi_u(\xi_i, Y_1)\big) \right\} \right] \;=\; \sum_{i=1}^{k} \gamma_i \, J_\beta^*(\xi_i),$$

which implies the concavity of the function J_β^*. □

Remark 3.2. This gives the second way of proving the concavity of the optimal cost functions. In the literature, an alternative proof of concavity of J_β^* assumes the case of discrete spaces and uses a backward induction argument [Åst69]. In Section 3.7, we will extend the backward induction argument to the general space case.

More Complicated Inequalities

We can define more complicated fictitious processes by altering the model of the first k stages of the original POMDP, instead of only the first stage model as we did. Correspondingly we can derive more complicated inequality expressions. These fictitious processes can still be viewed as fictitious processes as we defined earlier, if we consider the equivalent POMDP problem that has each one of its stages identical to k stages of the original POMDP, (and enlarge the state, observation and control spaces accordingly). For simplicity, we will not discuss these inequalities here, although we will mention a few examples in the later chapters.

3.4 Inequalities for Finite-Horizon Case

The inequality for the optimal k-stage cost function follows from the same argument as in the proof of Prop. 3.1:

Proposition 3.2. *For a given ξ, let $\{Q, \tilde{S}_0, \tilde{U}_0, (\tilde{S}_t, \tilde{Y}_t, \tilde{U}_t)_{t>0}\}$ be a fictitious process satisfying Condition 3.1. Then*

$$J_k^*(\xi) \geq \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + E^{\widetilde{P}_0^{\xi,u}} \left\{ J_{k-1}^* \left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1)) \right) \right\} \right]. \tag{3.11}$$

Our interest of the finite-horizon problem is in its relation to the average cost problem. From the finite-stage inequality (3.11) we can derive lower bounds of the optimal liminf average cost function. To this end, noticing that the fictitious processes and their associated inequalities are defined for one initial distribution ξ, we now consider the set of fictitious processes for all ξ.

Definition 3.1. We call the set $\{\widetilde{P}_0^{\xi,u} \mid \xi \in \mathcal{P}(\mathcal{S}), u \in \mathcal{U}\}$ a *lower approximation scheme*, if for each ξ a fictitious process is defined with the induced probability measures satisfying Condition 3.1 and with $\widetilde{P}_0^{\xi,u}$ being the corresponding law of $(Q, \tilde{S}_0, \tilde{S}_1, \tilde{Y}_1)$, and furthermore, $\tilde{\phi}_u(\xi, (q, y_1))$ as a function of (ξ, u, q, y_1) is Borel-measurable.

Define $\mathcal{T}$, the DP mapping of the POMDP, and $\widetilde{\mathcal{T}}$, the mapping associated with a lower approximation scheme $\{\widetilde{P}_0^{\xi,u} \mid \xi \in \mathcal{P}(\mathcal{S}), u \in \mathcal{U}\}$ by

$$(\mathcal{T}J)(\xi) = \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + E^{P_0^{\xi,u}} \left\{ J\big(\phi_u(\xi, Y_1)\big) \right\} \right],$$
$$(\widetilde{\mathcal{T}}J)(\xi) = \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + E^{\widetilde{P}_0^{\xi,u}} \left\{ J \left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1)) \right) \right\} \right].$$

We assume that both $\mathcal{T}$ and $\widetilde{\mathcal{T}}$ preserve measurability, (which can be satisfied by making proper continuity or lower semi-analytic assumptions on J and g), so that $\mathcal{T}^k$ and $\widetilde{\mathcal{T}}^k$ are well defined. Clearly both $\mathcal{T}$ and $\widetilde{\mathcal{T}}$ have the *monotonicity* property, i.e.,

$$J_1 \geq J_2 \quad \Rightarrow \quad \mathcal{T}J_1 \geq \mathcal{T}J_2, \quad \widetilde{\mathcal{T}}J_1 \geq \widetilde{\mathcal{T}}J_2.$$

The Modified Belief MDP

Corresponding to a lower cost approximation scheme, we define a belief MDP on $\mathcal{P}(\mathcal{S})$ with the mapping $\widetilde{T}$ as its DP mapping. We call this belief MDP the *modified belief MDP*, or the modified MDP in short. Its per-stage cost function is defined by $\bar{g}(\xi, u)$ and its state transition probability, denoted by $\widetilde{P}_\xi$, is defined by

$$\widetilde{P}_\xi\left((\xi, u), A\right) = \int \mathbf{1}_A \left(\tilde{\phi}_u\big(\xi, (q, y_1)\big)\right) d\widetilde{P}^{\xi,u}_{q,y_1}, \quad \forall \xi \in \mathcal{P}(\mathcal{S}),\ u \in \mathcal{U},\ A \in \mathcal{B}(\mathcal{P}(\mathcal{S})),$$

where $\widetilde{P}^{\xi,u}_{q,y_1}$ denotes the marginal distribution of $(Q, \tilde{Y}_1)$ corresponding to the joint distribution $\widetilde{P}^{\xi,u}_0$ of $(Q, \tilde{S}_0, \tilde{S}_1, \tilde{Y}_1)$. (In other words, the modified belief MDP is constructed by repeating and joining together at every stage the first-stage model of the fictitious processes.)

Our First (Weaker) Average Cost Lower Bound

Define $J_0(\xi) = 0$ for all $\xi \in \mathcal{P}(\mathcal{S})$. Then the k-stage optimal cost J^*_k satisfies $J^*_k = \mathcal{T}^k J_0$ by the finite-stage optimality equation (i.e., Bellman equation). Define

$$\tilde{J}^*_k(\cdot) = (\widetilde{\mathcal{T}}^k J_0)(\cdot), \quad k \geq 0.$$

The function $\tilde{J}^*_k$ is the k-stage optimal cost function of the modified belief MDP. We claim that $J^*_k \geq \tilde{J}^*_k$. To see this, we use induction: first, $J^*_0 = \tilde{J}^*_0 = J_0$; suppose $J^*_{k-1} \geq \tilde{J}^*_{k-1}$, and it follows then from inequality (3.11) and the monotonicity of $\widetilde{\mathcal{T}}$ that

$$J^*_k \geq \widetilde{\mathcal{T}} J^*_{k-1} \geq \widetilde{\mathcal{T}} \tilde{J}^*_{k-1} = \tilde{J}^*_k. \tag{3.12}$$

Recall from the definition of the optimal liminf average cost

$$J^*_-(\xi) = \inf_{\pi \in \Pi} \liminf_{k \to \infty} \frac{1}{k} J^\pi_k(\xi) \geq \liminf_{k \to \infty} \inf_{\pi \in \Pi} \frac{1}{k} J^\pi_k(\xi) = \liminf_{k \to \infty} \frac{1}{k} J^*_k(\xi). \tag{3.13}$$

Hence we have the following proposition.

Proposition 3.3. *Let $J_0 = 0$ and $\widetilde{\mathcal{T}}$ defined by a lower approximation scheme from Definition 3.1. Then*

$$\liminf_{k \to \infty} \frac{1}{k} \left(\widetilde{\mathcal{T}}^k J_0\right)(\xi) \leq J^*_-(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Remark 3.3. When specialized to discretized approximation schemes in POMDPs with finite spaces, the quantity of the left-hand side can be computed exactly, and hence we obtain lower bounds on the optimal liminf cost function of a POMDP.

Remark 3.4. For the β-discounted case, it can be seen that $\widetilde{\mathcal{T}}$ is also a contraction mapping. Hence by a similar argument, $\tilde{J}^*_\beta \leq J^*_\beta$, where $\tilde{J}^*_\beta$ and J^*_β are the optimal discounted cost functions of the modified belief MDP and the original POMDP, respectively.

Proposition 3.3 will be strengthened later in Section 3.8. Let us address here one motivation for strengthening it. Note that similar to Eq. (3.13), it can be shown that by definition the optimal liminf average cost function $\tilde{J}^*_-$ of the modified belief MDP satisfies

$$\tilde{J}^*_-(\xi) \geq \liminf_{k \to \infty} \frac{1}{k} (\widetilde{\mathcal{T}}^k J_0)(\xi).$$

Thus, in order for us to claim $\tilde{J}^*_- \leq J^*_-$ using Prop. 3.3, we will have to rely on the convergence of value iteration $\frac{1}{k}\widetilde{\mathcal{T}}^k J_0$ to $\tilde{J}^*_-$, which does not hold for general cases. Similar issues also arise in the total cost case as well as other cases. We will see more motivations in the next section on examples of fictitious processes.

3.5 Examples

We give more examples of fictitious processes that correspond to several approximation schemes from the literature. We categorize the design methods into two conceptually different types. One type we refer to as the "replacing" approach and the other type as the "information oracle" approach – the categorization is more from the point of view of designing a lower approximation scheme, than that of proving a certain scheme being a lower approximation scheme. There can be more methods than those we are aware of at present.

The way we construct the fictitious process in Example 3.1 belongs to the "replacing" approach. In that example we replace the state variable of the original POMDP by a new state variable with the same marginal distribution and generated from a mixture of distributions. Example 3.2 is another example of this approach. However, instead of replacing the initial state, it replaces the first state by another variable with the same (conditional) marginal distribution and generated from a mixture of distributions. Example 3.2 is a fictitious-process interpretation of the approximation scheme proposed initially by Zhou and Hanson [ZH01] for discounted problems. When a fixed set of component distributions in the mixture are chosen for all initial distributions, the "replacing" approach naturally quantizes the belief space, and is one straightforward way of generating discretized approximation schemes. These schemes are also called grid-based approximations and will be the focus of the subsequent chapters.

The "information oracle" approach to constructing fictitious processes is to assume the presence of an "oracle" that reveals to the controller certain information of the hidden states of the original POMDP. Example 3.3 is an example of this approach, in which the oracle reveals the subset of state space that contains the true state. It is initially proposed by Zhang and Liu [ZL97] as a continuous cost approximation method for discounted problems. The oracle approach is a fairly intuitive way for constructing approximation schemes. Besides subsets of state space, one can also let the oracle reveal the previous state, or states k-step earlier, or components of the state variable, etc.

The conditions of the fictitious process will be automatically satisfied. The free parameters in the model can be further chosen in a way suitable for approximating the original POMDP.

Example 3.2. The graphical model of the fictitious process is shown in Fig. 3-3 (right), (comparing the graphical model of Example 3.1 shown on the left). We will generate $\tilde{S}_1$ from a mixture of distributions ξ_i, $i = 1, \ldots, k$. The model parameters are as follows.

Let $P_{\tilde{S}_0} = \xi$, and $Q \in \{1, 2, \ldots, k\}$. To preserve the marginal distributions when marginalized over Q, we need to define transition probabilities $P_{\tilde{Y}_1}(u, \tilde{Y}_1 \in \cdot)$, $P_Q((y_1, u), Q \in \cdot)$ and $P_{\tilde{S}_1}(q, \tilde{S}_1 \in \cdot)$ such that they satisfy the condition

$$\widetilde{\mathbb{P}}^{\xi,\pi}\left((\tilde{S}_1, \tilde{Y}_1) \in \cdot \mid \tilde{U}_0 = u\right) = \mathbb{P}^{\xi,\pi}((S_1, Y_1) \in \cdot \mid U_0 = u), \quad \forall \pi \in \Pi,$$

i.e., Eq. (3.5) of Condition 3.1. To this end, we first let the marginal distributions of Y_1 and $\tilde{Y}_1$ be equal (conditioned on the control) by defining

$$P_{\tilde{Y}_1}(u, \cdot) = P_y^{\xi,u}(\cdot),$$

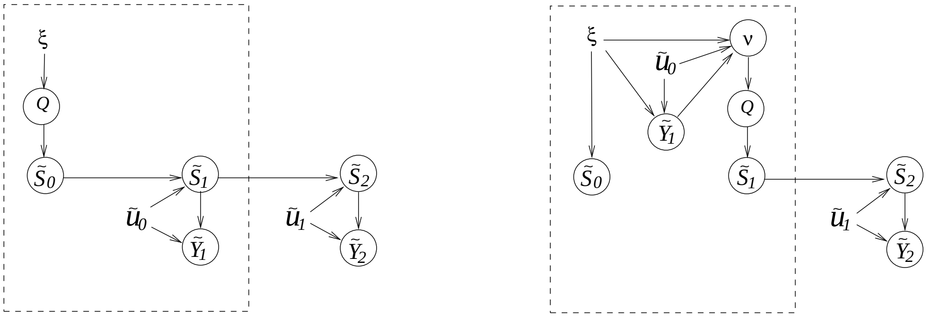

Figure 3-3: Examples of fictitious processes constructed by replacing distributions with mixtures of distributions. Left: a process corresponding to Eq. (3.10); right: a process corresponding to Eq. (3.14).

where $P_y^{\xi,u}$ is the marginal distribution of Y_1 in the original POMDP with initial distribution ξ and initial control u. We then let the conditional distributions of $\tilde S_1$ and S_1, conditioned on $\tilde Y_1$ and Y_1, respectively, be equal by applying the following steps:

- Define $\nu = \phi_u(\xi, \tilde Y_1)$, and assume that for $P_y^{\xi,u}$-almost every y, $\phi_u(\xi, y)$ can be expressed as a convex combination of ξ_i with coefficients $\gamma_i\big(\phi_u(\xi, y)\big)$, i.e.,

$$\phi_u(\xi, \tilde Y_1) = \sum_{i=1}^{k} \gamma_i\big(\phi_u(\xi, y)\big)\, \xi_i.$$

- Define

$$P_Q\big((y,u), \{i\}\big) = \gamma_i\big(\phi_u(\xi, y)\big), \qquad P_{\tilde S_1}(i, \tilde S_1 \in \cdot\,) = \xi_i(\cdot), \quad i = 1, \dots, k.$$

Condition 3.1 is thus satisfied.

Using the relation

$$\begin{aligned} E^{\tilde P_0^{\xi,u}}\Big\{ J_\beta^*\Big(\tilde\phi_u(\xi,(Q,\tilde Y_1))\Big)\Big\} &= E^{\tilde P_0^{\xi,u}}\Big\{ E^{\tilde P_0^{\xi,u}}\Big\{ J_\beta^*\Big(\tilde\phi_u(\xi,(Q,\tilde Y_1))\Big)\,\Big|\, \tilde Y_1\Big\}\Big\} \\ &= E^{\tilde P_0^{\xi,u}}\left\{ \sum_{i=1}^{k} \gamma_i\big(\phi_u(\xi,\tilde Y_1)\big) J_\beta^*(\xi_i)\right\} \\ &= E^{P_y^{\xi,u}}\left\{ \sum_{i=1}^{k} \gamma_i\big(\phi_u(\xi,Y_1)\big) J_\beta^*(\xi_i)\right\} \\ &= E^{P_0^{\xi,u}}\left\{ \sum_{i=1}^{k} \gamma_i\big(\phi_u(\xi,Y_1)\big) J_\beta^*(\xi_i)\right\}, \end{aligned}$$

the inequality of Prop. 3.1 for the β-discounted case is then specialized to

$$J_\beta^*(\xi) \geq \inf_{u \in \mathcal{U}} \left[\bar g(\xi, u) + \beta\, E^{P_0^{\xi,u}}\left\{ \sum_{i=1}^{k} \gamma_i\big(\phi_u(\xi, Y_1)\big) J_\beta^*(\xi_i)\right\}\right]. \tag{3.14}$$

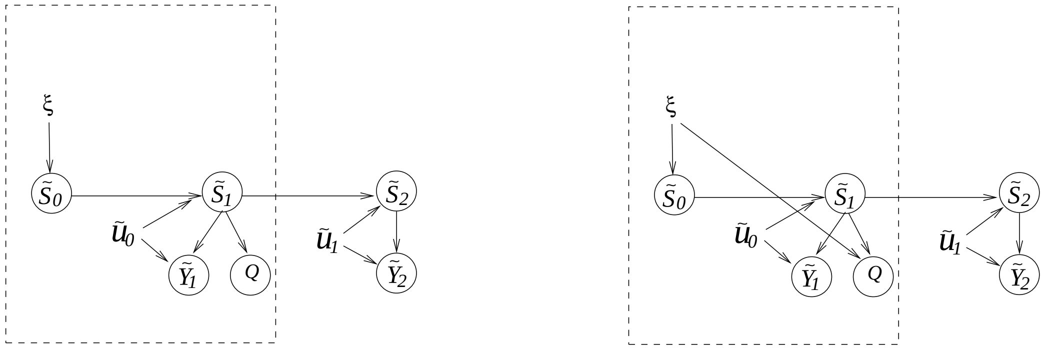

Figure 3-4: Examples of fictitious processes from the "information oracle" approach. Left: a process corresponding to the region-observable POMDP; right: a variant of the left.

Similarly, the inequality of Prop. 3.2 for the finite-stage case is specialized to

$$J_k^*(\xi) \geq \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + \beta \, E^{P_0^{\xi,u}} \left\{ \sum_{i=1}^{k} \gamma_i\big(\phi_u(\xi, Y_1)\big) J_{k-1}^*(\xi_i) \right\} \right].$$

The inequality (3.14) was proposed by Zhou and Hansen in [ZH01] for a discretized cost approximation scheme. It was derived in their paper directly from the optimality equation using the concavity of J_β^*. □

Example 3.3 (Region-Observable POMDP). Zhang and Liu [ZL97] proposed the "region-observable" POMDP as a continuous cost approximation scheme. Our description of it is slightly more general. The graphic model is as shown in Fig. 3-4 (left). The random variable Q indicates which subset of state space contains $\tilde{S}_1$ and may be viewed as an additional component of observation. Except for the transition probability P_Q, the rest of the transition probabilities are defined as they are in the original POMDP. Since Q is a leaf node in the graph, the marginal distributions of other random variables do not change when marginalized over Q. Thus, no matter how we define P_Q, the condition of the fictitious process will be satisfied, and P_Q is hence a free parameter that one can choose accordingly for a given problem.

More precisely, let $\{\mathcal{S}_k \mid \mathcal{S}_k \subset \mathcal{S}, k \in \mathcal{K}\}$, where the index set $\mathcal{K}$ is a set of integers, be a collection of subsets of states such that

$$\bigcup_{k \in \mathcal{K}} \mathcal{S}_k = \mathcal{S}.$$

The set $\{\mathcal{S}_k\}$ is called a *region system* by [ZL97]. Let Q be $\mathcal{K}$-valued, and let the transition probability $P_Q(s_1, Q \in \cdot)$ be defined. Then the marginal distribution $\widetilde{P}_0^{\xi,u}$ of $(Q, \tilde{S}_0, \tilde{S}_1, \tilde{Y}_1)$, when the initial distribution is ξ and initial control $\tilde{U}_0 = u$, satisfies

$$\widetilde{P}_0^{\xi,u}\left((\tilde{S}_0, \tilde{S}_1, Q, \tilde{Y}_1) \in A \right) = \iint \mathbf{1}_A(s_0, s_1, q, y_1) P_Q(s_1, dq) \, dP_0^{\xi,u}$$

for all Borel measurable sets A, where $P_0^{\xi,u}$ is the marginal distribution of (S_0, S_1, Y_1) in the original POMDP. The belief at time 1 in the fictitious process is

$$\tilde{\phi}_u(\xi, (q, y_1))(\cdot) = \widetilde{P}_0^{\xi,u}(\tilde{S}_1 \in \cdot \mid \tilde{Y}_1, Q)\Big|_{(Q,\tilde{Y}_1)=(q,y_1)}.$$

Thus we have defined the function for the next belief $\tilde{\phi}_u(\xi, (q, y_1))$ and the distribubtion $\widetilde{P}_0^{\xi,u}$ in the inequality of Prop. 3.1 for the β-discounted case:

$$J_\beta^*(\xi) \geq \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + \beta\, E^{\widetilde{P}_0^{\xi,u}} \left\{ J_\beta^* \left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1)) \right) \right\} \right],$$

and in the inequality of Prop. 3.2 for the finite-stage case.

An important property of region-observable POMDP is

$$\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1))(A) = 0, \quad \text{if } \; A \cap \mathcal{S}_Q = \emptyset,$$

which implies, when $\mathcal{S}$ is discrete, that the next belief $\tilde{\phi}$ is always on the affine space corresponding to some $\mathcal{S}_k$. This property is the key motivation in [ZL97] for proposing the region-observable POMDP, and it reduces the computational complexity of exact value iteration for discounted problems.

As for the choice of the region system, for example, Zhang and Liu [ZL97] choose the so called "radius-k" regions, with each region being a subset of states that are reachable from certain fixed state within k steps. The transition probability P_Q, which corresponds to how a region is revealed given the state s_1, can be chosen in a way that minimizes the information carried by Q, roughly speaking.

Variants of the region-observable POMDP can be formed by, e.g., letting Q depend on the control and observation in addition to the state, or letting Q depend on ξ in addition, as shown in Fig. 3-4 (right). □

Differences between Information Oracle and Non-Information Oracle Approaches

At least in our opinion, clearly there are conceptual differences between the information oracle type of schemes, such as region-observable POMDPs, and the non-information oracle type of schemes, such as those based on discretizing the belief space.

Mathematically, at first there seems to be a difference as well. For example, as mentioned earlier, the average cost lower bounds established by Prop. 3.3 rely on the DP mapping $\widetilde{T}$, and are in general smaller than the optimal average cost functions of the modified belief MDPs. On the other hand, one can claim a stronger result for an information oracle type of scheme: the optimal cost function of its associated modified belief MDP is a lower bound of the optimal of the original problem. The reason is that an information oracle type of scheme can be viewed as the original POMDP plus additional observation variables at every stage. Thus it has the same state evolution model as the original POMDP and however a larger policy space than the latter, and its optimal cost is consequently no greater than the optimal of the original problem.

This difference motivates us to strengthen the analysis given earlier. Indeed we will show that the stronger lower bound statement holds for non-information oracle type of schemes as well. So mathematically, there seem to be no differences between the information oracle and non-information oracle approaches.

3.6 Comparison of Approximation Schemes

Given two lower approximation schemes, we may be interested in comparing the lower bounds associated with them. In general it is hard to claim one scheme strictly dominating the other (i.e., better for every ξ), if the schemes are constructed to have different probabilistic dependence structures. We give such a comparison in the next proposition under restrictive conditions. The proposition says intuitively that the less information is revealed to the controller, the better is the approximation.

Proposition 3.4. *Let $\widetilde{\mathcal{T}}$ be defined by a lower approximation scheme as defined in Definition 3.1 with $Q = (Q_1, Q_2)$. For $i = 1, 2$, let $\tilde{\phi}^i_u(\xi, (Q_i, \tilde{Y}_1))$ be the conditional distributions of $\tilde{S}_1$ given $(Q_i, \tilde{Y}_1)$, respectively. Then for $i = 1$ or 2,*

$$J^*_\beta(\xi) \geq (\widetilde{\mathcal{T}}_i J^*_\beta)(\xi) \geq (\widetilde{\mathcal{T}} J^*_\beta)(\xi), \qquad \beta \in [0, 1),$$
$$J^*_k(\xi) \geq (\widetilde{\mathcal{T}}_i J^*_{k-1})(\xi) \geq (\widetilde{\mathcal{T}} J^*_{k-1})(\xi), \qquad \beta = 1,$$

where $\widetilde{\mathcal{T}}_i$ and $\widetilde{\mathcal{T}}$ are defined by

$$(\widetilde{\mathcal{T}}_i J)(\xi) = \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + \beta \, E^{\widetilde{P}^{\xi,u}_0} \left\{ J \left(\tilde{\phi}^i_u(\xi, (Q_i, \tilde{Y}_1)) \right) \right\} \right],$$
$$(\widetilde{\mathcal{T}} J)(\xi) = \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + \beta \, E^{\widetilde{P}^{\xi,u}_0} \left\{ J \left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1)) \right) \right\} \right].$$

Proof: The proof for $\beta < 1$ and $\beta = 1$ is similar, so we only consider $\beta < 1$. That $J^*_\beta(\xi) \geq (\widetilde{\mathcal{T}}_i J^*_\beta)(\xi)$ follows from the same argument as in the proof of Prop. 3.1. To show $(\widetilde{\mathcal{T}}_i J^*_\beta)(\xi) \geq (\widetilde{\mathcal{T}} J^*_\beta)(\xi)$, it is sufficient to show that for $i = 1$ and for any u,

$$E^{\widetilde{P}^{\xi,u}_0} \left\{ J^*_\beta \left(\tilde{\phi}^1_u(\xi, (Q_1, \tilde{Y}_1)) \right) \right\} \geq E^{\widetilde{P}^{\xi,u}_0} \left\{ J^*_\beta \left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1)) \right) \right\}.$$

Since

$$E^{\widetilde{P}^{\xi,u}_0} \left\{ \tilde{\phi}_u(\xi, (Q, \tilde{Y}_1)) \,\middle|\, Q_1, \tilde{Y}_1 \right\} = \tilde{\phi}^1_u(\xi, (Q_1, \tilde{Y}_1)),$$

and J^*_β is concave, it follows then from Jensen's inequality for conditional expectation that

$$\begin{aligned} E^{\widetilde{P}^{\xi,u}_0} \left\{ J^*_\beta \left(\tilde{\phi}^1_u(\xi, (Q_1, \tilde{Y}_1)) \right) \right\} &\geq E^{\widetilde{P}^{\xi,u}_0} \left\{ E^{\widetilde{P}^{\xi,u}_0} \left\{ J^*_\beta \left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1)) \right) \,\middle|\, Q_1, \tilde{Y}_1 \right\} \right\} \\ &= E^{\widetilde{P}^{\xi,u}_0} \left\{ J^*_\beta \left(\tilde{\phi}_u(\xi, (Q, \tilde{Y}_1)) \right) \right\}, \end{aligned}$$

and hence the claim. □

Remark 3.5. The above proposition does not imply $\widetilde{\mathcal{T}}^k_i J_0 \geq \widetilde{\mathcal{T}}^k J_0$, for which we need stronger conditions such as either $\widetilde{\mathcal{T}}_i$ or $\widetilde{\mathcal{T}}$ preserves concavity – usually this property does not hold for discretized approximation schemes.

Remark 3.6. The above proof gives an alternative way of proving the inequalities using the concavity of the optimal cost functions: For a lower approximation scheme $\widetilde{T}$ with Q, let $Q_1 = c$ be a constant dummy variable and let $Q_2 = Q$. Then $\widetilde{\mathcal{T}}_1 = \mathcal{T}$, the DP mapping of the POMDP; and the preceding proof establishes that $\mathcal{T} J \geq \widetilde{\mathcal{T}} J$ for any concave function J. This fact together with the concavity perserving property of $\mathcal{T}$ (i.e., $\mathcal{T} J$ is concave whenever

J is concave) then implies that the inequalities hold for the case where the per-stage cost function $\bar{g}(\xi, u)$ is concave in ξ.

3.7 Concavity Preserving Property of the DP Mapping $\mathcal{T}$

Recall that the DP mapping $\mathcal{T}$ with $\beta \in [0, 1]$ is defined as

$$(\mathcal{T}J)(\xi) = \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + \beta E^{P_0^{\xi,u}} \left\{ J\big(\phi_u(\xi, Y_1)\big) \right\} \right].$$

It is true that $\mathcal{T}J$ is concave whenever J is concave. Using this property together with a backward induction argument, one can conclude that the optimal discounted and finite-stage cost functions are concave. This is yet another alternative method of proving the concavity of those cost functions. In fact, chronologically it is the first method, given by Åström [Åst69], who proved for the case of discrete space models.

Here, for completeness, we give a proof of the concavity preserving property of $\mathcal{T}$ for general space models. The idea of proof being the same as that of [Åst69], technically the proof is however different. We also avoided the use of any fictitious process argument. This makes our proof different from the one in [RS94].

To show that $\mathcal{T}J$ is concave whenever J is concave, it is sufficient to show that for any given control u, $E^{P_0^{\xi,u}} \left\{ J\big(\phi_u(\xi, Y_1)\big) \right\}$ is a concave function of ξ.

Proposition 3.5. $E^{P_0^{\xi,u}}\left\{ J\big(\phi_u(\xi, Y_1)\big) \right\}$ *is a concave function of* ξ.

Proof: Suppose $\xi = \sum_{i=1}^k \alpha_i \xi_i$, where $\alpha_i \in (0, 1], \sum_{i=1}^k \alpha_i = 1$. Abusing notation, let us denote again by $P_0^{\xi,u}$ (respectively, $P_0^{\xi_i,u}$) the marginal distribution of (S_1, Y_1) under the initial control u and the initial distribution ξ (respectively, ξ_i), and denote by $P_y^{\xi,u}$ (respectively, $P_y^{\xi_i,u}$) the marginal distribution of Y_1. Then $P_0^{\xi,u} = \sum_{i=1}^k \alpha_i P_0^{\xi_i,u}$ and $P_y^{\xi,u} = \sum_{i=1}^k \alpha_i P_y^{\xi_i,u}$. There exist Radon-Nikodym derivatives $\frac{dP_0^{\xi_i,u}}{dP_0^{\xi,u}}(s_1, y_1)$ and $\frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)$, and they satisfy for any s_1 and y_1,

$$\sum_i \alpha_i \frac{dP_0^{\xi_i,u}}{dP_0^{\xi,u}}(s_1, y_1) = 1, \qquad \sum_i \alpha_i \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1) = 1.$$

First we prove for $P_y^{\xi,u}$-almost surely all y_1

$$\phi_u(\xi, y_1) = \sum_i \alpha_i \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)\, \phi_u(\xi_i, y_1).$$

Recall ϕ_u is the conditional distribution of S_1 given $Y_1 = y_1$. For any $A \in \mathcal{B}(\mathcal{S})$ and

$B \in \mathcal{B}(\mathcal{Y})$,

$$\begin{aligned}
P_0^{\xi,u}(S_1 \in A, Y_1 \in B) &= \int_B \phi_u(\xi, y_1)(A)\, dP_y^{\xi,u}(y_1), \\
P_0^{\xi_i,u}(S_1 \in A, Y_1 \in B) &= \int_B \phi_u(\xi_i, y_1)(A)\, dP_y^{\xi_i,u}(y_1) \\
&= \int_B \phi_u(\xi_i, y_1)(A)\, \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)\, dP_y^{\xi,u}(y_1),
\end{aligned}$$

$$\Rightarrow \quad \int_B \phi_u(\xi, y_1)(A)\, dP_y^{\xi,u}(y_1) = \int_B \sum_i \alpha_i \phi_u(\xi_i, y_1)(A)\, \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)\, dP_y^{\xi,u}(y_1). \tag{3.15}$$

For every y_1, since $\sum_i \alpha_i \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1) = 1$, $\sum_i \left(\alpha_i \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)\right) \phi_u(\xi_i, y_1)(\cdot)$ is also a probability measure on $\mathcal{B}(\mathcal{S})$. Hence by Eq. (3.15) for a fixed set A, $\sum_i \alpha_i \phi_u(\xi_i, y_1)(A) \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)$ can only disagree with $\phi_u(\xi, y_1)(A)$ on a set of y_1 with $P_y^{\xi,u}$-measure zero. Because $\mathcal{S}$ is a Borel set of a separable metric space, the Borel σ-algebra $\mathcal{B}(\mathcal{S})$ is generated by a countable number of open sets $\{A_j\}$. Thus, the set of all y_1 for which there exists some A_j such that the terms $\sum_i \alpha_i \phi_u(\xi_i, y_1)(A_j) \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)$ and $\phi_u(\xi, y_1)(A_j)$ disagree, has $P_y^{\xi,u}$-measure zero. In other words, for every y_1 except on a set with $P_y^{\xi,u}$-measure zero, it holds that for all A_j,

$$\phi_u(\xi, y_1)(A_j) = \sum_i \alpha_i \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)\, \phi_u(\xi_i, y_1)(A_j).$$

It follows from the uniqueness in the Caratheodory's extension theorem that

$$\phi_u(\xi, y_1)(A) = \sum_i \alpha_i \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)\, \phi_u(\xi_i, y_1)(A), \quad A \in \mathcal{B}(\mathcal{S}).$$

By the concavity of J, it then follows that

$$\begin{aligned}
\int J\big(\phi_u(\xi, y_1)\big)\, dP_y^{\xi,u}(y_1) &= \int J\Big(\sum_i \alpha_i \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)\, \phi_u(\xi_i, y_1)\Big)\, dP_y^{\xi,u}(y_1) \\
&\geq \int \sum_i \alpha_i \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)\, J\big(\phi_u(\xi_i, y_1)\big)\, dP_y^{\xi,u}(y_1) \\
&= \sum_i \alpha_i \int \frac{dP_y^{\xi_i,u}}{dP_y^{\xi,u}}(y_1)\, J\big(\phi_u(\xi_i, y_1)\big)\, dP_y^{\xi,u}(y_1) \\
&= \sum_i \alpha_i \int J\big(\phi_u(\xi_i, y_1)\big)\, dP_y^{\xi_i,u}(y_1),
\end{aligned}$$

and the proof is complete. □

Remark 3.7. The proof for the universally measurable case is the same, as the measure of the universal measurable sets is uniquely determined by the measure of the Borel sets.

Remark 3.8. Clearly the above proposition implies the more general statement that $\mathcal{T}$ preserves concavity when the per-stage cost function $\bar{g}$ is concave in the belief. Thus the lower bound results of the previous sections apply to the general case of concave per-stage cost models (see Remark 3.6).

3.8 A Strengthened Lower Bound Result

Let $\{\widetilde{P}_0^{\xi,u} \mid \xi \in \mathcal{P}(\mathcal{S}), u \in \mathcal{U}\}$ be a lower approximation scheme as defined by Definition 3.1. Consider its associated modified belief MDP problem. In this section we will strengthen the first average cost lower bound result, Prop. 3.3. In particular, we will prove the following theorem, which has been shown only for those lower approximation schemes constructed from the information oracle approach.

Theorem 3.1. *Let $\tilde{J}_-^*$ and $\tilde{J}_+^*$ be the optimal liminf and limsup average cost functions, respectively, of the modified belief MDP. Then,*

$$\tilde{J}_-^*(\xi) \leq J_-^*(\xi) \qquad \tilde{J}_+^*(\xi) \leq J_+^*(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Furthermore, it is not only for the average cost criterion but also for various expected cost criteria, (including constrained cases), that lower bounds analogous to the ones above hold. This will be implied by the following theorem, which is the main result of this section.

Theorem 3.2. *Given an initial distribution ξ_0, for any policy π of the original POMDP, there exists a policy $\tilde{\pi}$ of the modified belief MDP such that*

$$\tilde{J}_k^{\tilde{\pi}}(\xi_0) = J_k^{\pi}(\xi_0), \qquad \tilde{J}_\beta^{\tilde{\pi}}(\xi_0) = J_\beta^{\pi}(\xi_0), \quad \forall k \geq 1, \ \beta \in [0, 1),$$

for any bounded per-stage cost function g.

It can be seen that Theorem 3.1 is an immediate consequence of Theorem 3.2 applied to the average cost case (the cost of any policy of the original POMDP can be attained by a policy of the modified belief MDP). Theorem 3.2 was shown only for lower approximation schemes constructed form the information oracle approach.

In the rest of this section, we will prove Theorem 3.2. The method of proof is to construct a POMDP, to be called an *approximating POMDP*, that relates in a special way to both the original POMDP and the modified belief MDP associated with a lower approximation scheme. In particular, while similar to that in the information oracle approach, the approximating POMDP can be viewed as the original POMDP plus additional observations, the approximating POMDP is also equivalent to the modified belief MDP.

Theorem 3.2 and the construction of such an approximating POMDP for any lower approximation scheme also illustrate that the distinction between information oracle and non-information oracle approaches is more conceptual, rather than mathematical. On the other hand, the approximating POMDP is introduced here mainly as an intermediate proof device, and its interpretation of a non-information oracle type lower approximation scheme can be highly non-intuitive, and therefore not helpful for designing approximation schemes in our opinion.

In what follows, we define the approximating POMDP in Section 3.8.1. We then analyze the relations of the three processes, namely the approximating POMDP, the original POMDP and the modified belief MDP in Sections 3.8.1 and 3.8.2. Finally we prove the main theorem as well as its extension to the concave per-stage cost case in Section 3.8.3.

3.8.1 Definition of an Approximating POMDP

Define a POMDP

$$\{S_0, U_0, (S_t, Y_t, Q_t, U_t)_{t\geq 1}\}$$

on the canonical sample space $\Omega = \mathcal{S} \times \mathcal{U} \times \prod_{t\geq 1}(\mathcal{S} \times \mathcal{Y} \times \mathcal{M} \times \mathcal{U})$, where Q_t are $\mathcal{M}$-valued observation variables in addition to Y_t. We refer to this POMDP as the approximating POMDP and define its evolution model in what follows.

The evolution of the state S_t and observation Y_t is the *same* as in the original POMDP. The evolution of Q_t is defined for each initial distribution ξ_0 *differently.* The control can depend on Q_ts in addition to Y_ts and U_ts. To make these precise, we first give certain notation to be used later. We order the random variables in the following order:

$$S_0, U_0, S_1, Y_1, Q_1, U_1, \dots$$

The observable random variables are U_ts, Y_ts and Q_ts. Define $\mathcal{F}_t^q$ and $\mathcal{F}_t^u$ to be the σ-algebras generated by the observable random variables prior to (and not including) Q_t and U_t, respectively, and denote the corresponding random variables by H_t^q and H_t^u for short, i.e.,

$$\mathcal{F}_t^q = \sigma(U_0, Y_1, Q_1, U_1, \dots, U_{t-1}, Y_t), \qquad \mathcal{F}_t^u = \sigma(U_0, Y_1, Q_1, U_1, \dots, U_{t-1}, Y_t, Q_t),$$
$$H_t^q = (U_0, Y_1, Q_1, U_1, \dots, U_{t-1}, Y_t), \qquad H_t^u = (U_0, Y_1, Q_1, U_1, \dots, U_{t-1}, Y_t, Q_t).$$

Similarly, define $\mathcal{G}_t^q$ and $\mathcal{G}_t^u$ to be the σ-algebras generated by all the random variables before Q_t and U_t, respectively. Denote by $\mathcal{F}_1 \vee \mathcal{F}_2$ the σ-algebra generated by sets in $\mathcal{F}_1 \cup \mathcal{F}_2$, then $\mathcal{G}_t^q$ and $\mathcal{G}_t^u$ can be equivalently expressed as

$$\mathcal{G}_t^q = \mathcal{F}_t^q \vee \sigma\big((S_k)_{k\leq t}\big), \qquad \mathcal{G}_t^u = \mathcal{F}_t^u \vee \sigma\big((S_k)_{k\leq t}\big).$$

Policy Space

Let $\mathcal{H}_t^u$ be the space of H_t^u, the observed history prior to and not including the control U_t. For each initial distribution, the common set of admissible policies is $\widehat{\Pi} = \{\pi = (\mu_t)_{t\geq 0}\}$, where μ_t are transition probabilities from $\mathcal{H}_t^u$ to the control space $\mathcal{U}$:

$$\mu_t(h_t^u, \cdot) \in \mathcal{P}(\mathcal{U}), \quad \forall h_t^u \in \mathcal{H}_t^u,$$

i.e., U_t depends on the observable variables Y_k and Q_k for $k \leq t$, as well as U_k for $k < t$.

Probabilistic Independence Structure of Q_t

We now define the evolution of Q_t. First we define the belief process in the approximating POMDP. Let $\mathbb{P}^{\xi_0,\pi}$ be the probability measure induced by initial distribution ξ_0 and policy $\pi \in \widehat{\Pi}$. Define the $\mathcal{P}(\mathcal{S})$-valued random variable ξ_t to be a version of the conditional distribution of S_t given H_t^u:

$$\xi_t(\omega)(\cdot) = \mathbb{P}^{\xi_0,\pi}\left(S_t \in \cdot \mid \mathcal{F}_t^u\right)(\omega).$$

The random variable ξ_t is a function of (ξ_0, h_t^u) – the inititial distribution and the sample trajectory, (and note that ξ_t is not a function of π). We call $\{\xi_t\}$ the belief process of the approximating POMDP.

For a given ξ_0, we let Q_t depend on the state S_t and the observed history H_t^q prior to Q_t:

$$\mathbb{P}^{\xi_0,\pi}(Q_t \mid \mathcal{G}_t^q) = \mathbb{P}^{\xi_0,\pi}(Q_t \mid \sigma(S_t) \vee \mathcal{F}_t^q), \tag{3.16}$$

where $\sigma(S_t) \vee \mathcal{F}_t^q = \sigma(S_t, H_t^q)$. Since Q_t is an additional observation variable generated by the state and the past observable variables, the approximating POMDP has

- the same evolution model of the states S_t as in the original POMDP,
- the same evolution model of the observations Y_t as in the original POMDP, and
- a policy space that includes the policy space of the original POMDP, (when the policy of the latter is viewed as a policy of the approximating POMDP that does not functionally depend on Q_t).

These properties of the construction are simple, yet powerful, (as they have been in the information oracle approach). We state their implications in the following lemma.

Denote by $\hat{J}_k^\pi(\xi_0)$ and $\hat{J}_\beta^\pi(\xi_0)$ the k-stage cost and β-discounted cost, respectively, of a policy π of the approximating POMDP.

Lemma 3.2. *Given an initial distribution ξ_0, for any policy π of the original POMDP, there exists a policy $\hat{\pi}$ of the approximating POMDP such that*

$$\hat{J}_k^{\hat{\pi}}(\xi_0) = J_k^\pi(\xi_0), \qquad \hat{J}_\beta^{\hat{\pi}}(\xi_0) = J_\beta^\pi(\xi_0), \qquad \forall k \geq 1, \ \beta \in [0,1),$$

for any bounded per-stage cost function g.

Hence, it can be seen that with respect to the same per-stage cost function $g(s,u)$, *the optimal expected cost of this approximating POMDP is no greater than that of the original POMDP for various expected cost criteria.*

Transition Model of Q_t and Link to the Modified Belief MDP

We now specify the transition model of Q_t, which links the approximating POMDP to the modified belief MDP. The evolution of Q_t is defined *differently* for a different initial distribution ξ_0, and furthermore, Q_t depends on (S_t, H_t^q) in a way that is stationary with respect to $(\xi_{t-1}, U_{t-1}, S_t, Y_t)$. More precisely, we define P_{Q_t}, a transition probability from $\mathcal{P}(\mathcal{S}) \times \mathcal{U} \times \mathcal{S} \times \mathcal{Y}$ – the space of $(\xi_{t-1}, U_{t-1}, S_t, Y_t)$ – to the space of Q_t, to be

$$P_{Q_t}((\xi, u, s, y), \cdot) = \widetilde{P}_0^{\xi,u}(Q \in \cdot \mid \tilde{S}_1 = s, \tilde{Y}_1 = y), \tag{3.17}$$

where $\{\widetilde{P}_0^{\xi,u} \mid \xi \in \mathcal{P}(\mathcal{S}), u \in \mathcal{U}\}$ is the lower approximation scheme (recall that for each (ξ, u), $\widetilde{P}_0^{\xi,u}$ is the law of $(\tilde{S}_0, \tilde{S}_1, \tilde{Y}_1, Q)$ in the fictitious process with initial distribution ξ and initial control u). Thus, with $h_t^q = (h_{t-1}^u, u_{t-1}, y_t)$, the evolution of Q_t in the approximating POMDP can be specified by, (comparing Eq. (3.16)),

$$\begin{aligned} \mathbb{P}^{\xi_0,\pi}(Q_t \in \cdot \mid \sigma(S_t) \vee \mathcal{F}_t^q)\big|_{(H_t^q(\omega), S_t(\omega)) = (h_t^q, s_t)} &= P_{Q_t}\left((\xi_{t-1}(\xi_0, h_{t-1}^u), u_{t-1}, s_t, y_t), \cdot \right) \\ &= \widetilde{P}_0^{\xi_{t-1}, u_{t-1}}(Q \in \cdot \mid \tilde{S}_1 = s_t, \tilde{Y}_1 = y_t), \end{aligned} \tag{3.18}$$

where, abusing notation, in the first equation we write ξ_{t-1} as $\xi_{t-1}(\xi_0, h_{t-1}^u)$ to emphasize that ξ_{t-1} is a function of (ξ_0, h_{t-1}^u). Notice that the transition model of Q_t does not depend on t, and is stationary with respect to $(\xi_{t-1}, U_{t-1}, S_t, Y_t)$.

3.8.2 Equivalence to the Modified Belief MDP

We now show that the approximating POMDP is equivalent to the modified belief MDP associated with the lower approximation scheme $\{\widetilde{P}_0^{\xi,u}\}$. The arguments that we are about to go through, are similar to those for establishing the sufficient statistic for control in POMDPs. For the following analysis, it is hepful to consider the joint process

$$\{S_0, \xi_0, U_0, S_1, Y_1, Q_1, \xi_1, U_1, \ldots\}$$

that includes the beliefs.

Step 1: Change to Per-Stage Cost $\bar{g}$

First, we express the expected cost of a policy π as the expected cost with respect to the per-stage cost $\bar{g}(\xi, u)$.

Lemma 3.3. *In the approximating POMDP, given an initial distribution ξ_0, for any policy π, the expected cost of π with respect to any bounded per-stage cost function g satisfies for all $k \geq 1$ and $\beta \in [0, 1)$,*

$$\hat{J}_k^\pi(\xi_0) = E^{\mathbb{P}^{\xi_0,\pi}} \left\{ \sum_{t=0}^{k-1} \bar{g}(\xi_t, U_t) \right\}, \qquad \hat{J}_\beta^\pi(\xi_0) = E^{\mathbb{P}^{\xi_0,\pi}} \left\{ \sum_{t=0}^{\infty} \beta^t \bar{g}(\xi_t, U_t) \right\}.$$

Proof: Taking iterative conditional expectations, we have for any k,

$$E^{\mathbb{P}^{\xi_0,\pi}} \left\{ \sum_{t=0}^{k-1} g(S_t, U_t) \right\} = \sum_{t=0}^{k-1} E^{\mathbb{P}^{\xi_0,\pi}} \left\{ E^{\mathbb{P}^{\xi_0,\pi}} \{ g(S_t, U_t) \mid \mathcal{F}_t^u \} \right\} = \sum_{t=0}^{k-1} E^{\mathbb{P}^{\xi_0,\pi}} \left\{ \bar{g}(\xi_t, U_t) \right\},$$

where the second equality follows from the conditional independence of S_t and U_t given H_t^u. Similarly, for any discount factor β,

$$\begin{aligned} E^{\mathbb{P}^{\xi_0,\pi}} \left\{ \sum_{t=0}^{\infty} \beta^t g(S_t, U_t) \right\} &= \sum_{t=0}^{\infty} \beta^t E^{\mathbb{P}^{\xi_0,\pi}} \left\{ E^{\mathbb{P}^{\xi_0,\pi}} \{ g(S_t, U_t) \mid \mathcal{F}_t^u \} \right\} \\ &= \sum_{t=0}^{\infty} \beta^t E^{\mathbb{P}^{\xi_0,\pi}} \left\{ \bar{g}(\xi_t, U_t) \right\} = E^{\mathbb{P}^{\xi_0,\pi}} \left\{ \sum_{t=0}^{\infty} \beta^t \bar{g}(\xi_t, U_t) \right\}, \end{aligned}$$

where the interchange of summation and expectation is justified by the dominated convergence theorem, (since the per-stage cost function is assumed bounded). □

Thus, in the approximating POMDP, the expected cost of π can be equivalently defined as the expected cost of π with respect to the per-stage cost $\bar{g}(\xi, u)$.

Step 2: A Markovian Property of the Belief Process

Next, we show that with $(U_{t-1}, Y_t, Q_t) = (u_{t-1}, y_t, q_t)$, ξ_t is a function of $(\xi_{t-1}, u_{t-1}, y_t, q_t)$, and this function does not depend on t. Therefore, $\{\xi_t\}$ evolves in a stationary and Markov way in the approximating POMDP. Generally speaking, this property is due to the evolution model of Q_t as defined by Eq. (3.17). We now give the detailed analysis as follows.

Recall that the $\mathcal{P}(\mathcal{S})$-valued random variable ξ_t is defined as

$$\xi_t(\omega)(\cdot) = \mathbb{P}^{\xi_0,\pi}\left(S_t \in \cdot \mid \mathcal{F}_t^u\right)(\omega).$$

Consider the (random) law $P_t(\omega)$ of (S_{t-1}, S_t, Y_t, Q_t) defined by

$$P_t(\omega)\left((S_{t-1}, S_t, Y_t, Q_t) \in \cdot\right) = \mathbb{P}^{\xi_0,\pi}\big((S_{t-1}, S_t, Y_t, Q_t) \in \cdot \mid \mathcal{F}_{t-1}^u \vee \sigma(U_{t-1})\big)(\omega). \tag{3.19}$$

Since $H_t^u = (H_{t-1}^u, U_{t-1}, Y_t, Q_t)$ and $\mathcal{F}_t^u = \mathcal{F}_{t-1}^u \vee \sigma(U_{t-1}, Y_t, Q_t)$, comparing the preceding two relations, we have, with $(Y_t(\omega), Q_t(\omega)) = (y_t, q_t)$,

$$\xi_t(\omega)(\cdot) = P_t(\omega)(S_t \in \cdot \mid Y_t = y_t, Q_t = q_t). \tag{3.20}$$

Lemma 3.4. *In the approximating POMDP, given an initial distribution ξ_0, there exists a transition probability P_ξ from $\mathcal{P}(\mathcal{S}) \times \mathcal{U} \times \mathcal{Y} \times \mathcal{M}$ to $\mathcal{P}(\mathcal{S})$ such that for any policy π and $t \geq 1$, with $(U_{t-1}(\omega), Y_t(\omega), Q_t(\omega)) = (u_{t-1}, y_t, q_t)$,*

$$\mathbb{P}^{\xi_0,\pi}\left(\xi_t \in \cdot \mid \mathcal{F}_t^u\right)(\omega) = P_\xi\big((\xi_{t-1}, u_{t-1}, y_t, q_t), \cdot\big). \tag{3.21}$$

Proof: It is sufficient to show that ξ_t is a function of $(\xi_{t-1}, u_{t-1}, y_t, q_t)$ and this function is the same for all t. To show this, by Eq. (3.20), it is sufficient to show that the random law P_t is a function of (ξ_{t-1}, u_{t-1}) and this function is the same for all t. To this end, let $\big(H_t^u(\omega), U_{t-1}(\omega)\big) = (h_{t-1}^u, u_{t-1})$. It follows from Eq. (3.18) on the evolution of Q_t and the conditional independence of S_{t-1} and U_{t-1} given H_{t-1}^u that

$$\begin{aligned} P_t(\omega)\big((S_{t-1}, S_t, Y_t, Q_t) \in A\big) = \int \cdots \int \mathbf{1}_A(s_{t-1}, s_t, y_t, q_t) P_{Q_t}\big((\xi_{t-1}, u_{t-1}, s_t, y_t), dq_t\big) \\ P_Y\big((s_t, u_{t-1}), dy_t\big) P_S\big((s_{t-1}, u_{t-1}), ds_t\big) \xi_{t-1}(ds_{t-1}) \end{aligned} \tag{3.22}$$

for all Borel sets A. Since by Eq. (3.17) the transition probability P_{Q_t} does not depend on time t, it can be seen from Eq. (3.22) that P_t is a function of (ξ_{t-1}, u_{t-1}) and this function is the same for all t. Hence by Eq. (3.20) ξ_t is a function of $(\xi_{t-1}, u_{t-1}, y_t, q_t)$ with the function being the same for all t. In other words, consider the joint process

$$\{S_0, \xi_0, U_0, S_1, Y_1, Q_1, \xi_1, U_1, \ldots\},$$

and there exists some transition probability P_ξ from $\mathcal{P}(\mathcal{S}) \times \mathcal{U} \times \mathcal{Y} \times \mathcal{M}$ to $\mathcal{P}(\mathcal{S})$ such that

$$\begin{aligned} \mathbb{P}^{\xi_0,\pi}\left(\xi_t \in \cdot \mid \mathcal{F}_t^u\right)(\omega) &= \mathbb{P}^{\xi_0,\pi}\big(\xi_t \in \cdot \mid \mathcal{F}_t^u \vee \sigma(\xi_0, \ldots, \xi_{t-1})\big)(\omega) && (3.23) \\ &= \mathbb{P}^{\xi_0,\pi}(\xi_t \in \cdot \mid \xi_{t-1}, U_{t-1}, Y_t, Q_t)(\omega) && (3.24) \\ &= P_\xi\big((\xi_{t-1}, u_{t-1}, y_t, q_t), \cdot\big). && (3.25) \end{aligned}$$

The proof is complete. □

Step 3: Same Evolution of the Belief Processes

We now show that the evolution model of ξ_t is the same as that of the belief process in the modified belief MDP. This is stated by the next lemma.

Lemma 3.5. *The following two relations hold:*

(i) Given $H_t^u(\omega) = (h_{t-1}^u, u_{t-1}, y_t, q_t)$,

$$\xi_t = \tilde{\phi}_{u_{t-1}}(\xi_{t-1}, (y_t, q_t)), \tag{3.26}$$

where $\tilde{\phi}$ is the function for the next belief as defined in the fictitious process.

(ii) Given $\big(H_{t-1}^u(\omega), U_{t-1}(\omega)\big) = (h_{t-1}^u, u_{t-1})$, the marginal distribution of (Y_t, Q_t) corresponding to $P_t(\omega)$ satisfies

$$P_t(\omega)((Y_t, Q_t) \in \cdot) = \widetilde{P}_0^{\xi_{t-1}, u_{t-1}}\big((\tilde{Y}_1, Q) \in \cdot\big). \tag{3.27}$$

Proof: Consider the fictitious process associated with the modified belief MDP. By the construction of the fictitious process, it is true that marginalized over Q, the marginal distribution of $(\tilde{S}_1, \tilde{Y}_1)$ is the same as that of (S_1, Y_1) of the original POMDP. In other words, denote by $\widetilde{P}_{s,y}^{\xi_{t-1}, u_{t-1}}$ the marginal distribution of $(\tilde{S}_1, \tilde{Y}_1)$ in the fictitious process with initial distribution ξ_{t-1} and initial control u_{t-1}, then

$$\begin{aligned}\widetilde{P}_{s,y}^{\xi_{t-1}, u_{t-1}}\left((\tilde{S}_1, \tilde{Y}_1) \in A\right) &= \widetilde{P}_0^{\xi_{t-1}, u_{t-1}}\left((\tilde{S}_1, \tilde{Y}_1) \in A\right) \\ &= \int\int\int \mathbf{1}_A(s_1, y_1) P_Y\big((s_1, u_{t-1}), dy_1\big) P_S\big((s_0, u_{t-1}), ds_1\big)\xi_{t-1}(ds_0)\end{aligned}$$

for all Borel sets A. Thus, denote by $\widetilde{P}_{q|s,y}^{\xi_{t-1}, u_{t-1}}\left(Q \in \cdot \mid \tilde{S}_1, \tilde{Y}_1\right)$ the conditional distribution of Q given $(\tilde{S}_1, \tilde{Y}_1)$ in the fictitious process, then the marginal distribution of $(\tilde{S}_1, \tilde{Y}_1, Q)$ can be expressed as

$$\begin{aligned}&\widetilde{P}_0^{\xi_{t-1}, u_{t-1}}\left((\tilde{S}_1, \tilde{Y}_1, Q) \in A\right) \\ &\quad = \int\int \mathbf{1}_A(s_1, y_1, q)\widetilde{P}_{q|s,y}^{\xi_{t-1}, u_{t-1}}\left(dq \mid \tilde{S}_1 = s_1, \tilde{Y}_1 = y_1\right)\widetilde{P}_{s,y}^{\xi_{t-1}, u_{t-1}}\left(d(s_1, y_1)\right) \\ &\quad = \int \cdots \int \mathbf{1}_A(s_1, y_1, q)\widetilde{P}_{q|s,y}^{\xi_{t-1}, u_{t-1}}\left(dq \mid \tilde{S}_1 = s_1, \tilde{Y}_1 = y_1\right) P_Y\big((s_1, u_{t-1}), dy_1\big) \\ &\qquad\qquad P_S\big((s_0, u_{t-1}), ds_1\big)\xi_{t-1}(ds_0).\end{aligned} \tag{3.28}$$

By Eq. (3.22) and Eq. (3.17) (the definition of P_{Q_t}), we have

$$\begin{aligned}&P_t(\omega)\big((S_t, Y_t, Q_t) \in A\big) \\ &\quad = \int \cdots \int \mathbf{1}_A(s_t, y_t, q_t) P_{Q_t}\big((\xi_{t-1}, u_{t-1}, s_t, y_t), dq_t\big) P_Y\big((s_t, u_{t-1}), dy_t\big) \\ &\qquad\qquad P_S\big((s_{t-1}, u_{t-1}), ds_t\big)\xi_{t-1}(ds_{t-1}) \\ &\quad = \int \cdots \int \mathbf{1}_A(s_t, y_t, q_t)\widetilde{P}_{q|s,y}^{\xi_{t-1}, u_{t-1}}\big(dq_t \mid \tilde{S}_1 = s_t, \tilde{Y}_1 = y_t\big) P_Y\big((s_t, u_{t-1}), dy_1\big) \\ &\qquad\qquad P_S\big((s_{t-1}, u_{t-1}), ds_t\big)\xi_{t-1}(ds_{t-1}).\end{aligned} \tag{3.29}$$

Comparing Eq. (3.28) and (3.29), we thus have

$$P_t(\omega)\big((S_t, Y_t, Q_t) \in \cdot\big) = \widetilde{P}_0^{\xi_{t-1}, u_{t-1}}\left((\tilde{S}_1, \tilde{Y}_1, Q) \in \cdot\right),$$

which implies that

$$P_t(\omega)\big((Y_t, Q_t) \in \cdot\big) = \widetilde{P}_0^{\xi_{t-1}, u_{t-1}}\Big((\tilde{Y}_1, Q) \in \cdot\Big),$$

$$P_t(\omega)\big(S_t \in \cdot \mid Y_t = y_t, Q_t = q_t\big) = \widetilde{P}_0^{\xi_{t-1}, u_{t-1}}\big(\tilde{S}_1 \in \cdot \mid \tilde{Y}_1 = y_t, Q = q_t\big),$$

and the second equation is equivalent to $\xi_t = \tilde{\phi}_{u_{t-1}}(\xi_{t-1}, (y_t, q_t))$ by the definitions of ξ_t and $\tilde{\phi}$. Thus Eq. (3.26) and (3.27) hold. □

Step 4: Sufficient Statistic for Control

Finally, we show that ξ_t is a sufficient statistic for control in the approximating POMDP. This implies that we can without loss of generality consider the set of Markov policies with respect to ξ_t in the approximating POMDP. Together with the analysis from the previous steps, this will then establish the equivalence of the approximating POMDP to the modified belief MDP.

Lemma 3.6. *In the approximating POMDP, given an initial distribution ξ_0, for any policy π, there is a policy $\pi_{\xi_0} = (\hat{\mu}_t)_{t \geq 0}$, with $\hat{\mu}_t$ functionally depending only on the belief ξ_t, that has the same expected cost as π for any bounded per-stage cost function g.*

Proof: Define a transition probability $\tilde{\mu}_t$ from $\mathcal{P}(\mathcal{S})$ to $\mathcal{U}$ by

$$\tilde{\mu}_t(\xi, \cdot) = \mathbb{P}^{\xi_0, \pi}(U_t \in \cdot \mid \xi_t)\big|_{\xi_t(\omega) = \xi}, \qquad \forall \xi \in \mathcal{P}(\mathcal{S}). \tag{3.30}$$

(If ξ is not realizable as ξ_t, $\tilde{\mu}_t(\xi, \cdot)$ can be defined arbitrarily, subject to measurability conditions.) Define a policy $\pi_{\xi_0} = (\hat{\mu}_t)_{t \geq 0}$ depending on ξ_0 by

$$\hat{\mu}_t(h_t^u, \cdot) = \tilde{\mu}_t\big(\xi_t(\xi_0, h_t^u), \cdot\big), \qquad \forall h_t^u \in \mathcal{H}_t^u. \tag{3.31}$$

We now show by induction that

$$\mathbb{P}^{\xi_0, \pi}(\xi_t, U_t) = \mathbb{P}^{\xi_0, \pi_{\xi_0}}(\xi_t, U_t), \qquad \forall\, t \geq 0. \tag{3.32}$$

Let ω and $\hat{\omega}$ denote a sample of the sample space of the stochastic process induced by π and π_{ξ_0}, respectively. By the definition of $\hat{\mu}_t$, Eq. (3.32) holds for $t = 0$. Assume that $\mathbb{P}^{\xi_0, \pi}(\xi_t, U_t) = \mathbb{P}^{\xi_0, \pi_{\xi_0}}(\xi_t, U_t)$ holds for t. By the definition of $\hat{\mu}_{t+1}$, to show that Eq. (3.32) holds for $t+1$, it is sufficient to show that for all (ξ, u),

$$\mathbb{P}^{\xi_0, \pi}(\xi_{t+1} \mid \xi_t, U_t)\big|_{\left(\xi_t(\omega), U_t(\omega)\right) = (\xi, u)} = \mathbb{P}^{\xi_0, \pi_{\xi_0}}(\xi_{t+1} \mid \xi_t, U_t)\big|_{\left(\xi_t(\hat{\omega}), U_t(\hat{\omega})\right) = (\xi, u)}. \tag{3.33}$$

This relation is evident from the evolution models of Y_{t+1} and Q_{t+1} in the approximating POMDP. Therefore Eq. (3.32) holds.

By Lemma 3.3, for any bounded per-stage cost function $g(s, u)$, the expected cost of a policy in the approximating POMDP is equal to its expected cost with respect to the per-stage cost function $\bar{g}(\xi, u)$. Therefore, Eq. (3.32) implies that

$$\hat{J}_k^{\pi}(\xi_0) = \hat{J}_k^{\pi_{\xi_0}}(\xi_0), \qquad \hat{J}_\beta^{\pi}(\xi_0) = \hat{J}_\beta^{\pi_{\xi_0}}(\xi_0), \qquad \forall\, k \geq 1,\ \beta \in [0, 1),$$

for any bounded per-stage cost function g. The proof is complete. □

3.8.3 Proof of Main Theorem

We can now prove Theorem 3.2. For convenience we restate the theorem.

Main Theorem (Theorem 3.2). *Given an initial distribution ξ_0, for any policy π of the original POMDP, there exists a policy $\tilde{\pi}$ of the modified belief MDP such that*

$$\tilde{J}_k^{\tilde{\pi}}(\xi_0) = J_k^{\pi}(\xi_0), \qquad \tilde{J}_\beta^{\tilde{\pi}}(\xi_0) = J_\beta^{\pi}(\xi_0), \quad \forall k \geq 1,\ \beta \in [0,1),$$

for any bounded per-stage cost function g.

Proof of Theorem 3.2: For a given ξ_0 and any policy π of the original POMDP, by Lemma 3.2 there exists a policy $\hat{\pi}$ of the approximating POMDP such that

$$\hat{J}_k^{\hat{\pi}}(\xi_0) = J_k^{\pi}(\xi_0), \qquad \hat{J}_\beta^{\hat{\pi}}(\xi_0) = J_\beta^{\pi}(\xi_0), \qquad k \geq 1,\ \beta \in [0,1), \tag{3.34}$$

for any bounded per-stage cost function g. This relation together with Lemma 3.6 implies that there exists a policy in the approximating POMDP, denoted here by $\hat{\pi}_{\xi_0}$, that functionally depends on ξ_t at each stage and has the same expected cost:

$$\hat{J}_k^{\hat{\pi}_{\xi_0}}(\xi_0) = J_k^{\pi}(\xi_0), \qquad \hat{J}_\beta^{\hat{\pi}_{\xi_0}}(\xi_0) = J_\beta^{\pi}(\xi_0), \qquad k \geq 1,\ \beta \in [0,1),$$

for any bounded per-stage cost function g. As can be seen in the proof of Lemma 3.6, the policy $\hat{\pi}_{\xi_0} = (\hat{\mu}_t)_{t\geq 0}$ can be viewed equivalently as a policy of the modified belief MDP, and we denote the latter by $\tilde{\pi}_{\xi_0}$. By Lemma 3.5, $\hat{\pi}_{\xi_0}$ and $\tilde{\pi}_{\xi_0}$ induce in the approximating POMDP a belief-control process $\{(\xi_t, U_t)_{t\geq 0}\}$ and in the modified belief MDP $\{(\xi_t, \tilde{U}_t)_{t\geq 0}\}$, respectively, with the same joint distribution (marginalized over the rest of random variables). Therefore,

$$\hat{J}_k^{\hat{\pi}_{\xi_0}}(\xi_0) = \tilde{J}_k^{\tilde{\pi}_{\xi_0}}(\xi_0), \qquad \hat{J}_\beta^{\hat{\pi}_{\xi_0}}(\xi_0) = \tilde{J}_\beta^{\tilde{\pi}_{\xi_0}}(\xi_0), \qquad k \geq 1,\ \beta \in [0,1), \tag{3.35}$$

for any bounded per-stage cost function g. Combining this with Eq. (3.34), we thus have

$$\tilde{J}_k^{\tilde{\pi}_{\xi_0}}(\xi_0) = J_k^{\pi}(\xi_0), \qquad \tilde{J}_\beta^{\tilde{\pi}_{\xi_0}}(\xi_0) = J_\beta^{\pi}(\xi_0), \qquad k \geq 1,\ \beta \in [0,1),$$

for any bounded per-stage cost function g. □

It can be seen that Theorem 3.1 is a corollary of the preceding main theorem applied to the average cost case.

Proof of Main Theorem for the Case of a Concave Per-Stage Cost Function

The statement of Theorem 3.2 holds with "$\leq$" replacing "$=$" for the more general case where the per-stage cost function $\bar{g}(\xi, u)$ is a concave function of beliefs for each value of u. The proof needs a slight modification, which we address in what follows. First, note that since the approximating POMDP is equivalent to the modified belief MDP, Eq. (3.35) in the preceding proof holds for the case of a concave per-stage cost function. The modification we need is to replace Lemma 3.2 with the following lemma, and consequently to have "$\leq$" replacing "=" in Eq. (3.34) in the preceding proof.

Lemma 3.2'. *Given an initial distribution ξ_0, for any policy π of the original POMDP, there exists a policy $\hat{\pi}$ of the approximating POMDP such that*

$$\hat{J}_k^{\hat{\pi}}(\xi_0) \leq J_k^{\pi}(\xi_0), \qquad \hat{J}_\beta^{\hat{\pi}}(\xi_0) \leq J_\beta^{\pi}(\xi_0), \qquad \forall k \geq 1,\ \beta \in [0,1),$$

for any bounded per-stage cost function $\bar{g}(\xi, u)$ that is concave in ξ for each value of u.

Proof: We specify some notation first. For the approximating POMDP and the original POMDP, denote by $\hat{\mathbb{P}}^{\xi_0,\hat{\pi}}$ and $\mathbb{P}^{\xi_0,\pi}$ the probability measures induced by policy $\hat{\pi}$ and π in the two processes, respectively, and denote by $\hat{\omega}$ and ω a sample of the two processes, respectively. Let ξ_t be the belief process as defined before for the approximating POMDP. Denote the belief process in the original POMDP by $\{\xi_t^o\}$. Correspondingly, also define $\mathcal{P}(\mathcal{S})$-valued random variables $\hat{\xi}_t^o$ in the approximating POMDP by

$$\hat{\xi}_t^o(\hat{\omega})(\cdot) = \hat{\mathbb{P}}^{\xi_0,\hat{\pi}}\left(S_t \in \cdot \mid U_0, (Y_k, U_k)_{k<t}, Y_t\right)(\hat{\omega}).$$

For a given policy π, we let $\hat{\pi}$ be the same policy π viewed as a policy of the approximating POMDP (i.e., $\hat{\pi}$ ignores Q_ts). Thus the marginal distributions of $(S_0, U_0, (S_t, Y_t, U_t)_{t>0})$ are the same in the approximating POMDP controlled by $\hat{\pi}$ and in the original POMDP controlled by π. This implies

$$\hat{\mathbb{P}}^{\xi_0,\hat{\pi}}\left((\hat{\xi}_t^o, U_t) \in \cdot\right) = \mathbb{P}^{\xi_0,\pi}\left((\xi_t^o, U_t) \in \cdot\right), \qquad t \geq 0$$

Consequently, by the concavity of $\bar{g}(\cdot, u)$ for all u and Jensen's inequality, for all $t \geq 0$,

$$\begin{aligned} E^{\hat{\mathbb{P}}^{\xi_0,\hat{\pi}}}\left\{\bar{g}(\xi_t, U_t)\right\} &= E^{\hat{\mathbb{P}}^{\xi_0,\hat{\pi}}}\left\{E^{\hat{\mathbb{P}}^{\xi_0,\hat{\pi}}}\left\{\bar{g}(\xi_t, U_t) \mid U_0, (Y_k, U_k)_{k<t}, Y_t\right\}\right\} \\ &\leq E^{\hat{\mathbb{P}}^{\xi_0,\hat{\pi}}}\left\{\bar{g}(\hat{\xi}_t^o, U_t)\right\} \\ &= E^{\mathbb{P}^{\xi_0,\pi}}\left\{\bar{g}(\xi_t^o, U_t)\right\}, \end{aligned}$$

where, to derive the second inequality, we have also used the conditional independence of U_t and ξ_t given $\{U_0, (Y_k, U_k)_{k<t}, Y_t\}$, (which is true due to our choice of $\hat{\pi}$). Hence $\hat{J}_\beta^{\hat{\pi}}(\xi_0) \leq J_\beta^{\pi}(\xi_0), \beta \in [0,1)$ and $\hat{J}_k^{\hat{\pi}}(\xi_0) \leq J_k^{\pi}(\xi_0), k \geq 1$ for any bounded per-stage cost function $\bar{g}(\xi, u)$ that is concave in ξ for each value of u. □

Thus we have the following theorem.

Theorem 3.2'. *Given an initial distribution ξ_0, for any policy π of the original POMDP, there exists a policy $\tilde{\pi}$ of the modified belief MDP such that*

$$\tilde{J}_k^{\tilde{\pi}}(\xi_0) \leq J_k^{\pi}(\xi_0), \qquad \tilde{J}_\beta^{\tilde{\pi}}(\xi_0) \leq J_\beta^{\pi}(\xi_0), \qquad \forall k \geq 1,\ \beta \in [0,1),$$

for any bounded per-stage cost function $\bar{g}(\xi, u)$ that is concave in ξ for each value of u.

3.9 Summary

In this chapter we have established essentially the lower approximating property for a class of processes as a whole, and characterized them in a unified way that links to the presence of hidden states in a POMDP. This theme is to be further carried out in more details or extended in the subsequent chapters.

Chapter 4

Discretized Lower Approximations for Discounted Cost Criterion

We now consider POMDP with *finite* state, observation and control spaces. Our interests are on those lower approximation schemes that can be computed exactly for finite space models. We will refer to them generally as discretized lower approximation schemes. We will consider computational issues, for discounted cost in this chapter, and for average cost in Chapter 5.

The discounted case is well-studied. This chapter is thus more for the preparation of the subsequent chapters, and for the sake of completeness. We will summarize discretized lower approximation schemes, which are applicable not only for the discounted but also the average cost and many other cases, as has been essentially established in Chapter 3. We will present, somewhat in detail, the algorithms and approximating functions, as well as asymptotic convergence analysis for the discounted case.

We shall simplify notation, since we will be dealing with DP equations instead of induced stochastic processes for most of the time. The distribution symbol $P_0^{\xi,u}$ in the expectation $E^{P_0^{\xi,u}}$ will be dropped. Instead we write $E_{V|\xi,u}$ for the conditional expectation over the random variable V with respect to the conditional distribution of V given the initial distribution ξ and control u. As to the symbol μ, instead of denoting a transition probability from the history set to the space of control, it will be used to denote a stationary and deterministic policy (with respect to the belief MDP), as a convention in DP problems.

4.1 Approximation Schemes

First we consider "grid"-based approximation schemes associated with the inequalities (3.14) and (3.10). Let $G = \{\xi_i\}$ be a finite set of beliefs such that their convex hull is $\mathcal{P}(\mathcal{S})$. A simple choice is to discretize $\mathcal{P}(\mathcal{S})$ into a regular grid, so we refer to ξ_i as *grid points*. By choosing different ξ_i and $\gamma_i(\cdot)$ in the inequalities (3.14) and (3.10), we obtain lower cost approximations that are functionally determined by their values at a finite number of beliefs.

Definition 4.1 (ϵ-Discretization Scheme). Call $(G, \underline{\gamma})$ an *ϵ-discretization scheme* where $G = \{\xi_i\}$ is a set of n beliefs, $\underline{\gamma} = (\gamma_1(\cdot), \ldots, \gamma_n(\cdot))$ is a convex representation scheme such that $\xi = \sum_i \gamma_i(\xi)\xi_i$ for all $\xi \in \mathcal{P}(\mathcal{S})$, and ϵ is a scalar characterizing the fineness of the discretization, and defined by

$$\epsilon = \max_{\xi \in \mathcal{P}(\mathcal{S})} \max_{\substack{\xi_i \in G \\ \gamma_i(\xi) > 0}} \|\xi - \xi_i\|.$$

Given (G, γ), let $\widetilde{\mathfrak{T}}_{D_i}, i = 1, 2$, be the mappings corresponding to the right-hand sides of inequalities (3.14) and (3.10), respectively:

$$(\widetilde{\mathfrak{T}}_{D_1} J)(\xi) = \min_{u \in \mathcal{U}} \Big[\bar{g}(\xi, u) + \beta \sum_i E_{Y|\xi,u}\big\{\gamma_i\big(\phi_u(\xi, Y)\big)\big\} J(\xi_i) \Big], \tag{4.1}$$

$$(\widetilde{\mathfrak{T}}_{D_2} J)(\xi) = \min_{u \in \mathcal{U}} \Big[\bar{g}(\xi, u) \; + \beta \sum_i \gamma_i(\xi) E_{Y|\xi_i,u}\big\{J\big(\phi_u(\xi_i, Y)\big)\big\} \Big]. \tag{4.2}$$

Associated with these mappings are the modified belief MDPs. The optimal cost functions $\tilde{J}_i$ in these modified problems satisfy, respectively,

$$(\widetilde{\mathfrak{T}}_{D_i} \tilde{J}_i)(\xi) = \tilde{J}_i(\xi) \leq J^*_\beta(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}), \; i = 1, 2.$$

The function $\tilde{J}_1$ was proposed by Zhou and Hansen [ZH01] as a grid-based approximation to improve the lower bound on the optimal cost function proposed by Lovejoy [Lov91]. Both $\tilde{J}_i$ are functionally determined by their values at a finite number of beliefs, which will be called *supporting points*, and whose set is denoted by $\mathcal{C}$. In particular, the function $\tilde{J}_1$ can be computed by solving a corresponding finite-state MDP on $\mathcal{C} = G = \{\xi_i\}$, and the function $\tilde{J}_2$ can be computed by solving a corresponding finite-state MDP on $\mathcal{C} = \{\phi_u(\xi_i, y) | \xi_i \in G, u \in \mathcal{U}, y \in \mathcal{Y}\}$.[1] The computation can thus be done efficiently by common algorithms for finite-state MDPs, e.g., variants of value iteration or policy iteration, and linear programming.

Usually $\mathcal{P}(\mathcal{S})$ is partitioned into convex regions and beliefs in a region are represented as the convex combinations of its vertices. The function $\tilde{J}_1$ is then piecewise linear on each region, and the function $\tilde{J}_2$ is piecewise linear and concave on each region. To see the latter, let $f(\xi_i, u) = E_{Y|\xi_i,u}\{\tilde{J}_2\big(\phi_u(\xi_i, Y)\big)\}$, and we have $\tilde{J}_2(\xi) = \min_u[\bar{g}(\xi, u) + \beta \sum_s \gamma_i(\xi) f(\xi_i, u)]$.

The simplest case for both mappings is when G consists of vertices of the belief simplex, i.e.,

$$G = \{e_s | s \in \mathcal{S}\}, \quad \text{where } e_s(s) = 1, \;\; e_s(s') = 0, \quad \forall s, s' \in \mathcal{S}, s \neq s'.$$

Denote the corresponding mappings by $\widetilde{\mathfrak{T}}_{D_i^0}, i = 1, 2$, respectively, i.e.,

$$(\widetilde{\mathfrak{T}}_{D_1^0} J)(\xi) = \min_{u \in \mathcal{U}} \Big[\bar{g}(\xi, u) + \beta \sum_s p(s|\xi, u) J(e_s) \Big], \tag{4.3}$$

$$(\widetilde{\mathfrak{T}}_{D_2^0} J)(\xi) = \min_{u \in \mathcal{U}} \Big[\bar{g}(\xi, u) + \beta \sum_s \xi(s) E_{Y|e_s,u}\big\{J\big(\phi_u(e_s, Y)\big)\big\} \Big]. \tag{4.4}$$

The mapping $\widetilde{\mathfrak{T}}_{D_1^0}$ gives the so called QMDP approximation suggested by Littman, Cassandra and Kaelbling [LCK95], and its corresponding function $\tilde{J}_1$ is simply

$$\tilde{J}_1(\xi) = \min_{u \in \mathcal{U}} \Big[\bar{g}(\xi, u) + \beta \sum_s p(s|\xi, u) J_\beta^{MDP}(s) \Big], \tag{4.5}$$

[1] More precisely, $\mathcal{C} = \{\phi_u(\xi_i, y) | \xi_i \in G, u \in \mathcal{U}, y \in \mathcal{Y}, \text{such that } p(Y_1 = y | \xi_i, u) > 0\}$.

where $J_\beta^{MDP}(s)$ is the optimal cost for the completely observable MDP. In the belief MDP associated with $\widetilde{\mathcal{T}}_{D_1^0}$, the states are observable after the initial step. In the belief MDP associated with $\widetilde{\mathcal{T}}_{D_2^0}$, the *previous* state is revealed at each step. Both approximating processes can be viewed as POMDPs, and derived using the information oracle argument.

Comparison of $\widetilde{\mathcal{T}}_{D_1}$ and $\widetilde{\mathcal{T}}_{D_2}$

First, we can apply Prop. 3.4 to compare the QMDP approximation $\widetilde{\mathcal{T}}_{D_1^0}$ with $\widetilde{\mathcal{T}}_{D_2^0}$. The operator $\widetilde{\mathcal{T}}_{D_1^0}$ corresponds to a lower approximation scheme with $Q = (Q_1, Q_2) = ((\tilde{S}_0, \tilde{Y}_1), \tilde{S}_1))$, while $\widetilde{\mathcal{T}}_{D_2^0}$ corresponds to $Q = Q_1$, or in other words, the "$\widetilde{\mathcal{T}}_1$" in Prop. 3.4. Both mappings preserve concavity, because the fictitious processes are POMDPs. Thus by Prop. 3.4 and Prop. 3.1 we have

$$J_\beta^*(\xi) \geq (\widetilde{\mathcal{T}}_{D_2^0} J_\beta^*)(\xi) \geq (\widetilde{T}_{D_1^0} J_\beta^*)(\xi), \quad J_\beta^*(\xi) \geq \tilde{J}_2(\xi) \geq \tilde{J}_1(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}), \tag{4.6}$$

where $\tilde{J}_i, i = 1, 2$ are the fixed points of $\widetilde{\mathcal{T}}_{D_i^0}$ respectively. By a similar argument for the finite-stage case, we have that $\widetilde{\mathcal{T}}_{D_2^0}$ gives a better *cost approximation* than $\widetilde{\mathcal{T}}_{D_1^0}$, under both discounted and average cost cases.

For the comparison of $\widetilde{\mathcal{T}}_{D_i}$ in general, one can show that by relaxing the inequality $J_\beta^* \geq \widetilde{\mathcal{T}}_{D_2} J_\beta^*$ using the concavity of J_β^*, we obtain an inequality of *the same form* as the inequality $J_\beta^* \geq \widetilde{T}_{D_1} J_\beta^*$. Note, however, this does not imply $(\widetilde{\mathcal{T}}_{D_2} J_\beta^*)(\xi) \geq (\widetilde{\mathcal{T}}_{D_1} J_\beta^*)(\xi), \forall \xi \in \mathcal{P}(\mathcal{S})$. We must also be aware that a better cost approximation does not imply a better greedy policy.

Other discretization schemes

Grid-based approximations are not the only schemes that have finite representations. One may combine the information oracle and the grid-based approximations to obtain other discretization schemes. For example, while a region-observable POMDP [ZL97] gives a continuous cost approximation, we can apply it to all beliefs in the relative interior of $\mathcal{P}(\mathcal{S})$, and apply a grid-based approximation to the beliefs on the boundary of $\mathcal{P}(\mathcal{S})$. Recall that the next belief in a region-observable POMDP is always on a lower dimensional affine space corresponding to some $\mathcal{S}_k$ in the region system (see Example 3.3). So in this combined scheme we only need to put grid points on lower dimensional spaces, which can be computationally appealing, (even though by Prop. 3.4 the resulting approximating function is a worse bound than that from the region-observable POMDP.)

By concatenating mappings we also obtain other discretized lower approximation schemes. For example,

$$\mathcal{T} \circ \widetilde{\mathcal{T}}_{D_i}, \ i = 1, 2; \text{ and } \widetilde{\mathcal{T}}_I \circ \widetilde{\mathcal{T}}_{D_2}, \tag{4.7}$$

where $\widetilde{\mathcal{T}}_I$ denotes the mapping in a region-observable POMDP. While concatenating two lower approximating mappings does not give a better approximation than concatenating $\mathcal{T}$ with one of them, the concatenated mapping $\widetilde{\mathcal{T}}_I \widetilde{\mathcal{T}}_{D_2}$ has the same property that we only need grid points ξ_i on some low dimensional affine spaces. Thus, the number of grid points needed for discretizing the entire belief space is traded with a more complex minimization problem as well as the approximation quality.

Discretized approximation schemes can also be obtained using a "pure" information oracle approach. The simplest schemes of this kind are $\widetilde{\mathcal{T}}_{D_1^0}$ and $\widetilde{\mathcal{T}}_{D_2^0}$, in which the states are revealed immediately or after one step. In order to obtain better approximations, one can design more complicated information patterns revealed by the oracle. For example, for some fixed k, assume that every state S_{ik} would be revealed at time $(i+1)k$; within every k stage interval, one can further make certain region-observable type of assumptions, to reduce computation when needed.

Mathematically, we can represent these combined or concatenated schemes using one single mapping $\widetilde{\mathcal{T}}$, by considering multiple stages as one single stage and enlarging the state, observation and control spaces accordingly. So there is no loss of generality in addressing all the discretized schemes, including the more complicated ones, simply by $\widetilde{\mathcal{T}}$.

4.2 Asymptotic Convergence

For discounted problems asymptotic convergence of cost approximations is naturally expected and a detailed analysis can be done in some standard way, (see e.g., [Ber75, Bon02]), to show the sufficient resolution in the discretization for obtaining an ϵ-optimal policy. We will leave these details out and give only a simple proof for completeness, using the uniform continuity property of $J^*_\beta(\cdot)$.

We first give some conventional notation related to policies, to be used throughout the text. Let μ be a stationary policy, and J_μ be its cost. Define the mapping $\mathcal{T}_\mu$ by

$$(\mathcal{T}_\mu J)(\xi) = \bar{g}(\xi, \mu(\xi)) + \beta\, E_{Y|\xi,\mu(\xi)}\{J\big(\phi_{\mu(\xi)}(\xi, Y)\big)\}.$$

Similarly, abusing notation, for any control u, we define $\mathcal{T}_u$ to be the mapping that has the same single control u in place of $\mu(\xi)$ in $\mathcal{T}_\mu$. Similarly, for the modified belief MDP, let $\widetilde{\mathcal{T}}_\mu$ and $\widetilde{\mathcal{T}}_u$ be the mappings correspond to a policy μ and control u, respectively.

The function $J^*_\beta(\xi)$ is continuous on $\mathcal{P}(\mathcal{S})$. For any continuous function $v(\cdot)$ on $\mathcal{P}(\mathcal{S})$, the function $E_{Y|\xi,u}\{v\big(\phi_u(\xi, Y)\big)\}$ as a function of ξ is also continuous on $\mathcal{P}(\mathcal{S})$. As $\mathcal{P}(\mathcal{S})$ is compact, by the uniform continuity of corresponding functions, we have the following simple lemma.

Lemma 4.1. *Let $v(\cdot)$ be a continuous function on $\mathcal{P}(\mathcal{S})$. Let $\beta \in [0,1]$. For any $\delta > 0$, there exists $\bar{\epsilon} > 0$ such that for any ϵ-discretization scheme $(G, \underline{\gamma})$ with $\epsilon \le \bar{\epsilon}$,*

$$|(\mathcal{T}_u v)(\xi) - (\widetilde{\mathcal{T}}_u v)(\xi)| \le \delta, \ \forall \xi \in \mathcal{P}(\mathcal{S}), \ \forall u \in \mathcal{U},$$

where $\widetilde{\mathcal{T}}$ is either $\widetilde{\mathcal{T}}_{D_1}$ or $\widetilde{T}_{D_2}$ associated with $(G, \underline{\gamma})$.

Proof: For $\widetilde{\mathcal{T}} = \widetilde{\mathcal{T}}_{D_1}$,

$$(\mathcal{T}_u v)(\xi) - (\widetilde{\mathcal{T}}_u v)(\xi) = \beta\, E_{Y|\xi,u}\Big\{v\big(\phi_u(\xi, Y)\big) - \sum_i \gamma_i\big(\phi_u(\xi, Y)\big)v(\xi_i)\Big\},$$

and the conclusion follows from the uniform continuity of $v(\cdot)$ over $\mathcal{P}(\mathcal{S})$. For $\widetilde{\mathcal{T}} = \widetilde{T}_{D_2}$,

$$(\mathcal{T}_u v)(\xi) - (\widetilde{\mathcal{T}}_u v)(\xi) = \beta E_{Y|\xi,u}\left\{v\big(\phi_u(\xi, Y)\big)\right\} - \beta \sum_i \gamma_i(x) E_{Y|\xi_i,u}\left\{v\big(\phi_u(\xi_i, Y)\big)\right\},$$

and the conclusion follows from the uniform continuity of $E_{Y|\cdot,u}\{v(\phi_u(\cdot, Y))\}$ over $\mathcal{P}(\mathcal{S})$. □

Lemma 4.1 implies that for any $\delta > 0$, there exists $\bar{\epsilon} > 0$ such that for any ϵ-discretization scheme with $\epsilon < \bar{\epsilon}$, the corresponding $\widetilde{\mathcal{T}}$ satisfies

$$(\widetilde{\mathcal{T}} J_\beta^*)(\xi) \leq J_\beta^*(\xi) \leq (\widetilde{\mathcal{T}} J_\beta^*)(\xi) + \delta, \ \forall \xi \in \mathcal{P}(\mathcal{S}). \tag{4.8}$$

Using this implication and the standard error bounds, one can show the following theorem which states that the lower approximation and the cost of its look-ahead policy, as well as the cost of the policy that is optimal with respect to the modified belief MDP, all converge to the optimal cost of the original POMDP.

Theorem 4.1. *Let $\widetilde{\mathcal{T}}_\epsilon$ be either $\widetilde{\mathcal{T}}_{D_1}$ or $\widetilde{\mathcal{T}}_{D_2}$ associated with an ϵ-discretization scheme. Define the function $\tilde{J}_\epsilon$, the policy μ_ϵ and $\tilde{\mu}_\epsilon$ to be such that*

$$\tilde{J}_\epsilon = \widetilde{\mathcal{T}}_\epsilon \tilde{J}_\epsilon, \qquad \mathcal{T}_{\mu_\epsilon} \tilde{J}_\epsilon = \mathcal{T} \tilde{J}_\epsilon, \qquad \widetilde{\mathcal{T}}_{\tilde{\mu}_\epsilon} \tilde{J}_\epsilon = \widetilde{\mathcal{T}}_\epsilon \tilde{J}_\epsilon.$$

Then

$$\tilde{J}_\epsilon \to J_\beta^*, \quad J_{\mu_\epsilon} \to J_\beta^*, \quad J_{\tilde{\mu}_\epsilon} \to J_\beta^*, \quad \textit{as } \epsilon \to \infty,$$

and the convergence is uniform over $\mathcal{P}(\mathcal{S})$.

Proof: Let $\delta > 0$. When ϵ is sufficiently small, by Eq. (4.8) we have $\|\widetilde{\mathcal{T}}_\epsilon J_\beta^* - J_\beta^*\|_\infty \leq \delta$. Viewing J_β^* as the approximate cost-to-go in the modified problem associated with $\widetilde{\mathcal{T}}_\epsilon$, we have by the standard MDP error bounds (see e.g., [Ber01]) that

$$\|J_\beta^* - \tilde{J}_\epsilon\|_\infty \leq \frac{\|\widetilde{\mathcal{T}}_\epsilon J_\beta^* - J_\beta^*\|_\infty}{1-\beta} = \frac{\delta}{1-\beta}.$$

This implies $\lim_{\epsilon \to 0} \|J_\beta^* - \tilde{J}_\epsilon\|_\infty = 0$, which in turn implies $\lim_{\epsilon \to 0} \|J_\beta^* - J_{\mu_\epsilon}\|_\infty = 0$, where μ_ϵ is the one-step look-ahead policy with respect to $\tilde{J}_\epsilon$. We now show $J_{\tilde{\mu}_\epsilon} \to J_\beta^*$, where $\tilde{\mu}_\epsilon$ is the optimal policy of the modified problem. We have $J_{\tilde{\mu}_\epsilon} \geq J_\beta^* \geq \tilde{J}_\epsilon$ and

$$\|J_\beta^* - J_{\tilde{\mu}_\epsilon}\|_\infty \leq \|\tilde{J}_\epsilon - J_{\tilde{\mu}_\epsilon}\|_\infty \leq \frac{\|\mathcal{T}_{\tilde{\mu}_\epsilon} \tilde{J}_\epsilon - \tilde{J}_\epsilon\|_\infty}{1-\beta},$$

where the second inequality follows from the standard MDP error bounds by viewing $\tilde{J}_\epsilon$ as the approximate cost-to-go in the original POMDP controlled by $\tilde{\mu}_\epsilon$. Using triangle inequality, the last term in the preceding equation can be further relaxed as

$$\|\mathcal{T}_{\tilde{\mu}_\epsilon} \tilde{J}_\epsilon - \tilde{J}_\epsilon\|_\infty \leq \|\mathcal{T}_{\tilde{\mu}_\epsilon} \tilde{J}_\epsilon - \mathcal{T}_{\tilde{\mu}_\epsilon} J_\beta^*\|_\infty + \|\mathcal{T}_{\tilde{\mu}_\epsilon} J_\beta^* - \widetilde{\mathcal{T}}_{\tilde{\mu}_\epsilon} J_\beta^*\|_\infty + \|\widetilde{\mathcal{T}}_{\tilde{\mu}_\epsilon} J_\beta^* - \tilde{J}_\epsilon\|_\infty \ .$$

As $\epsilon \to 0$, the second term on the right-hand side diminishes by Lemma 4.1, and the rest terms diminish due to the fact $\tilde{J}_\epsilon \to J_\beta^*$. Thus the claimed convergence is proved. Furthermore, since the convergence is in sup-norm, the convergence of the approximating functions is uniform in ξ. □

4.3 Summary

This chapter summarizes discretized lower approximation schemes and their computation issues for finite space POMDPs with the discounted cost criterion. The asymptotic convergence analysis is provided for completeness.

Chapter 5

Discretized Lower Approximations for Average Cost Criterion

In this chapter we will establish the application of discretized lower approximation schemes to finite space POMDPs with the average cost criterion. While for discounted problems several lower approximation schemes have been proposed earlier, ours seems the first of its kind for average cost POMDP problems. We show in Section 5.2 that the corresponding approximation can be computed efficiently using multichain algorithms for finite-state MDPs. We show in Section 5.3 that in addition to providing a lower bound to the optimal liminf average cost function, the approximation obtained from a discretized lower approximation scheme can also be used to calculate an upper bound on the limsup optimal average cost function, as well as bounds on the cost of executing the stationary policy associated with the approximation.

We prove in Section 5.4 the asymptotic convergence of the cost approximation, under the condition that the optimal average cost is constant and the optimal differential cost is continuous. We will also discuss the restriction of this condition and issues relating to the convergence of policies. We show in Section 5.5 experimental results of applying the discretized lower approximation approach to average cost problems.

5.1 Introduction

As reviewed in Section 2.4, exact solutions for POMDPs with the average cost criterion are substantially more difficult to analyze than those with the discounted cost criterion. Consider the belief MDP equivalent to the POMDP, and denote the belief at time 1 by $\tilde{X}$, a random variable. For a POMDP with average cost, in order that a stationary optimal policy exists, it is sufficient that the following functional equations, in the belief MDP notation,

$$J(\xi) = \min_u E_{\tilde{X}|\xi,u}\{J(\tilde{X})\}, \qquad U(\xi) = \operatorname*{argmin}_{u \in \mathcal{U}} E_{\tilde{X}|\xi,u}\{J(\tilde{X})\},$$

$$J(\xi) + h(\xi) = \min_{u \in U(\xi)} \left[\bar{g}(\xi, u) + E_{\tilde{X}|\xi,u}\{h(\tilde{X})\} \right], \tag{5.1}$$

admit a bounded solution $(J^*(\cdot), h^*(\cdot))$. The stationary policy that attains the minima of the right-hand sides of both equations is then average cost optimal with its average cost being $J^*(\xi)$. However, there are no finite computation algorithms to obtain it.

We now extend the application of the discretized approximations to the average cost case. First, note that solving the corresponding average cost problem in the discretized approach is much easier. Let $\widetilde{\mathcal{T}}$ be any of the mappings from discretized lower approximation schemes as listed in Section 4.1. For its associated modified belief MDP, we have the following average cost optimality equations:

$$J(\xi) = \min_u \tilde{E}_{\tilde{X}|\xi,u}\{J(\tilde{X})\}, \qquad U(\xi) = \operatorname*{argmin}_{u \in \mathcal{U}} \tilde{E}_{\tilde{X}|\xi,u}\{J(\tilde{X})\},$$
$$J(\xi) + h(\xi) = \min_{u \in U(\xi)} \left[\bar{g}(\xi,u) + \tilde{E}_{\tilde{X}|\xi,u}\{h(\tilde{X})\} \right], \tag{5.2}$$

and we have denoted by $\tilde{X}$ the belief at time 1, and used $\tilde{E}$ to indicate that the expectation is taken with respect to the distributions $\tilde{p}(\tilde{X}|\xi,u)$ of the modified MDP, which satisfy for all (ξ,u),

$$\tilde{p}(\tilde{X} = \tilde{\xi} \mid \xi, u) = 0, \quad \forall \tilde{\xi} \notin \mathcal{C},$$

with $\mathcal{C}$ being the finite set of supporting beliefs (see Section 4.1). There are bounded solutions $(\tilde{J}(\cdot), \tilde{h}(\cdot))$ to the optimality equations (5.2) for the following reason: Every finite-state MDP problem admits a solution to its average cost optimality equations. Furthermore if $\xi \notin \mathcal{C}$, ξ is transient and unreachable from $\mathcal{C}$, and the next belief $\tilde{X}$ belongs to $\mathcal{C}$ under any control u in the modified MDP. It follows that the optimality equations (5.2) restricted on $\{\xi\} \cup \mathcal{C}$ are the optimality equations for the finite-state MDP with $|\mathcal{C}| + 1$ states, so the solution $(\tilde{J}(\bar{\xi}), \tilde{h}(\bar{\xi}))$ exists for $\bar{\xi} \in \{\xi\} \cup \mathcal{C}$ with their values on $\mathcal{C}$ independent of ξ. This is essentially the algorithm to solve $\tilde{J}(\cdot)$ and $\tilde{h}(\cdot)$ in two stages, and obtain an optimal stationary policy for the modified MDP.

Concerns arise, however, about using any optimal policy for the modified MDP as suboptimal control in the original POMDP. Although all average cost optimal policies behave equally optimally in the asymptotic sense, they do so in the *modified MDP*, in which all the states $\xi \notin \mathcal{C}$ are transient. As an illustration, suppose for the completely observable MDP, the optimal average cost is constant over all states, then at any belief $\xi \notin \mathcal{C}$ any control will have the same asymptotic average cost in the modified MDP corresponding to the QMDP approximation scheme. The situation worsens, if even the completely observable MDP itself has a large number of states that are transient under its optimal policies. We therefore propose that for the modified MDP, we should aim to compute policies with additional optimality guarantees, relating to their finite-stage behaviors. Fortunately for finite-state MDPs, there are efficient algorithms for computing such policies. Furthermore, by adopting these algorithms, we automatically take care of the total cost case in which both the POMDP and the modified MDP have finite optimal total costs.

5.2 Algorithm

Review of n-discount optimality

We first briefly review related results, called sensitive optimality in literature, for *finite-state* MDPs. A reference can be found in Puterman [Put94].

Since average cost measures the asymptotic behavior of a policy, given two policies having the same average cost, one can incur significantly larger cost over a finite horizon than the other. The concept of n-discount optimality is useful for differentiating between such policies. It is also closely related to Blackwell optimality. Let J_β^π be the cost of policy

π in a β-discounted problem. A policy π^* is called *n-discount optimal* if its discounted costs satisfy

$$\limsup_{\beta \to 1} \, (1-\beta)^{-n} (J_\beta^{\pi^*}(s) - J_\beta^{\pi}(s)) \leq 0, \quad \forall s, \ \forall \pi.$$

By definition an $(n+1)$-discount policy is also k-discount optimal for $k = -1, 0, \ldots, n$. A policy is called *Blackwell optimal*, if it is optimal for all the discounted problems with discount factor $\beta \in [\bar{\beta}, 1)$ for some $\bar{\beta} < 1$. For finite-state MDPs, a policy is Blackwell optimal if and only if it is ∞-discount optimal. By contrast, any (-1)-discount optimal policy is average cost optimal.

For any finite-state MDP, there exist stationary average cost optimal policies and furthermore, stationary n-discount optimal and Blackwell optimal policies. In particular, denoting by $\tilde{S}$ the state at time 1, there exist functions $J(\cdot), h(\cdot)$ and $w_k(\cdot)$, $k = 0, \ldots, n+1$, with $w_0 = h$ such that they satisfy the following nested equations:

$$\begin{aligned} J(s) &= \min_{u \in U(s)} E_{\tilde{S}|s,u}\{J(\tilde{S})\}, \qquad (5.3) \\ J(s) + h(s) &= \min_{u \in U_{-1}(s)} [\, g(s,u) + E_{\tilde{S}|s,u}\{h(\tilde{S})\}], \\ w_{k-1}(s) + w_k(s) &= \min_{u \in U_{k-1}(s)} E_{\tilde{S}|s,u}\{w_k(\tilde{S})\}, \end{aligned}$$

where

$$\begin{aligned} U_{-1}(s) &= \operatorname*{arg\,min}_{u \in U(s)} E_{\tilde{S}|s,u}\{J(\tilde{S})\}, \\ U_0(s) &= \operatorname*{arg\,min}_{u \in U_{-1}(s)} [\, g(s,u) + E_{\tilde{S}|s,u}\{h(\tilde{S})\}], \\ U_k(s) &= \operatorname*{arg\,min}_{u \in U_{k-1}(s)} E_{\tilde{S}|s,u}\{w_k(\tilde{S})\}. \end{aligned}$$

Any stationary policy that attains the minimum in the right-hand sides of the equations in (5.3) is an n-discount optimal policy.

For finite-state MDPs, a stationary n-discount optimal policy not only exists, but can also be efficiently computed by multichain algorithms. Furthermore, in order to obtain a Blackwell optimal policy, which is ∞-discount optimal, it is sufficient to compute a $(N-2)$-discount optimal policy, where N is the number of states in the finite-state MDP. We refer readers to the book by Puterman [Put94], Chapter 10, especially Section 10.3 for details of the algorithm as well as theoretical analysis.

Algorithm

This leads to the following algorithm for computing an n-discount optimal policy for the *modified* MDP defined on the continuous belief space. We first solve the average cost problem on $\mathcal{C}$, then determine optimal controls on transient states $\xi \notin \mathcal{C}$. Note there are no conditions (such as unichain) on the recurrence structure of the modified belief MDP, which is a preferred property, since the modified process is, after all, an artificial process, that may lack nice properties usually found in real problems.

The algorithm solving the modified MDP

1. Compute an n-discount optimal solution for the finite-state MDP problem associated with $\mathcal{C}$. Let $\tilde{J}^*(\xi)$, $\tilde{h}(\xi)$, and $\tilde{w}_k(\xi)$, $k = 1, \ldots, n+1$, with $\xi \in \mathcal{C}$, be the corresponding functions obtained that satisfy Eq. (5.3) on $\mathcal{C}$.

2. For any belief $\xi \in \mathcal{P}(\mathcal{S})$, let the control set U_{n+1} be computed at the last step of the sequence of optimizations, (note $\tilde{X} \in \mathcal{C}$ with probability one):

$$\begin{aligned} U_{-1} &= \arg\min_{u} \tilde{E}_{\tilde{X}|\xi,u}\{\tilde{J}^*(\tilde{X})\}, \\ U_0 &= \arg\min_{u \in U_{-1}} \left[\bar{g}(\xi, u) + \tilde{E}_{\tilde{X}|\xi,u}\{\tilde{h}(\tilde{X})\} \right], \\ U_k &= \arg\min_{u \in U_{k-1}} \tilde{E}_{\tilde{X}|\xi,u}\{\tilde{w}_k(\tilde{X})\}, \; 1 \le k \le n+1. \end{aligned}$$

   Let u be any control in U_{n+1}, and let $\tilde{\mu}^*(x) = u$. Also if $\xi \notin \mathcal{C}$, define

$$\begin{aligned} \tilde{J}^*(\xi) &= \tilde{E}_{\tilde{X}|\xi,u}\{\tilde{J}^*(\tilde{X})\}, \\ \tilde{h}(\xi) &= \bar{g}(\xi, u) + \tilde{E}_{\tilde{X}|\xi,u}\{\tilde{h}(\tilde{X})\} - \tilde{J}^*(\xi). \end{aligned}$$

With the above algorithm we obtain an n-discount optimal policy for the modified MDP. When $n = |\mathcal{C}| - 1$, we obtain an ∞-discount optimal policy for the modified MDP,[1] since the algorithm essentially computes a Blackwell optimal policy for every finite-state MDP restricted on $\{\xi\} \cup \mathcal{C}$, for all ξ. Thus, for the *modified MDP*, for any other policy π, and any $\xi \in \mathcal{P}(\mathcal{S})$,

$$\limsup_{\beta \to 1} \, (1-\beta)^{-n} (\tilde{J}_\beta^{\tilde{\mu}^*}(\xi) - \tilde{J}_\beta^{\pi}(\xi)) \le 0, \quad \forall n \ge -1.$$

It is also straightforward to see that

$$\tilde{J}^*(\xi) = \lim_{\beta \to 1} \, (1-\beta) \tilde{J}_\beta^*(x), \quad \forall \xi \in \mathcal{P}(\mathcal{S}), \tag{5.4}$$

where $\tilde{J}_\beta^*(\xi)$ are the optimal discounted costs for the modified MDP, and the convergence is uniform over $\mathcal{P}(\mathcal{S})$, since $\tilde{J}_\beta^*(\xi)$ and $\tilde{J}^*(\xi)$ are piecewise linear interpolations of the function values on a finite set of beliefs.

5.3 Analysis of Error Bounds

We now show how to bound the optimal average cost of the original POMDP, and how to bound the cost of executing the suboptimal policy, that is optimal to the modified MDP, in the original POMDP.

[1] Note that ∞-discount optimality and Blackwell optimality are equivalent for finite-state MDPs, however, they are not equivalent in the case of a continuous state space. In the modified MDP, although for each ξ there exists an $\beta(\xi) \in (0, 1)$ such that $\tilde{\mu}^*(\xi)$ is optimal for all β-discounted problems with $\beta(\xi) \le \beta < 1$, we may have $\sup_\xi \beta(\xi) = 1$ due to the continuity of the belief space. In some literature, this is called weak Blackwell optimality, while the definition of Blackwell optimality we use in this thesis is called strong Balckwell optimality.

The fact that $\tilde{J}^*$ is a lower bound of J_-^* follows from the stronger lower bound result, Theorem 3.1 of Section 3.4. It can also be established by the weaker lower bound result, Prop. 3.3, using the fact that value iteration converges for finite state and control MDPs with average cost. We describe this line of analysis briefly as a comparison. Let $\tilde{V}_N^*(\xi)$ be the optimal N-stage cost function of the modified belief MDP. By Prop. 3.3, $\liminf_{N\to\infty} \frac{1}{N}\tilde{V}_N^*(\xi) \leq J_-^*(\xi)$. Recall in the modified problem for each ξ it is sufficient to consider the finite state MDP on $\{\xi\} \cup \mathcal{C}$. By Theorem 9.4.1. b of [Put94], the limit of $\frac{1}{N}\tilde{V}_N^*(\xi)$ exists and $\tilde{J}^*(\xi) = \lim_{N\to\infty}\frac{1}{N}\tilde{V}_N^*(\xi)$. Hence it follows that $\tilde{J}^*(\xi) \leq J_-^*(\xi)$.

Next we give a simple upper bound on $J_+^*(\cdot)$, which is an upper bound on the cost of a suboptimal policy, hence may be loose.

Theorem 5.1. *The optimal liminf and limsup average cost functions satisfy*

$$\tilde{J}^*(\xi) \leq J_-^*(\xi) \leq J_+^*(\xi) \leq \max_{\bar{\xi}\in\mathcal{C}} \tilde{J}^*(\bar{\xi}) + \delta,$$

$$\textit{where} \qquad \delta = \max_{\xi\in\mathcal{P}(\mathcal{S})} \left[(\mathcal{T}\tilde{h})(\xi) - \tilde{J}^*(\xi) - \tilde{h}(\xi)\right],$$

and $\tilde{J}^*(\xi)$, $\tilde{h}(\xi)$ and $\mathcal{C}$ are defined as in the modified MDP.

The upper bound is a consequence of the following lemma, the proof of which follows by bounding the expected cost per stage in the summation of the N-stage cost. This is a standard and useful bound for average cost MDPs, (not only for POMDPs).

Lemma 5.1. *Let $J(\xi)$ and $h(\xi)$ be any bounded functions on $\mathcal{P}(\mathcal{S})$, and μ be any stationary and deterministic policy. Define constants δ^+ and δ^- by*

$$\delta^+ = \max_{\xi\in\mathcal{P}(\mathcal{S})} \left[\bar{g}(\xi,\mu(\xi)) + E_{\tilde{X}|\xi,\mu(\xi)}\{h(\tilde{X})\} - J(\xi) - h(\xi)\right],$$

$$\delta^- = \min_{\xi\in\mathcal{P}(\mathcal{S})} \left[\bar{g}(\xi,\mu(\xi)) + E_{\tilde{X}|\xi,\mu(\xi)}\{h(\tilde{X})\} - J(\xi) - h(\xi)\right].$$

Then $V_N^\mu(\xi)$, the N-stage cost of executing policy μ, satisfies

$$\alpha^-(\xi) + \delta^- \leq \liminf_{N\to\infty} \tfrac{1}{N} V_N^\mu(\xi) \leq \limsup_{N\to\infty} \tfrac{1}{N} V_N^\mu(\xi) \leq \alpha^+(\xi) + \delta^+, \ \forall \xi \in \mathcal{P}(\mathcal{S}),$$

where $\alpha^+(\xi)$, $\alpha^-(\xi)$ are defined by

$$\alpha^+(\xi) = \max_{\bar{\xi}\in\mathcal{D}_\xi^\mu} J(\bar{\xi}), \quad \alpha^-(\xi) = \min_{\bar{\xi}\in\mathcal{D}_\xi^\mu} J(\bar{\xi}),$$

and $\mathcal{D}_\xi^\mu$ denotes the set of beliefs reachable under policy μ from ξ.

Proof: Let $\{\xi_t\}$ be the Markov process of beliefs in the original problem under the control of μ. It can be seen that

$$V_N^\mu(\xi) = E\{\sum_{t=0}^{N-1} \bar{g}(\xi_t, \mu(\xi_t))\}.$$

By definitions of δ^- and δ^+,

$$\bar{g}(\xi_t,\mu(\xi_t)) \leq J(\xi_t) + h(\xi_t) - E_{\xi_{t+1}|\xi_t,\mu(\xi_t)}\{h(\xi_{t+1})\} + \delta^+,$$

$$\bar{g}(\xi_t,\mu(\xi_t)) \geq J(\xi_t) + h(\xi_t) - E_{\xi_{t+1}|\xi_t,\mu(\xi_t)}\{h(\xi_{t+1})\} + \delta^-.$$

Summing up the inequalities over t, respectively, it follows that

$$\begin{aligned}E\{\sum_{t=0}^{N-1}\bar{g}(\xi_t,\mu(\xi_t))\} &\leq E\{\sum_{t=0}^{N-1}J(\xi_t)\}+h(\xi_0)-E\{h(\xi_N)\}+N\delta^+\\ &\leq N(\alpha^+(\xi_0)+\delta^+)+h(\xi_0)-E\{h(\xi_N)\},\\ E\{\sum_{t=0}^{N-1}\bar{g}(\xi_t,\mu(\xi_t))\} &\geq E\{\sum_{t=0}^{N-1}J(\xi_t)\}+h(\xi_0)-E\{h(\xi_N)\}+N\delta^-\\ &\geq N(\alpha^-(\xi_0)+\delta^-)+h(\xi_0)-E\{h(\xi_N)\}.\end{aligned}$$

Since h is a bounded function, we have

$$\alpha^-(\xi_0)+\delta^- \leq \liminf_{N\to\infty}\tfrac{1}{N}V_N^\mu(\xi_0)\leq \limsup_{N\to\infty}\tfrac{1}{N}V_N^\mu(\xi_0)\leq \alpha^+(\xi_0)+\delta^+,$$

and this proves the claim. □

Let $\tilde{\mu}^*$ be the stationary policy that is optimal for the modified MDP. We can use Lemma 5.1 to bound the liminf and limsup average cost of $\tilde{\mu}^*$ in the original POMDP. For example, consider the modified process corresponding to the QMDP approximation. If the optimal average cost J^*_{MDP} of the completely observable MDP problem equals the constant λ^* over all states, then we also have $\tilde{J}^*(\xi)=\lambda^*$, $\forall \xi \in \mathcal{P}(\mathcal{S})$. The cost of executing the policy $\tilde{\mu}^*$ in the original POMDP can therefore be bounded by

$$\lambda^*+\delta^- \leq \liminf_{N\to\infty}\tfrac{1}{N}V_N^{\tilde{\mu}^*}(\xi)\leq \limsup_{N\to\infty}\tfrac{1}{N}V_N^{\tilde{\mu}^*}(\xi)\ \leq\ \lambda^*+\delta^+.$$

The quantities δ^+ and δ^- can be hard to calculate exactly in general, since $\tilde{J}^*(\cdot)$ and $\tilde{h}(\cdot)$ obtained from the modified MDP are piecewise linear functions. Furthermore, the bounds may be loose, since they are worst-case bounds. On the other hand, these functions may indicate the structure of the original problem, and help us to refine the discretization scheme in the approximation.

5.4 Analysis of Asymptotic Convergence

In this section we will prove that as the resolution in discretization increases, the optimal cost of the modified problem asymptotically converges to the optimal cost function of the original POMDP, under the condition that there is a bounded solution (J^*,h^*) to the optimality equations (5.1) with J^* being constant and h^* continuous. We note that this is a fairly stringent condition. Because, first, it is easy to construct examples of POMDP with a non-constant optimal average cost (Section 2.4), and secondly, questions such as when h^* is continuous, how to verify it for a given problem, and several others, still lack satisfactory answers.

Let $(G,\underline{\gamma})$ be an ϵ-discretization scheme (Definition 4.1). Let $\tilde{J}_\epsilon$ and $\tilde{J}_{\beta,\epsilon}$ be the optimal average cost and discounted cost, respectively, in the modified MDP associated with $(G,\underline{\gamma})$ and either $\widetilde{\mathcal{T}}_{D_1}$ or $\widetilde{\mathcal{T}}_{D_2}$ as defined in Section 4.1. We address the question whether $\tilde{J}_\epsilon(\xi)\to J^*(\xi)$, as $\epsilon\to 0$, when $J^*(\xi)=J^*_-(\xi)=J^*_+(\xi)$ exists. Recall that in the discounted case for a fixed discount factor β, we have asymptotic convergence to optimality (Theorem 4.1):

$$\lim_{\epsilon \to 0} \tilde{J}_{\beta,\epsilon}(\xi) = J^*_\beta(\xi).$$

However, even under the conditions guaranteeing a constant optimal average cost $J^*(\xi) = \lambda^*$ and the convergence of discounted costs with vanishing discount factors, i.e.,

$$\lambda^* = \lim_{\beta \to 1} (1-\beta) J^*_\beta(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}),$$

in general we have

$$\lim_{\epsilon \to 0} \tilde{J}_\epsilon(\xi) = \lim_{\epsilon \to 0} \lim_{\beta \to 1} (1-\beta) \tilde{J}_{\beta,\epsilon}(\xi) \neq \lim_{\beta \to 1} \lim_{\epsilon \to 0} (1-\beta) \tilde{J}_{\beta,\epsilon}(\xi) = \lambda^*.$$

To ensure that $\tilde{J}_\epsilon \to \lambda^*$, we therefore assume stronger conditions than those that guarantee the existence of λ^*. We now show that a sufficient condition is the continuity of the optimal differential cost $h^*(\cdot)$.

Theorem 5.2. *Suppose the average cost optimality equations (5.1) admit a bounded solution $(J^*(\xi), h^*(\xi))$ with $J^*(\xi)$ equal to a constant λ^*. Then, if the differential cost $h^*(\xi)$ is continuous on $\mathcal{P}(\mathcal{S})$, we have*

$$\lim_{\epsilon \to 0} \tilde{J}_\epsilon(\xi) = \lambda^*, \quad \forall \xi \in \mathcal{P}(\mathcal{S}),$$

and the convergence is uniform, where $\tilde{J}_\epsilon$ is the optimal average cost function for the modified MDP corresponding to either $\widetilde{\mathcal{T}}_{D_1}$ or $\widetilde{\mathcal{T}}_{D_2}$ associated with an ϵ-discretization scheme $(G, \underline{\gamma})$.

Proof: Let $\tilde{\mu}^*_\epsilon$ be the optimal policy for the modified MDP associated with an ϵ-discretization scheme. Let $\widetilde{\mathcal{T}}$ be the mapping corresponding to the modified MDP, defined by $(\widetilde{\mathcal{T}} v)(\xi) = \min_{u \in \mathcal{U}} [\bar{g}(\xi, u) + \tilde{E}_{\tilde{X}|\xi,u}\{v(\tilde{X})\}]$. Since $h^*(\xi)$ is continuous on $\mathcal{P}(\mathcal{S})$, by Lemma 4.1 in Section 4.2, we have that for any $\delta > 0$, there exists $\bar{\epsilon} > 0$ such that for all ϵ-discretization schemes with $\epsilon < \bar{\epsilon}$,

$$|(\mathcal{T}_{\tilde{\mu}^*_\epsilon} h^*)(\xi) - (\widetilde{\mathcal{T}}_{\tilde{\mu}^*_\epsilon} h^*)(\xi)| \leq \delta, \quad \forall \xi \in \mathcal{P}(\mathcal{S}). \tag{5.5}$$

We now apply the result of Lemma 5.1 *in the modified MDP* with $J(\cdot) = \lambda^*, h = h^*$, and $\mu = \tilde{\mu}^*_\epsilon$. That is, by the same argument as in Lemma 5.1, the N-stage cost $\tilde{V}_N$ in the modified MDP satisfies

$$\tilde{J}_\epsilon(\xi) = \liminf_{N \to \infty} \tfrac{1}{N} \tilde{V}_N^{\tilde{\mu}^*_\epsilon}(\xi) \geq \lambda^* + \eta, \ \forall \xi \in \mathcal{P}(\mathcal{S}),$$

where $\eta = \min_{\xi \in \mathcal{P}(\mathcal{S})} \left[(\widetilde{\mathcal{T}}_{\tilde{\mu}^*_\epsilon} h^*)(\xi) - \lambda^* - h^*(\xi) \right]$. Since

$$\lambda^* + h^*(\xi) = (\mathcal{T} h^*)(\xi) \leq (\mathcal{T}_{\tilde{\mu}^*_\epsilon} h^*)(\xi),$$

and $|(\mathcal{T}_{\tilde{\mu}^*_\epsilon} h^*)(\xi) - (\widetilde{\mathcal{T}}_{\tilde{\mu}^*_\epsilon} h^*)(\xi)| \leq \delta$ by Eq. (5.5), we have

$$(\widetilde{\mathcal{T}}_{\tilde{\mu}^*_\epsilon} h^*)(\xi) - \lambda^* - h^*(\xi) \geq -\delta.$$

Hence $\eta \geq -\delta$, and $\tilde{J}_\epsilon(\xi) \geq \lambda^* - \delta$ for all $\epsilon \leq \bar{\epsilon}$, and $\xi \in \mathcal{P}(\mathcal{S})$, which proves the uniform convergence of $\tilde{J}_\epsilon$ to λ^*. □

Remark 5.1. The proof does not generalize to the case when $J^*(\xi)$ is not constant; and the inequality $\tilde{J}_\epsilon \leq J^*$ is crucial in the preceding proof. If a stronger condition is assumed that there exists a sequence of $\beta_k \uparrow 1$ such that the discounted optimal cost functions $\{J^*_{\beta_k} | k \geq 1\}$ are equicontinuous, then one can show the asymptotic convergence for other discretized approximations $\tilde{J}_\epsilon$ that are not necessarily lower bounds of J^*, using a vanishing discount argument as follows. For any $\delta > 0$, we can choose an ϵ-discretization scheme such that the β_k-discounted optimal cost $\tilde{J}^*_{\beta_k,\epsilon}$ of the modified problem satisfies

$$\|J^*_{\beta_k} - \tilde{J}^*_{\beta_k,\epsilon}\|_\infty \leq (1-\beta_k)^{-1}\delta/3, \quad \forall k.$$

Since

$$\lim_{k\to\infty}(1-\beta_k)\tilde{J}^*_{\beta_k,\epsilon} = \tilde{J}^*_\epsilon, \qquad \lim_{k\to\infty}(1-\beta_k)J^*_{\beta_k} = \lambda^*,$$

and the convergence in both equations is uniform, (the uniform convergence in the second equation is due to the equicontinuity of $\{J^*_{\beta_k}\}$), we can choose a $\bar{k}$ sufficiently large such that

$$\|\tilde{J}^*_\epsilon - \lambda^*\|_\infty \leq \|\tilde{J}^*_\epsilon - (1-\beta_{\bar{k}})\tilde{J}^*_{\beta_{\bar{k}},\epsilon}\|_\infty + (1-\beta_{\bar{k}})\|\tilde{J}^*_{\beta_{\bar{k}},\epsilon} - J^*_{\beta_{\bar{k}}}\|_\infty + \|(1-\beta_{\bar{k}})J^*_{\beta_{\bar{k}}} - \lambda^*\|_\infty \leq \delta,$$

where $\tilde{J}^*_\epsilon$ is the optimal average cost of the modified problem. This shows the asymptotic convergence of discretized approximation schemes that are not necessarily lower approximation schemes, when $\{J^*_\beta\}$ are equicontinuous.

Remark 5.2. Conditions in [HCA05] guarantee a bounded but not necessarily continuous h^*. Platzman [Pla80] shows an example in which the optimal differential cost has to be discontinuous. Platzman's reachability and detectability conditions [Pla80] guarantee a continuous h^*, but these conditions are quite strong and also not easily verifiable for a given problem.

Convergence Issues for Policies

An important issue that we have not addressed so far is how to obtain ϵ-optimal control policies, assuming that the optimal average cost is constant. A worst-case error bound analysis unfortunately fails to establish the asymptotic convergence of the cost of the suboptimal control policies obtained in solving the average cost modified problem.

A different approach is to use a near-optimal policy for the β-discounted problem with β sufficiently close to 1, as a suboptimal control policy for the average cost problem. This approach was taken by Runggaldier and Stettner [RS94]. We describe it here for completeness.

Assume that the constant average cost DP equation admits a bounded solution (λ^*, h^*). This assumption is equivalent to $\{h^*_\beta \mid \beta \in [0,1)\}$ being uniformly bounded (Theorem 7 of [HCA05]), where

$$h^*_\beta(\xi) = J^*_\beta(\xi) - J^*_\beta(\bar{\xi})$$

is the relative cost with respect to some fixed reference belief $\bar{\xi}$. Furthermore, this assumption implies $\lim_{\beta\uparrow 1}(1-\beta)J^*_\beta(\xi) = \lambda^*, \forall \xi \in \mathcal{P}(\mathcal{S})$.

Similar to the vanishing discount approach, first, by subtracting the term $\beta J^*_\beta(\bar{\xi})$ from both sides of the DP equation for the discounted problem, we can write the DP equation as

$$(1-\beta)J^*_\beta(\bar{\xi}) + h^*_\beta(\xi) = (\mathcal{T}_\beta h^*_\beta)(\xi),$$

where we have used $\mathcal{T}_\beta$ to denote the DP mapping for a β-discounted problem to its dependence on β explicit. Consider the policy μ such that

$$\|\mathcal{T}_\beta J^*_\beta - \mathcal{T}_{\beta,\mu} J^*_\beta\|_\infty \le \delta, \tag{5.6}$$

where $\mathcal{T}_{\beta,\mu}$ is the DP mapping associated with μ in the discounted problem. Then

$$(1-\beta)J^*_\beta(\bar{\xi}) + h^*_\beta(\xi) = (\mathcal{T}_\beta h^*_\beta)(\xi) \ge (\mathcal{T}_{\beta,\mu} h^*_\beta)(\xi) - \delta \ge (\mathcal{T}_\mu h^*_\beta)(\xi) - (1-\beta)\|h^*_\beta\|_\infty - \delta, \tag{5.7}$$

where $\mathcal{T}_\mu$ is the average cost DP mapping associated with μ. By Eq. (5.7) and Lemma 5.1, the limsup average cost $J^\mu_+(\xi)$ of μ is thus bounded by

$$J^\mu_+(\xi) \le (1-\beta)J^*_\beta(\bar{\xi}) + \delta + (1-\beta)\|h^*_\beta\|_\infty, \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

If β is sufficiently large so that

$$(1-\beta)\|h^*_\beta\|_\infty \le \delta, \qquad (1-\beta)J^*(\bar{\xi}) - \lambda^* \le \delta,$$

and furthermore μ satisfies Eq. (5.6), then μ is 3δ-optimal for the average cost problem.

One method to obtain a policy satisfying Eq. (5.6) is to compute the optimal policy of a discounted modified problem corresponding to an ϵ-discretization scheme with $\epsilon = O((1-\beta)\delta)$, roughly speaking. Another method is to use value iteration to compute an approximate cost-to-go function $\hat{J}_\beta$ such that $\|\mathcal{T}_\beta \hat{J}_\beta - \hat{J}_\beta\|_\infty = O((1-\beta)\delta)$ and then take the greedy policy with respect to $\hat{J}_\beta$. For β close to 1, both methods are of course computationally intensive.

5.5 Preliminary Experiments

We demonstrate our approach on a set of test problems: **Paint**, **Bridge-repair**, and **Shuttle**. The sizes of the problems are summarized in Table 5.1. Their descriptions and parameters are as specified in A. Cassandra's POMDP File Repository,[2] and we define costs to be negative rewards when a problem has a reward model.

Paint	Bridge	Shuttle
4×4×2	5×12×5	8×3×5

Table 5.1: Sizes of problems in terms of $|\mathcal{S}|\times|\mathcal{U}|\times|\mathcal{Y}|$.

We used some simple grid patterns. One pattern, referred to as **k-E**, consists of k grid points on each edge, in addition to the vertices of the belief simplex. Another pattern, referred to as **n-R**, consists of n randomly chosen grid points, in addition to the vertices of the simplex. The combined pattern is referred to as k-E+n-R. Thus the grid pattern for QMDP

[2] http://cs.brown.edu/research/ai/pomdp/examples

approximation is 0-E, for instance, and 2-E+10-R is a combined pattern. The grid pattern then induces a partition of the belief space and a convex representation (interpolation) scheme, which we kept implicitly and computed by linear programming on-line.

The algorithm for solving the modified finite-state MDP was implemented by solving a system of linear equations for each policy iteration. This may not be the most efficient way. No higher than 5-discount optimal policies were computed, when the number of supporting points became large.

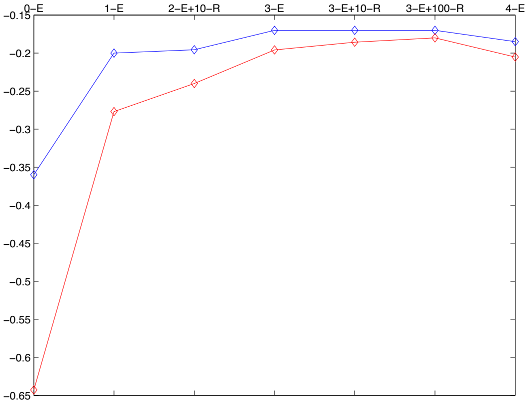

Figure 5-1: Average cost approximation for problem Paint using various grid patterns. The upper blue curve corresponds to $\widetilde{\mathcal{T}}_{D_2}$, and the lower red curve $\widetilde{\mathcal{T}}_{D_1}$. Linear interpolations between data points are plotted for reading convenience.

Fig. 5-1 shows the average cost approximation of $\widetilde{\mathcal{T}}_{D_1}$ and $\widetilde{\mathcal{T}}_{D_2}$ with a few grid patterns for the problem Paint. In all cases we obtained a constant average cost for the modified MDP. The horizontal axis is labeled by the grid pattern, and the vertical axis is the approximate cost. The red curve is obtained by $\widetilde{\mathcal{T}}_{D_1}$, and the blue curve $\widetilde{\mathcal{T}}_{D_2}$. As will be shown below, the approximation obtained by $\widetilde{\mathcal{T}}_{D_2}$ with 3-E is already near optimal. The policies generated by $\widetilde{\mathcal{T}}_{D_2}$ are not always better, however. We also notice, as indicated by the drop in the curves when using grid pattern 4-E, that the improvement of cost approximation does not solely depend on the number of grid points, but also on where they are positioned.

Problem	**LB**	**N. UB**	**S. Policy**
Paint	−0.170	-0.052	−0.172 ±0.002
Bridge	241.798	241.880	241.700 ±1.258
Shuttle	−1.842	−1.220	−1.835 ±0.007

Table 5.2: Average cost approximations and simulated average cost of policies.

In Table 5.2 we summarize the cost approximations obtained (column **LB**) and the simulated cost of the policies (column **S. Policy**) for the three problems. The approximation schemes that attained **LB** values in Table 5.2, as well as the policies simulated, are listed in Table 5.3. The column **N. UB** shows the numerically computed upper bound of the optimal – we calculate δ in Theorem 5.1 by sampling the values of $(\mathcal{T}\tilde{h})(\xi) - \tilde{h}(\xi) - \tilde{J}(\xi)$

Problem	**LB**	**S. Policy**
Paint	$\widetilde{\mathfrak{T}}_{D_2}$ w/ 3-E	$\widetilde{\mathfrak{T}}_{D_1}$ w/ 1-E
Bridge	$\widetilde{\mathfrak{T}}_{D_2}$ w/ 0-E	$\widetilde{\mathfrak{T}}_{D_2}$ w/ 0-E
Shuttle	$\widetilde{\mathfrak{T}}_{D_{1,2}}$ w/ 2-E	$\widetilde{\mathfrak{T}}_{D_1}$ w/ 2-E

Table 5.3: Approximation schemes in LB and simulated policies in Table 5.2.

at hundreds of beliefs generated randomly and taking the maximum over them. Thus the **N. UB** values are under-estimates of the exact upper bound. For both Paint and Shuttle the number of trajectories simulated is 160, and for Bridge 1000. Each trajectory has 500 steps starting from the same belief. The first number in **S. Policy** in Table 5.2 is the mean over the average cost of simulated trajectories, and the standard error listed as the second number is estimated from bootstrap samples – we created 100 pseudo-random samples by sampling from the empirical distribution of the original sample and computed the standard deviation of the mean estimator over these 100 pseudo-random samples.

As shown in Table 5.2, we find that some policy from the discretized approximation with very coarse grids can already be comparable to the optimal. This is verified by simulating the policy (**S. Policy**) and comparing its average cost against the lower bound of the optimal (**LB**), which in turn shows that the lower approximation is near optimal.

We find that in some cases the upper bounds may be too loose to be informative. For example, in the problem Paint we know that there is a simple policy achieving zero average cost, therefore a near-zero upper bound does not tell much about the optimal. In the experiments we also observe that an approximation scheme with more grid points does not necessarily provide a better upper bound of the optimal.

5.6 Summary

For average cost POMDP with finite space models, we have shown that lower bounds of the optimal liminf average cost function can be computed by using discretized lower approximation schemes and multichain n-discount MDP algorithms. Standard error bounds can be applied to the resulting approximating cost functions to bound the error of cost approximation as well as the cost of the suboptimal control obtained from the cost approximation. These results apply to POMDP problems in general, regardless of the equality of the optimal liminf and limsup cost functions, or the existence of solutions to the optimality equations.

We have also proved the asymptotic convergence of the average cost approximation under the restrictive condition of a constant optimal average cost and a continuous differential cost function. The question of the asymptotic convergence of the cost of the policies obtained from the approximation is, however, still open.

Chapter 6

Extension of Lower Approximations to Partially Observable Semi-Markov Decision Processes

Consider first continuous-time semi-Markov decision processes (SMDPs). An SMDP is similar to an MDP, except that there are random variables $\{\tau_n\}$, called *decision epochs*, which are the only times when controls can be applied. The time interval $\tau_{n+1} - \tau_n$, called the *sojourn time*, between transition from state S_n at time τ_n to state S_{n+1} at time τ_{n+1}, is random and depends on S_n, S_{n+1} as well as the control U_n.

A partially observable SMDP (POSMDP) is defined as an SMDP with hidden states and observations Y_n generated by (S_n, U_{n-1}). A graphical model of POSMDP is shown in Fig. 6-1.

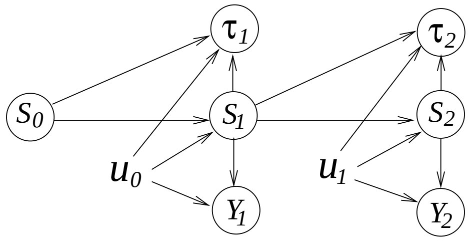

Figure 6-1: The graphical model of a partially observable semi-Markov decision process.

Though as a model, an SMDP is more general than an MDP, algorithmically many SMDP problems (e.g., policy iteration and value iteration) share similar structures with their MDP counterparts, and therefore can be solved by algorithms developed for MDPs after proper transformations of the SMDP problems. (For theories and algorithms of discrete space SMDPs one can see e.g., Puterman [Put94] and Sennott [Sen99].)

Similarly, our analyses from Chapter 3 to Chapter 5 for POMDPs also find their parallels in POSMDP problems, with, nevertheless, subtleties and differences in the latter worth to account for. In particular, we will show in this chapter the following results.

- The fictitious processes, inequalities for the optimal cost functions, lower cost approximations and modified belief MDPs proposed for POMDPs extend straightforwardly to the POSMDP case. Correspondingly, there are two lines of analysis that lead first to a weaker and then to a stronger lower bound result, and the latter states that the optimal cost of the modified problem is a lower bound of the optimal cost of the original problem (Theorem 6.1 and Corollary 6.1).

- For discounted or average cost POSMDPs with finite state, control and observation spaces, the lower bounds are computable using finite-state and control SMDP algorithms (Section 6.3).

- As an application of the POSMDP results, one can compute lower bounds of the optimal cost function of a POMDP problem over a subclass of policies that we refer to as hierarchical controllers, which induce a semi-Markov structure in the POMDP problem (Section 6.4).

Like the POMDP, the average cost POSMDP problem is still not well understood. The main contribution of this chapter is thus the proposal and analysis of lower bounds for the average cost case.

6.1 Review of POSMDPs

We will review the model and induced stochastic processes of a POSMDP, the belief SMDP equivalent to a POSMDP, expected cost criteria and optimality equations. We will follow the notation of Chapter 2 and simplify it in later sections where we consider discretized approximations.

6.1.1 Model and Assumptions

The model of a POSMDP can be specified by a seven-tuple $< \mathcal{S}, \mathcal{Y}, \mathcal{U}, P_{\tau,S}, P_Y, g, \alpha >$ with the terms defined as follows.

- $P_{\tau,S}$ is the state and sojourn time transition probability, and $P_{\tau,S}((s,u),A)$ denotes the conditional probability that $(s',\tau) \in A$, where s' is the next state and τ the sojourn time, given that at the current decision epoch the state is s and the control u.

- $P_Y((s,u),\cdot)$ is the observation probability, the same as in POMDP.

- $g(s,u)$ is the per-stage cost function, and $\alpha \in [0,\infty)$ the *discounting rate*. With a given α, $g(s,u)$ is defined as the sum of two parts:

$$g(s,u) = c_1(s,u) + E\left\{ \int_0^{\tau_1} e^{-\alpha t} c_2(W_t, S_0, U_0)\, dt \,\middle|\, S_0 = s, U_0 = u \right\} \qquad (6.1)$$

  where the first part of the cost $c_1(s,u)$ occurs immediately at the decision time, and the second part involving a "natural" process $\{W_t\}$ and a cost rate c_2 is the expected discounted cumulative costs during two consecutive decision epochs. Note that the per-stage cost $g(s,u)$ depends on the discounting rate α. We assume that $g(s,u)$ are given as part of the model.

The set of admissible policies are policies depending on past controls, decision epochs and observations. Let $\mathcal{H}_n$ be the space of $\{(U_{k-1}, \tau_k, Y_k)_{k \leq n}\}$, which is recursively defined as

$$\mathcal{H}_0 = \emptyset, \quad \mathcal{H}_n = \mathcal{H}_{n-1} \times \mathcal{U} \times \mathcal{R}_+ \times \mathcal{Y}.$$

A history dependent randomized policy π is a collection $(\mu_n)_{n \geq 0}$, where μ_n are transition probabilities from the history set $\mathcal{H}_n$ to the control space $\mathcal{U}$ that map an observed history up to the n-th decision epoch to a law on the control space. The set of all history dependent randomized policies are denoted by Π.

We need a few assumptions to ensure that expected costs are well-defined. Let $F(t|s,u)$ be the conditional cumulative distribution of the length between two consecutive decision epochs, i.e.,

$$F(\, t \mid s, u) = P_{\tau,S}\big((s,u), (-\infty, t] \times \mathcal{S}\big).$$

Ross [Ros70] introduced the following important assumption, which ensures that only finite number of decision epochs can happen during a finite period of time.

Assumption 6.1. *There exist $a > 0$ and $\delta \in (0,1)$ such that for all $s \in \mathcal{S}$, $u \in \mathcal{U}$,*

$$F(\, a \mid s, u) < \delta. \tag{6.2}$$

We will also impose the boundedness of the per-stage cost g.

Assumption 6.2. *There exists a constant L such that for $\alpha = 0$, $\sup_{s \in \mathcal{S}, u \in \mathcal{U}} |g(s,u)| < L$.*

The boundedness assumption is satisfied, if c_1, c_2 in Eq. (6.1) are bounded and $E\{\tau_1 \mid S_0 = s, U_0 = u\} < \infty$ for all s, u. Therefore we also make the following assumption.

Assumption 6.3. $\sup_{s \in \mathcal{S}, u \in \mathcal{U}} E\{\tau_1 \mid S_0 = s, U_0 = u\} < \infty$.

6.1.2 Discounted and Average Cost Criteria

Discounted Cost

Consider a discounted problem with discounting rate α. Let $\mathbb{P}^{\xi,\pi}$ be the joint distribution of the stochastic process $\{S_0, U_0, (S_n, \tau_n, Y_n, U_n)_{n \geq 1}\}$ induced by the initial distribution ξ and policy $\pi \in \Pi$. The discounted cost of π with discounting rate α is defined by

$$J_\alpha^\pi(\xi) = E^{\mathbb{P}^{\xi,\pi}} \left\{ \sum_{n=0}^{\infty} e^{-\alpha \tau_n} \, g(S_n, U_n) \right\},$$

and the optimal cost function is defined by

$$J_\alpha^*(\xi) = \inf_{\pi \in \Pi} J_\alpha^\pi(\xi).$$

Assumption 6.1 together with the boundedness of g ensures that for $\mathbb{P}^{\xi,\pi}$-almost all sample paths the infinite summation $\sum_{n=0}^{\infty} e^{-\alpha \tau_n} g(S_n, U_n)$ is well-defined and bounded. So J_α^π and J_α^* are well-defined.

The Equivalent Belief SMDP and Optimality Equations

We have for a POSMDP the equivalent belief SMDP and other definitions analogous to those in the POMDP case. For a given initial distribution ξ, define $\xi_0 = \xi$, and define the belief ξ_n, a $\mathcal{P}(\mathcal{S})$-valued random variable, to be the conditional distribution of S_n given the observed history up to the n-th decision epoch, i.e., $\{U_0, \tau_1, Y_1, U_1, \ldots, \tau_n, Y_n\}$, prior to control U_n being applied. Let $P_0^{\xi,u}$ be the marginal distribution of (S_0, S_1, τ_1, Y_1) when the initial distribution is ξ and initial control u. Define the function for the next belief $\phi_u(\xi, (\tau_1, Y_1))$ to be the conditional distribution of S_1 given (τ_1, Y_1):

$$\phi_u(\xi, (\tau, y))(\cdot) = P_0^{\xi,u}\left(S_1 \in \cdot \mid \tau_1 = \tau, Y_1 = y\right).$$

Following the same line of analysis for POMDPs (see e.g., [BS78]), one can show that the beliefs ξ_n are sufficient statistics for control, and the POSMDP can be viewed as a belief SMDP $\{\xi_0, U_0, (\xi_n, \tau_n, U_n)_{n \geq 1}\}$ with state space $\mathcal{P}(\mathcal{S})$ and per-stage cost function $\bar{g}(\xi, u)$ defined by

$$\bar{g}(\xi, u) = \int g(s, u)\, \xi(ds).$$

In this belief SMDP the transition probabilities for (τ_n, ξ_n), denoted by $P_{\tau,\xi}$, are defined as

$$\begin{aligned} P_{\tau,\xi}\Big((\xi_{n-1}, u_{n-1}), (\tau_n, \xi_n) \in A\Big) \; &= \iiint 1_A\Big(\tau_n, \phi_{u_{n-1}}\big(\xi_{n-1}, (\tau_n, y_n)\big)\Big) P_Y\Big((s_n, u_{n-1}), dy_n\Big) \\ &\qquad P_{\tau,S}\Big((s_{n-1}, u_{n-1}), d(\tau_n, s_n)\Big) \xi_{n-1}(ds_{n-1}), \end{aligned}$$

for all values of (ξ_{n-1}, u_{n-1}) and all Borel measurable subsets A of $\mathcal{R}_+ \times \mathcal{P}(\mathcal{S})$. Similar to the POMDP case, we can either view the belief SMDP as an SMDP by itself or view it as a process embedded in the joint process

$$\{S_0, \xi_0, U_0, \tau_1, S_1, Y_1, \xi_1, U_1, \ldots\}$$

of the POSMDP.

Furthermore, under proper measurability assumptions, there exists an optimal or ϵ-optimal policy that is deterministic and stationary with respect to the belief SMDP, for the discounted cost criterion. The optimal discounted cost satisfies the following optimality equation

$$J_\alpha^*(\xi) = \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + E^{P_0^{\xi,u}}\Big\{e^{-\alpha\tau_1}\, J_\alpha^*\Big(\phi_u\big(\xi, (\tau_1, Y_1)\big)\Big)\Big\}\right]. \tag{6.3}$$

There are analogous statements for the finite stage case (i.e., finite decision epochs), although the expected finite-stage cost may not be a natural cost criterion for certain problems, because the expected time length τ_N of N decision epochs for a fixed N depends on the policy.

Average Cost

There are two definitions of average cost for SMDP problems, which are equivalent for the SMDP case under certain conditions, but are not equivalent for the POSMDP case. We will describe both definitions in what follows.

The first definition of average cost, also the one we will use, is the limit of the expected cost up to time T divided by T, as T goes to infinity. Define the n_T-th decision epoch by

$$n_T = \max\{\, k \mid \tau_k \leq T \,\}.$$

For initial distribution ξ, the cost of policy π up to time T is defined by

$$\begin{aligned} J_C^\pi(\xi, T) &= E^{\mathbb{P}^{\xi,\pi}} \left\{ \sum_{n=0}^{n_T} \left(c_1(S_n, U_n) + \int_{\tau_n}^{\tau_{n+1} \wedge T} c_2(W_t, S_n, U_n)\, dt \right) \right\} \\ &= E^{\mathbb{P}^{\xi,\pi}} \left\{ \sum_{n=0}^{n_T} g(S_n, U_n) + \int_{\tau_{n_T}}^{T} c_2(W_t, S_{n_T}, U_{n_T})\, dt \right\}, \end{aligned} \tag{6.4}$$

where the subscript "C" stands for "continuous", and $a \wedge b = \min\{a, b\}$. The optimal liminf and limsup cost functions, denoted by J_{C-}^* and J_{C+}^*, respectively, are defined by

$$J_{C-}^*(\xi) = \inf_{\pi \in \Pi} \liminf_{T \to \infty} \frac{J_C^\pi(\xi, T)}{T}, \qquad J_{C+}^*(\xi) = \inf_{\pi \in \Pi} \limsup_{T \to \infty} \frac{J_C^\pi(\xi, T)}{T}. \tag{6.5}$$

For an SMDP there is also a second definition of average cost. Consider the expected cost up to the N-th decision epoch divided by the expected length of N decision epochs:

$$J^\pi(\xi, N) = \frac{E^{\mathbb{P}^{\xi,\pi}} \left\{ \sum_{n=0}^{N-1} g(S_n, U_n) \right\}}{E^{\mathbb{P}^{\xi,\pi}} \{\tau_N\}}.$$

Correspondingly, the optimal liminf and limsup cost functions are defined by

$$J_-^*(\xi) = \inf_{\pi \in \Pi} \liminf_{N \to \infty} J^\pi(\xi, N), \qquad J_+^*(\xi) = \inf_{\pi \in \Pi} \limsup_{N \to \infty} J^\pi(\xi, N).$$

For finite-state SMDPs the two definitions are equivalent for unichain models and stationary policies [Ros70]. This facilitates the computation of the average cost, because while the first definition does not give much structure to exploit, the second definition is easier for computing after relating the problem to the average cost MDP $\{(S_n, U_n)\}$. For multichain models, although the two definitions are not equal on transient states, the value of J_C^π at a transient state can be determined through the average cost optimality equation from the values of J_C^π at recurrent states, on which the two definitions agree. (For reference on multichain SMDP, see e.g., [Put94].)

For POSMDPs, because in general a policy does not induce an ergodic Markov chain, the two definitions of average cost are not equivalent. Furthermore, it is the first definition that is sensible. So **we will use the first definition of the average cost, Eq. (6.4)**.

Similar to the case of POMDPs, there are many open questions relating to the average cost POSMDP problem. For example, the important issue of when we have $J_{C-}^* = J_{C+}^*$ is not fully understood.

6.2 Lower Cost Approximations

6.2.1 Fictitious Processes, Inequalities and Lower Approximation Schemes

With a straightforward extension we can apply the various lower bound results of POMDPs to POSMDPs. This is done by viewing τ_n as part of the observation variable.[1]

[1] Even though τ_n depends on both S_{n-1} and S_n, and the observation Y_n in a POMDP depends only on S_n, this difference is not essential, because the Markov structure of the process is preserved. We could have defined the observation of a POMDP to depend on S_{n-1} as well, and all the analyses go through.

First, for a given initial distribution ξ, we can define a fictitious process

$$\{Q, \tilde{S}_0, \tilde{U}_0, \tilde{S}_1, (\tilde{\tau}_1, \tilde{Y}_1), \ldots\}$$

that satisfy Condition 3.1, with the pair of observation variables $(\tilde{\tau}_n, \tilde{Y}_n)$ replacing the single observation variable $\tilde{Y}_n$ in the POMDP case. For every policy π of the POSMDP, let $\widetilde{\mathbb{P}}^{\xi,\pi}$ be the induced joint distribution of the fictitious process. By the construction of the fictitious process, the marginal distribution of $(\tilde{S}_0, \tilde{U}_0)$ and the conditional distribution of $(\tilde{S}_1, \tilde{\tau}_1, \tilde{Y}_1, \tilde{U}_1 ...)$ given $\tilde{U}_0 = u$ are the same as those in the POSMDP (Lemma 3.1). So we have

$$J^\pi_\alpha(\xi) = E^{\widetilde{\mathbb{P}}^{\xi,\pi}} \left\{ \sum_{n=0}^{\infty} e^{-\alpha \tilde{\tau}_n} \, g(\tilde{S}_n, \tilde{U}_n) \right\}, \qquad \pi \in \Pi.$$

Consequently, by an argument identical to that in proving Prop. 3.1 for POMDPs, we have the inequality satisfied by the optimal discounted cost function:

$$J^*_\alpha(\xi) \geq \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, u) + E^{\widetilde{P}^{\xi,u}_0} \left\{ e^{-\alpha \tilde{\tau}_1} \, J^*_\alpha\Big(\tilde{\phi}_u(\xi, (Q, \tilde{\tau}_1, \tilde{Y}_1)) \Big) \right\} \right].$$

Here, $\widetilde{P}^{\xi,u}_0$ is the marginal distribution of $(Q, \tilde{S}_0, \tilde{S}_1, \tilde{\tau}_1, \tilde{Y}_1)$ when the initial distribution is ξ and control u, and $\tilde{\phi}_u(\xi, (Q, \tilde{\tau}_1, \tilde{Y}_1))$ is the conditional distribution of $\tilde{S}_1$ given $(Q, \tilde{\tau}_1, \tilde{Y}_1)$.

Similarly, we have the inequality for the finite-stage case (i.e., finite decision epochs), and can also show that these optimal cost functions are concave.

Lower Approximation Schemes and Modified Belief SMDP

We can analogously define lower approximation schemes $\{\widetilde{P}^{\xi,u}_0 \mid \xi \in \mathcal{P}(\mathcal{S}), u \in \mathcal{U}\}$ and the modified belief SMDP associated with the first stage model of the fictitious processes. The function for the next belief is now $\tilde{\phi}_u(\xi, (q, \tau, y))$, instead of $\tilde{\phi}_u(\xi, (q, y))$ in the POMDP case.

For later use, it will also be helpful to define observation variables $\{(Q_n, \tilde{Y}_n)_{n \geq 1}\}$ in the modified belief SMDP. They are defined to be generated by the belief state and the control. In particular, we define in the modified belief SMDP the transition probability $\widetilde{P}_{Q,\tilde{\tau},\tilde{Y}}$ for the random variables $(Q_n, \tilde{\tau}_n, \tilde{Y}_n)$ by

$$\widetilde{P}_{Q,\tilde{\tau},\tilde{Y}}\Big((\xi_{n-1}, u_{n-1}), (Q_n, \tilde{\tau}_n, \tilde{Y}_n) \in \cdot \Big) = \widetilde{P}^{\xi_{n-1}, u_{n-1}}_0 \big((Q, \tilde{\tau}_1, \tilde{Y}_1) \in \cdot \big).$$

With $(Q_n, \tilde{\tau}_n, \tilde{Y}_n) = (q_n, \tau_n, y_n)$, we then define the n-th belief state to be

$$\xi_n = \tilde{\phi}_{u_{n-1}}\big(\xi_{n-1}, (q_n, \tau_n, y_n) \big).$$

In other words, the transition probability $\widetilde{P}_\xi$ for the belief state can be expressed as

$$\widetilde{P}_\xi\Big((\xi_{n-1}, u_{n-1}, q_n, \tau_n, y_n), \xi_n \in A \Big) = \mathbf{1}_A\Big(\tilde{\phi}_{u_{n-1}}\big(\xi_{n-1}, (q_n, \tau_n, y_n) \big) \Big)$$

for all values of $(\xi_{n-1}, u_{n-1}, q_n, \tau_n, y_n)$ and all Borel measurable sets A of $\mathcal{P}(\mathcal{S})$.

6.2.2 Lower Bounds

Define the DP mapping of the modified belief SMDP to be $\widetilde{\mathcal{T}}$, and we can use the contraction and monotonicity properties of $\widetilde{\mathcal{T}}$ to derive lower bounds for the discounted and average cost problems. This is similar to the first line of analysis for the weaker lower bound result in the POMDP case. The conclusion of this approach applied to POSMDPs is, however, different from that in the POMDP case. We will comment on this issue later and also provide the corresponding analysis in Appendix A.1.

Our focus now is on a stronger lower bound result analogous to Theorem 3.2 in the POMDP case.

Theorem 6.1. *Given an initial distribution ξ_0, for every policy π of the POSMDP, there exists a policy $\tilde{\pi}$ of the modified belief SMDP such that*

$$\tilde{J}_\alpha^{\tilde{\pi}}(\xi_0) = J_\alpha^\pi(\xi_0), \qquad E^{\widetilde{\mathbb{P}}^{\xi_0,\tilde{\pi}}}\left\{\sum_{n=0}^{n_T} \bar{g}(\xi_n, \tilde{U}_n)\right\} = E^{\mathbb{P}^{\xi_0,\pi}}\left\{\sum_{n=0}^{n_T} g(S_n, U_n)\right\}, \quad \forall \alpha > 0,\ T \geq 0,$$

for any bounded per-stage cost function g.

Proof: We construct an approximating POSMDP in the same way as in the POMDP case (Section 3.8.1), treating (τ_n, Y_n) jointly as the observation variable. By this construction, for any given initial distribution ξ_0 and every policy π of the original POSMDP, there exists a policy $\hat{\pi}$ of the approximating POSMDP that has the same expected costs. Furthermore, we can choose $\hat{\pi}$ such that it depends only on the beliefs ξ_n in the approximating POSMDP. We denote by $\tilde{\pi}$ the same policy applied to the modified belief SMDP. It suffices to show that the joint distribution of the process $\{\xi_0, U_0, (\tau_1, Y_1), \xi_1, U_1, \ldots\}$ in the approximating POSMDP controlled by $\hat{\pi}$ is the same as that of $\{\xi_0, \tilde{U}_0, (\tilde{\tau}_1, \tilde{Y}_1), \xi_1, \tilde{U}_1, \ldots\}$ in the modified belief SMDP controlled by $\tilde{\pi}$. This is true by Lemma 3.5 and the fact that $\hat{\pi}$ depends functionally only on ξ_n. Thus, in particular, the marginal distribution of the process of beliefs, controls and sojourn times, $\{\xi_0, U_0, \tau_1, \xi_1, U_1, \tau_2, \ldots\}$ in the approximating POSMDP is the same as that of $\{\xi_0, \tilde{U}_0, \tilde{\tau}_1, \xi_1, \tilde{U}_1, \tilde{\tau}_2, \ldots\}$ in the modified belief SMDP. This implies that $\hat{\pi}$ and $\tilde{\pi}$ have the same expected cost, and consequently $\tilde{\pi}$ and π, the policy of the original POSMDP, have the same expected cost. □

Since the difference between the expected cost up to the random time τ_{n_T} and the expected cost up to time T is bounded, we have as an immediate consequence of the preceding theorem the following lower bound result for the average cost case.

Corollary 6.1. *Consider the modified belief SMDP associated with a lower cost approximation scheme. Then*

$$\tilde{J}_{C-}^*(\xi) \leq J_{C-}^*(\xi), \qquad \tilde{J}_{C+}^*(\xi) \leq J_{C+}^*(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Remark 6.1. The different line of analysis that is based on the inequalities for the optimal cost functions, leads to a lower bound result weaker than the preceding corollary. We include the analysis in Appendix A.1 for comparison of the two lines of analysis, as well as for ease of reference.

6.3 Discretized Lower Cost Approximations for Finite Space Models

Consider POSMDPs with finite state, control and observation spaces. The lower cost approximations for both discounted and average cost cases are computable by solving the modified belief SMDPs associated with those discretized approximation schemes. These modified belief SMDPs are essentially finite-state and control belief SMDPs, because there is a finite set $\mathcal{C}$ of supporting beliefs containing all the recurrent states. In particular, denote the next belief state by $\tilde{\xi}$ and the sojourn time by τ, and we can write the state and sojourn time transition probability as

$$\tilde{P}(\,\tau \leq t, \tilde{\xi} \in A \mid \xi, u) = \int_0^t \tilde{P}(\,\tilde{\xi} \in A \mid \xi, u, \tau)\, \tilde{P}(\,d\tau \mid \xi, u),$$

where the conditional distribution $\tilde{P}(\,\tau \leq t \mid \xi, u)$ of the sojourn time given (ξ, u) is the same as in the original POSMDP. The support of $\tilde{P}(\,\tilde{\xi} \mid \xi, u, \tau)$ is $\mathcal{C}$, i.e.,

$$\tilde{P}(\,\tilde{\xi} \in \mathcal{C} \mid \xi, u, \tau) = 1, \qquad \tilde{P}(\,\tilde{\xi} \in A \mid \xi, u, \tau) = 0, \quad \forall A, A \cap \mathcal{C} = \emptyset.$$

One example of such approximation schemes, when the sojourn time takes continuous values, is Zhou and Hansen's scheme [ZH01] (see Example 3.2 of Chapter 3). The set $\mathcal{C}$ is the set of grid points. If the sojourn time takes a finite number of values, then all discretized approximation schemes for POMDPs, when applied to the POSMDPs (view τ_n as additional observations), also have essentially finite space modified belief SMDPs.

To solve the modified belief SMDP problems, we can use the finite space SMDP algorithms. The algorithms are slightly different for the discounted case and for the average cost case, which we describe next.

6.3.1 Discounted Cost

For the discounted case, we mention a few issues in applying the finite space SMDP algorithms.

First, to solve the modified belief SMDP associated with a discretized lower approximation scheme, we do not need to compute certain probabilities explicitly, which can be especially hard to compute when the sojourn time takes continuous values. To illustrate this point, take Zhou and Hansen's approximation scheme for instance. The mapping $\widetilde{\mathcal{T}}$ is of the form

$$(\widetilde{\mathcal{T}}J)(\xi) = \inf_{u \in \mathcal{U}} \left[\bar{g}(\xi, s) + \sum_{k=1}^{m} E^{P_0^{\xi,u}} \left\{ e^{-\alpha\tau_1}\, \gamma_k\Big(\phi_u\big(\xi, (\tau_1, Y_1)\big)\Big) \right\} J(\xi_k) \right] \tag{6.6}$$

where $\{\xi_k\}$ is a set of m discretizing points, and $\gamma_k(\cdot)$ is the linear coefficients in the convex representation $\phi_u\big(\xi, (\tau_1, Y_1)\big) = \sum_k \gamma_k\Big(\phi_u\big(\xi, (\tau_1, Y_1)\big)\Big)\, \xi_k$. To solve the modified belief SMDP, we do not need to keep explicitly the probabilities $\phi_u(\xi, (\tau_1, Y_1))$ for all values of τ_1. Instead, we only need the values of the expectation terms

$$E^{P_0^{\xi,u}} \left\{ e^{-\alpha\tau_1}\, \gamma_k\Big(\phi_u\big(\xi, (\tau_1, Y_1)\big)\Big) \right\}, \quad k = 1, \ldots, m,$$

in the preceding equation. These can be computed by one-dimensional integration over τ_1 or by Monte carlo methods.

Secondly, in the special case where the length of interval τ_1 depends only on the control and not the state, the random discounting factor $e^{-\alpha\tau_1}$ in the optimality equation (6.3) can be replaced by its mean. The problem is then simplified to a POMDP with discounting factors $\beta(u)$ depending on controls:

$$\beta(u) = E^{P_0^{\xi,u}}\left\{e^{-\alpha\tau_1}\right\} = \int e^{-\alpha t} dF(\, t \mid u).$$

So despite that τ_n can take continuous values in this case, all the discretized lower approximation schemes for POMDPs can be applied.

Finally, with an identical argument as in Section 4.2 for POMDP, one can show the asymptotic convergence of the cost approximation and the cost of the greedy policies, using the fact that $\mathcal{T}, \widetilde{\mathcal{T}}$ are contraction mappings with respect to the sup-norm and with the contraction factor $\beta = \max_{s,u} E\{e^{-\alpha\tau_1} \mid S_0 = s, U_0 = u\} < 1$.

6.3.2 Average Cost

The approximation algorithm is similar to the average cost POMDP case. Because except for a finite set of beliefs, the rest belief states are transient, and from which we reach the set $\mathcal{C}$ in one step, it suffices to solve the finite-state belief SMDP and extends the solution to the entire belief space. It is also known from the the theory on finite-state and control SMDPs that the optimal average cost exists for every ξ, i.e., $\tilde{J}^*_{C-}(\xi) = \tilde{J}^*_{C+}(\xi) = \tilde{J}^*(\xi)$.

The multichain SMDP algorithm can be applied to compute the function $\tilde{J}^*(\xi)$ on $\mathcal{C}$. Indeed, the algorithm first transform the finite-state SMDP problem into an MDP using the so called “uniformization” transformation, and then solve it by a multichain algorithm for MDP. (For references on this transformation, see e.g., Chapter 11, Puterman [Put94].) The solution $(\tilde{J}^*, h)$ for the modified problem restricted on $\mathcal{C}$ satisfies the following optimality equation in the belief SMDP notation:

$$\begin{aligned} \tilde{J}^*(\xi) &= \min_{u\in\mathcal{U}} \tilde{E}_{\tilde{X}|\xi,u}\{\tilde{J}^*(\tilde{X})\}, \qquad U(\xi) = \operatorname*{argmin}_{u\in\mathcal{U}} \tilde{E}_{\tilde{X}|\xi,u}\{\tilde{J}^*(\tilde{X})\}, \\ \tilde{h}(\xi) &= \min_{u\in U(\xi)} \left[\bar{g}(\xi,u) - \bar{\tau}(\xi,u)\,\tilde{J}^*(\xi) + \tilde{E}_{\tilde{X}|\xi,u}\{\tilde{h}(\tilde{X})\}\right], \end{aligned} \tag{6.7}$$

where $\tilde{X}$ denotes the next belief, the distribution with respect to which the expectation $\tilde{E}_{\tilde{X}|\xi,u}$ is taken is the marginal state transition distribution $\tilde{P}(\tilde{\xi}|\xi,u)$ (marginalized over the sojourn time), and $\bar{\tau}(\xi,u)$ is the expected length between decision epochs, defined by

$$\tau(s,u) = E\{\tau_1 \mid S_0 = s, U_0 = u\}, \quad \bar{\tau}(\xi,u) = E\{\tau(S_0,u) \mid S_0 \sim \xi\},$$

(and note that $\bar{\tau}(\xi,u)$ is a linear function of ξ). Note that in computing the solution, we do not need explicitly the values of $\tilde{P}(\,\tau,\tilde{\xi} \mid \xi,u)$, which can be a fairly complicated distribution. Instead we only need the marginal state transition probabilities $\tilde{P}(\,\tilde{\xi} \mid \xi,u)$ and the expected sojourn time $\bar{\tau}(\xi,u)$ for $\xi,\tilde{\xi} \in \mathcal{C}$. These quantities should be computable with high precision in practice without too much difficulty.

The solution can then be extended to the entire belief space to obtain the optimal average cost function $\tilde{J}^*(\cdot)$ and a stationary optimal policy for the modified problem: For $\xi \notin \mathcal{C}$, compute $\tilde{J}^*(\xi)$ and then $\tilde{h}(\xi)$ by the above equations; the controls that attain the minima in the two nested equations above define a stationary optimal policy for the modified problem. The computation of this extension is more intensive than that in the case of POMDP. The overhead is mainly in computing for each ξ the marginal state transition probabilities $\tilde{P}(\tilde{\xi} \mid \xi, u)$, where $\tilde{\xi} \in \mathcal{C}$. Hence one may not be able to have an explicit representation of the function $\tilde{J}^*(\xi)$, but one can always have an implicit representation and compute $\tilde{J}^*(\xi)$ for any ξ of interest from the function $\tilde{J}^*$ restricted on $\mathcal{C}$.

Remark 6.2. The interpretation of the solution as a suboptimal policy for the original POSMDP need to be further studied, due to the concern of transience structure in the modified SMDP problem similar to the case of average cost POMDPs in Chapter 5. We may also solve the n-discount optimal problem for the transformed MDP corresponding to the modified SMDP problem. However, note that in this case the corresponding optimal policy is not n-discount optimal for that SMDP in the sense as n-discount optimality is defined.

6.4 Application to POMDPs with Hierarchical Control Strategies

We apply the lower bound results for POSMDPs to POMDP problems with certain structured policies that we call hierarchical control strategies, thus to obtain lower bounds of the cost function that is optimal over the subset of policies under consideration.

The Hierarchical Control Approach and its Motivations

Consider a POMDP with finite spaces. Suppose we are given a finite set of heuristic policies which we can apply one at a time and follow for a certain period before switch to another. Our goal is to find the best way to do so under either discounted or average cost criteria. This is a problem of a POMDP with a subset of structured policies. From the optimization point of view, one may hope that finding a good policy in a restricted policy space is more tractable than that in the entire policy space. There are motivations also in practice for considering this approach.

In certain large POMDP problems, such as navigation, while finding the exact solution is intractable, we have enough understanding of the structure of the problem to break a problem down to sub-problems. Subsequently, finding exact or approximate solutions to the sub-problems can be more tractable. The goal is then to design a high level controller which decides which sub-problem to invoke under various scenarios.

The approach of designing subproblems or hierarchical controllers which induce a semi-Markov structure, has been proposed and used for large scale MDPs (Sutton et al. [SPS99]) as well as POMDPs (Theocharous and Kaelbling [TK03]). The full scope of this approach and its applications is beyond the range of this section, and we will address only one specific type of hierarchical controllers.

Definitions of Controllers and Transformation to POSMDP

The controller has two levels, a high level controller which we intend to design, and base level heuristic policies which are given to us.

Formally, let us define a *heuristic control strategy* (abbreviated "heuristic") by $\{\pi = \big((\mu_t)_{t\geq 0}, (f_t)_{t\geq 1}\big)\}$ where μ_0 is a probability distribution on $\mathcal{U}$, $\mu_t, t > 0$ are transition probabilities from the history set $\mathcal{H}_t$ to $\mathcal{U}$, (recall the history set is the product space of the spaces of controls and observations), and f_t, a function from $\mathcal{H}_t$ to $\{0, 1\}$, defines the stopping condition: when $f_t(h_t) = 1$ the heuristic exits, and the control is given back to the high level controller.

The initial condition of a heuristic is as follows. If the high level controller calls a heuristic π at time t_0, then π is applied with its internal time index set to 0, that is, the heuristic does not depend on histories prior to t_0. Let $t_0 + \tau$ be the time the heuristic exits, and $S_{t_0+\tau}$ the state at that time. It then follows that

$$\mathbb{P}^{\pi}(\tau \leq t, S_{t_0+\tau} \mid (S_k, U_k, Y_k)_{k<t_0}, S_{t_0}) = \mathbb{P}^{\pi}(\tau \leq t, S_{t_0+\tau} \mid S_{t_0}). \tag{6.8}$$

We require that for all heuristic π, $E^{\pi}\{\tau \mid S_{t_0} = s\} < \infty, \forall s$, in order to satisfy the assumptions for SMDPs (Section 6.1).

At the exit time $t_0 + \tau$ of the heuristic, the high level controller receives the observation $\bar{Y}$, which can be either simply the sequence of the controls u_k and observations y_k during the period $[t_0, t_0 + \tau]$, or some message generated by the heuristic controller based on its observations during the same period. (We may restrict the range of τ or the length of the message so that the space of $\bar{Y}$ is finite.)

Denote by $\bar{u}$ a heuristic, and $\bar{\mathcal{U}}$ the finite set of such heuristic control strategies. Treat each $\bar{u}$ as a control and $\bar{\mathcal{U}}$ as the control space. Let $\{\tau_k\}$ be the exit times of the heuristics and let $\bar{S}_k = S_{\tau_k}$. At the decision epochs $\{\tau_k\}$ for the high level controller, we now have a POSMDP problem whose graphical model is as shown in Fig. 6-2. The model parameters of the POSMDP are as follows. The transition probabilities of $P_{\tau,\bar{S}}(\tau \leq t, s'|s, \bar{u})$ are defined by the right-hand side of Eq. (6.8); the per-stage cost of the POSMDP, denoted by g', is defined by

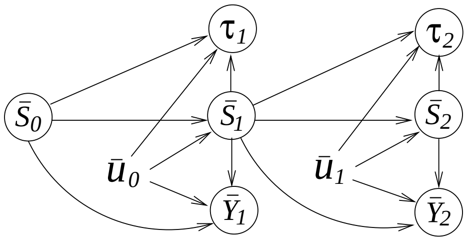

Figure 6-2: The graphical model of the POMDP with hierarchical controllers at the decision epochs $\{\tau_k\}$.

$$g'(s, \bar{u}) = E^{\pi}\{\sum_{t=0}^{\tau-1} \beta^t g(S_t, U_t) \mid S_0 = s\}, \qquad \beta \in [0, 1];$$

and the observation probabilities $P_{\bar{Y}}(\bar{Y}|s, s', \bar{u})$ are defined by the corresponding probabilities in the POMDP controlled by the heuristics. These parameters can be obtained from simulation, if we have the model of the POMDP. The parameters of the modified belief SMDP can also be obtained from simulation, as mentioned in the preceding sections.

Lower Bounds

The results established in the previous sections for POSMDPs then have the following implications. By solving the modified SMDP problem corresponding to a discretized lower approximation scheme, we obtain

- for average (discounted, respectively) cost criterion, a lower bound of the optimal liminf average cost function (the optimal discounted cost function, respectively) of the original POMDP problem over *the set of hierarchical control policies* that use the heuristics in the way described above, and
- a deterministic and stationary suboptimal policy that maps a belief to a heuristic control strategy.

We now discuss when this approach may be useful in practice. First, note that with the same discretized lower approximation scheme, the modified SMDP problem from the transformed POSMDP is more complex than the modified MDP problem from the original POMDP. So in order to gain from this approach, it has to be that in the POSMDP, coarse discretization can already yield good approximations, for otherwise, we may work directly with the POMDP without restricting the policy space. The second scenario when the approach can be useful is that some physical constraints require us to follow these heuristics for a certain period of time. In that case, the lower bounds obtained here are better than those obtained in the previous chapters, which are lower bounds on the optimal cost over all policies of the POMDP.

6.5 Summary

In this chapter we have extended the discretized lower approximation approach to the partially observable semi-Markov case with both the discounted and average cost criteria. We have given the algorithms and discussed their subtleties different to their counterparts for POMDPs. We have also shown that the POSMDP lower bound result allows one to compute for a POMDP lower bounds of the optimal cost function over a subset of hierarchical control policies that induce a semi-Markov structure in the original POMDP problem.

Chapter 7

Discretized Lower Approximations for Total Cost Criterion

In this chapter we consider the expected total cost criterion for finite space POMDPs. We show the application of our lower bound method and address the computational issues.

We first consider *non-negative* per-stage cost models and *non-positive* per-stage cost models. (In the reward maximization formulation, they are called negative models and positive models, respectively.) For these models the total costs, which may be infinite, are well defined.

Consider the modified problem associated with a discretized lower approximation scheme. By the stronger lower bound result, Theorem 3.2, the optimal total cost of the modified problem is a lower bound of the optimal total cost of the original POMDP. To obtain the optimal total cost of the modified problem, by the sensitive optimality theory for finite MDPs, we can solve the 0-discount optimality equation. If the solution $(\tilde{J}^*(\cdot), \tilde{h}^*(\cdot))$ satisfies $\tilde{J}^*(\xi) = 0$, then the optimal total cost of the modified problem with initial distribution ξ is bounded and equal to $\tilde{h}^*(\xi)$. (These will be proved rigorously later.) For non-negative cost models, under the assumption that the optimal total cost is finite, we also prove asymptotic convergence of the lower bounds.

We will address general per-stage cost models at the end. For that case we show that a zero optimal average cost of the modified problem yields a lower bound of the optimal limsup total cost of the original problem.

As a comparison of the stronger lower bound result and the weaker lower bound result that depends on the DP mapping, we will also sketch the line of analysis via the latter approach aiming for proving the same claims. This will show the technical limitation of the latter approach in various cases.

7.1 Non-Negative and Non-Positive Per-Stage Cost Models

7.1.1 Lower Bounds and Their Computation

Define $\tilde{V}^*(\xi)$ to be the optimal total cost of the modified problem with initial distribution ξ. To compute $\tilde{V}^*$, we use the average cost multichain policy iteration algorithm for the sensitive optimality criterion. Let $(\tilde{J}^*, \tilde{h}^*)$ satisfy the 0-discount optimality equations for the modified problem. As the following lemma indicates, if the optimal average cost $\tilde{J}^*$ is zero for initial distribution ξ, then $\tilde{h}^*(\xi) = \tilde{V}^*(\xi)$.

As for the proof of the lemma, if one assumes that the total cost of any policy in the modified problem is bounded, therefore $\tilde{J}^*$ is zero everywhere, then the proof can be found in Chapter 7 of [Put94]. Here, aiming for a pointwise lower bound on $V^*(\xi)$, we relaxed the finite total cost assumption for all policies. We provide a proof for completeness.

Lemma 7.1. *Assume either that the per-stage cost function is non-negative or that it is non-positive. For any $\xi \in \mathcal{P}(\mathcal{S})$, if $\tilde{J}^*(\xi) = 0$, then $\tilde{h}^*(\xi) = \tilde{V}^*(\xi)$.*

Proof: Recall that the modified problem is a belief MDP with its recurrent states contained in the finite set $\mathcal{C}$ of supporting beliefs, and that for every ξ it is sufficient to consider the finite state and control MDP on $\{\xi\} \cup \mathcal{C}$. For the latter MDP, by Theorem 7.1.9 of [Put94], there exists a total cost optimal policy that is stationary and deterministic. Denote this optimal policy by π_{tc}, and denote the 0-discount optimal policy by π.

First we show that π is total cost optimal. By the definition of 0-discount optimality,

$$\limsup_{\beta \uparrow 1} (1-\beta)^0 \big(\tilde{J}^\pi_\beta(\xi) - \tilde{J}^{\pi_{tc}}_\beta(\xi)\big) = \limsup_{\beta \uparrow 1} \big(\tilde{J}^\pi_\beta(\xi) - \tilde{J}^{\pi_{tc}}_\beta(\xi)\big) \leq 0. \tag{7.1}$$

By Lemma 7.1.8 of [Put94], the total cost of a policy equals the limit of its discounted costs when $\beta \uparrow 1$ (note that the limit can be infinite-valued):

$$\tilde{V}^\pi(\xi) = \lim_{\beta \uparrow 1} \tilde{J}^\pi_\beta(\xi), \qquad \tilde{V}^{\pi_{tc}}(\xi) = \lim_{\beta \uparrow 1} \tilde{J}^{\pi_{tc}}_\beta(\xi). \tag{7.2}$$

Thus Eq. (7.1) implies that

$$\tilde{V}^\pi(\xi) \leq \tilde{V}^{\pi_{tc}}(\xi) = \tilde{V}^*(\xi), \qquad \Rightarrow \quad \tilde{V}^\pi(\xi) = \tilde{V}^*(\xi). \tag{7.3}$$

Now suppose $\tilde{J}^*(\xi) = 0$ for a certain ξ. Since per-stage costs are bounded, by the result of Laurent series expansion (see Theorem 8.2.3 and Theorem 8.2.4 of [Put94]),

$$\lim_{\beta \uparrow 1} \tilde{J}^\pi_\beta(\xi) = \lim_{\beta \uparrow 1} (1-\beta)^{-1} \tilde{J}^*(\xi) + \tilde{h}(\xi) = \tilde{h}(\xi).$$

Thus, by the first equation in (7.2) and the second equation in (7.3), we conclude that $\tilde{h}(\xi) = \tilde{V}^*(\xi)$. □

Combining Lemma 7.1 with the stronger lower bound result, Theorem 3.2, we have the following theorem.

Theorem 7.1. *Assume either that the per-stage cost function is non-negative or that it is non-positive. Then*

$$\tilde{V}^*(\xi) \leq V^*(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Furthermore,

$$\tilde{V}^*(\xi) = \begin{cases} \tilde{h}^*(\xi), & \tilde{J}^*(\xi) = 0, \\ \infty, & \tilde{J}^*(\xi) > 0, \\ -\infty, & \tilde{J}^*(\xi) < 0. \end{cases} \tag{7.4}$$

Remark 7.1. As a consequence of the theorem, for non-positive per-stage cost models, since $\tilde{V}^*(\xi) \leq V^*(\xi) \leq 0$, we have a method to testify if the optimal total cost of the original POMDP is finite.

Remark 7.2. For non-negative and non-positive per-stage cost models, we note that one can prove the same claim using the weaker lower bound result based on the finite-stage inequalities. We sketch this line of analysis in what follows. Recall the finite-stage inequality from Section 3.4:

$$\tilde{V}_N^*(\xi) \le V_N^*(\xi), \qquad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

This implies

$$\lim_{N\to\infty} \tilde{V}_N^*(\xi) \le \lim_{N\to\infty} V_N^*(\xi) \le V^*(\xi).$$

Furthermore, it can be shown that for finite-stage and control MDPs, value iteration (with the initial cost function being 0) converges to the optimal total cost. Thus the preceding inequality and Lemma 7.1 imply that

$$\tilde{h}^*(\xi) = \tilde{V}^*(\xi) = \lim_{N\to\infty} \tilde{V}_N^*(\xi) \le V^*(\xi).$$

7.1.2 Asymptotic Convergence of Lower Bounds in Non-Negative Per-Stage Cost Models

Consider a POMDP with the per-stage cost $g(s,u) \ge 0$ for all s, u. Assume that the POMDP has the optimal total cost $V^*(\xi) < \infty$ for all $\xi \in \mathcal{P}(\mathcal{S})$. Then the optimal average cost of the POMDP is 0 for all initial distributions. Furthermore, we know the following facts:

1. The pair $(0, V^*)$, with 0 being the optimal average cost and V^* the optimal differential cost, is a bounded solution to the average cost optimality equation, i.e., the Bellman equation. (This is a standard result. See Theorem 7.1.3 of [Put94].)

2. The β-discounted optimal cost $J_\beta^* \uparrow V^*$ as $\beta \uparrow 1$. To see this, due to non-negativity of $g(s,u)$, the limiting function of the monotonically increasing sequence $J_{\beta_k}^*$, as $\beta_k \uparrow 1$, is positive and less than V^*, and satisfies the Bellman equation. Since V^* is the minimal positive solution to the Bellman equation (Theorem 7.3.2 of [Put94]), we conclude that this limiting function equals V^*.

3. In the modified problem, the optimal average cost $\tilde{J}^* = 0$, (since it is bounded both above and below by 0). The pair $(\tilde{J}^*, \tilde{h}^*) = (0, \tilde{V}^*)$ is the solution to the 0-discount optimality equation of the modified problem, as proved in the previous section.

Theorem 7.2. *Assume that the per-stage cost function is non-negative and the optimal total cost $V^* < \infty$. Let $\tilde{V}_\epsilon^*$ be the optimal total cost of a modified problem corresponding to either $\widetilde{\mathcal{T}}_{D_1}$ or $\widetilde{\mathcal{T}}_{D_2}$ of Section 4.1 associated with an ϵ-disretization scheme. Then*

$$\lim_{\epsilon\to 0} \tilde{V}_\epsilon^*(\xi) = V^*(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Proof: Prove by contradiction. Suppose the statement is not true. Then there exist $\delta > 0$, $\xi \in \mathcal{P}(\mathcal{S})$ and a sequence $\epsilon_k \to 0$ such that for all k sufficiently large

$$\tilde{V}_{\epsilon_k}^*(\xi) < V^*(\xi) - \delta.$$

Let J_β^* be the optimal β-discounted cost of the POMDP. Due to the second fact of the preceding discussion, for β sufficiently close to 1, $J_\beta^*(\xi) > V^*(\xi) - \frac{\delta}{2}$. Fix β. Thus for all k

sufficiently large, $\tilde{V}^*_{\epsilon_k}(\xi) < J^*_\beta(\xi) - \frac{\delta}{2}$. However, in the modified problem the β-discounted optimal cost $\tilde{J}^*_{\beta,\epsilon_k} \leq \tilde{V}^*_{\epsilon_k}$ due to non-negativity of per-stage costs. Hence for all k sufficiently large $\tilde{J}^*_{\beta,\epsilon_k}(\xi) < J^*_\beta(\xi) - \frac{\delta}{2}$. As $\epsilon_k \to 0$ when $k \to \infty$, this contradicts the asymptotic convergence of cost approximation for the discounted problems. □

Remark 7.3. The proof has used the fact that $\tilde{V}^* \leq V^*$. The finiteness assumption $V^* < \infty$ may be hard to verify. Note that when $\tilde{V}^* < \infty$, V^* can still take infinite value.

Remark 7.4. We comment on the issue of convergence of cost of policies.
(i) Like the average cost case, the above proposition shows the asymptotic convergence of *cost* approximations, but not that of the total cost of the policies obtained from the approximation schemes. The proposition does show, however, the asymptotic convergence to zero of the *average cost* of the policies, (and it also shows the asymptotic convergence of the differential cost function in the average cost context under the given conditions). As an example, consider a stochastic shortest path problem that has an absorbing destination state and a finite optimal total cost. The preceding theorem then implies that as the resolution of the discretization increases, the policies obtained from the corresponding modified problems have increasing probabilities of reaching the destination state.
(ii) Recall that in the MDP theory, for total cost criterion, there are no finite error bounds on the cost of a look-ahead policy, except when the DP mapping is a k-step contraction for some integer k. Thus in the case here it seems hard to us to prove the convergence of cost of policies without additional assumptions.

7.2 General Per-Stage Cost Models

For general per-stage cost models, the total cost may not be well defined. So we define the liminf and limsup total cost functions as in the average cost case:

$$V^*_-(\xi) = \inf_{\pi\in\Pi} \liminf_{N\to\infty} V^\pi_N(\xi), \qquad V^*_+(\xi) = \inf_{\pi\in\Pi} \limsup_{N\to\infty} V^\pi_N(\xi).$$

Consider the modified problem associated with a discretized lower approximation scheme. Since it is essentially a finite-state MDP, by the result of Denardo and Miller [DM68] (see also Chapter 10 of [Put94]), a stationary and deterministic 0-discount optimal policy, denoted by $\tilde{\pi}$, is average overtaking optimal, i.e., for all policy π of the modified problem,

$$\limsup_{N\to\infty} \frac{1}{N} \left(\sum_{k=1}^{N} \tilde{J}^{\tilde{\pi}}_k(\xi) - \sum_{k=1}^{N} \tilde{J}^{\pi}_k(\xi) \right) \leq 0, \qquad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Also by [DM68], if for some ξ_0, the 0-discount optimal solution $(\tilde{J}^*, \tilde{h}^*)$ satisfies $\tilde{J}^*(\xi_0) = 0$, then $\tilde{h}^*(\xi_0)$ is the Cesaro-limit of the finite-stage costs of $\tilde{\pi}$ for initial distribution ξ_0:

$$\tilde{h}^*(\xi_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \tilde{J}^{\tilde{\pi}}_k(\xi_0).$$

Since the Cesaro-limit is no greater than the limsup value, by the average overtaking optimality result of Denardo and Miller [DM68], we thus claim the following lower bound property for general per-stage cost models.

Theorem 7.3. *Consider a POMDP with a general per-stage cost model. If $\tilde{J}^*(\xi) = 0$, then the optimal limsup total cost $V_+^*(\xi)$ of the original POMDP satisfies*

$$\tilde{h}^*(\xi) \leq V_+^*(\xi).$$

Remark 7.5. For general per-stage cost model, the weaker lower bound result based on the finite-stage inequalities does not lead to the same claim. A technical difficulty there is that the finite-stage inequality result (Section 3.4) implies that for $J_0 = 0$,

$$\limsup_{k\to\infty} \widetilde{\mathcal{T}}^k J_0 \leq \limsup_{k\to\infty} \mathcal{T}^k J_0,$$

but in MDP with a general per-stage cost model, value iteration does not have to converge to the total cost. So it may happen that

$$\tilde{h}^* \geq \limsup_{k\to\infty} \widetilde{\mathcal{T}}^k J_0$$

in the modified problem.

7.3 Summary

We have shown that discretized lower approximations can be applied to compute lower bounds of the optimal total cost function for POMDPs with non-negative and non-positive per-stage cost models. For non-negative per-stage cost models, we have also shown the asymptotic convergence of cost approximations under a finite optimal total cost assumption on the original POMDP. We have also considered general per-stage cost models and shown that the discretized lower approximation schemes can be applied to compute lower bounds of the optimal limsup total cost of the original POMDP.

Chapter 8

Applications of Lower Bounds

In this chapter we show several applications, for which the use of the discretized lower cost approximations is not so much in providing suboptimal control policies, as it is in providing lower bounds and approximations to quantities of interest.

In Section 8.1 we consider problems of reaching, avoidance and model identification. These problems can be cast as POMDP problems, and lower bounds of the optimal average cost or total cost function can be used to bound probabilities or expected values of interest.

In Section 8.2 we consider the entropy rate of hidden Markov sources, a classic problem in information theory. We show that the discretized lower approximation provides a new and more efficient method of computing lower bounds of the entropy rate. Asymptotic convergence issues will be addressed and proved under certain conditions. Applications of the same type include sequential coding, which we briefly address in the summary.

8.1 Bounding Probabilities of Success in Reaching, Avoidance and Identification

We consider problems of reaching, avoidance, and model identification, and apply results from Chapters 5 and 7 to bound the probabilities of success, or certain expected values of interest. The bounds are mostly useful, in our opinion, for quantitative assessment of the risk, or the possibility of success, in accomplishing the corresponding control task. Especially, the bounds are informative when their indications are negative, for instance, when they indicate a high probability of failure in reaching or identification in a given problem.

Example 8.1 (Reaching). Consider a POMDP problem in which the goal is to reach a certain set $\mathcal{S}'$ of destination states. We assume $\mathcal{S}'$ is absorbing, but we do not assume the rest of the states are non-absorbing – there can be other absorbing states, which will be undesirable destinations. We can upper-bound the maximal probability of reaching and lower-bound the minimal expected reaching time of this problem as follows.

Define the per-stage cost function of this POMDP simply as,

$$g(s,u) = 0, \quad \forall s \in \mathcal{S}', \qquad g(s,u) = 1, \quad \forall s \in \mathcal{S}\backslash\mathcal{S}', \quad \forall u \in \mathcal{U}.$$

One can show that the optimal liminf and limsup average cost functions are equal, and furthermore, the optimal average cost function, denoted by $J^*(\xi)$, is the minimal probability of starting from the distribution ξ and never reaching the destination set $\mathcal{S}'$,[1] i.e.,

$$J^*(\xi) = \inf_{\pi \in \Pi} \mathbb{P}^\pi(\text{never reach } \mathcal{S}' \mid S_0 \sim \xi). \tag{8.1}$$

Choose a lower discretization scheme and compute the optimal average cost function $\tilde{J}^*$ for the corresponded modified problem. Then starting from a distribution ξ, the probability that $\mathcal{S}'$ is not reached is at least $\tilde{J}^*(\xi)$ under any policy, i.e.,

$$\mathbb{P}(\mathcal{S}' \text{ never reached} \mid S_0 \sim \xi) \geq \tilde{J}^*(\xi), \tag{8.2}$$

and the inequality is non-trivial when $\tilde{J}^*(\xi) \neq 0$.

If $\tilde{J}^*(\xi) = 0$, then we can bound the expected reaching time. (Note that this will be the case, if $\mathcal{S}'$ is reachable with probability 1 under any policy in the original problem.) The optimal total cost V^* of this problem is the minimal expected reaching time over all policies,

$$V^*(\xi) = \inf_{\pi \in \Pi} E^\pi\{\tau \mid S_0 \sim \xi\}$$

where $\tau = \min\{k \,|\, S_k \in \mathcal{S}'\}$ is the hitting time. Let $(\tilde{J}^*, \tilde{h}^*)$ be the solution pair to the 0-discount optimality equations of the modified problem. Since the per-stage cost is non-negative, by Theorem 7.1 of Section 7.1, $\tilde{h}^*$ is the optimal total cost function of the modified problem, and $\tilde{h}^*(\xi) \leq V^*(\xi)$. Thus,

$$\inf_{\pi \in \Pi} E^\pi\{\tau\} \geq \tilde{h}^*(\xi). \tag{8.3}$$

[1] To see this, for any initial distribution ξ, define the hitting time

$$\tau \overset{def}{=} \min\{k \mid S_k \in \mathcal{S}'\}.$$

Since $\mathcal{S}'$ is absorbing, from time 0 to time $N-1$, the N-stage cost V_N^π of any policy π satisfies the following relation, (which uses a standard trick in expressing expected value):

$$\begin{aligned} V_N^\pi = E^\pi\{\min\{\tau, N\}\} &= \sum_{k=0}^{N-1} k\, \mathbb{P}^\pi(\tau = k) + N\, \mathbb{P}^\pi(\tau \geq N) \\ &= \sum_{k=1}^{N-1} \sum_{j=1}^{k} \mathbb{P}^\pi(\tau = k) + N\mathbb{P}^\pi(\tau \geq N) = \sum_{j=1}^{N-1} \sum_{k=j}^{N-1} \mathbb{P}^\pi(\tau = k) + N\mathbb{P}^\pi(\tau \geq N) \\ &= \sum_{j=1}^{N} \mathbb{P}^\pi(\tau \geq j). \end{aligned}$$

Since $\mathbb{P}^\pi(\tau \geq j) \geq \mathbb{P}^\pi(\tau = \infty)$, it follows that

$$V_N^\pi \geq N\mathbb{P}^\pi(\tau = \infty) \qquad \Rightarrow \qquad \liminf_{N\to\infty} \frac{V_N^\pi}{N} \geq \mathbb{P}^\pi(\tau = \infty);$$

and since for any $M < N$, $\mathbb{P}^\pi(\tau \geq M) \geq \mathbb{P}^\pi(\tau \geq N)$, it follows that

$$\begin{aligned} V_N^\pi \leq M + (N-M)\, \mathbb{P}^\pi(\tau \geq M) \qquad &\Rightarrow \qquad \limsup_{N\to\infty} \frac{V_N^\pi}{N} \leq \mathbb{P}^\pi(\tau \geq M), \quad \forall M, \\ &\Rightarrow \qquad \limsup_{N\to\infty} \frac{V_N^\pi}{N} \leq \mathbb{P}^\pi(\tau = \infty), \end{aligned}$$

where the last inequality follows by letting $M \to \infty$. Thus $\lim_{N\to\infty} \frac{V_N^\pi}{N} = \mathbb{P}^\pi(\tau = \infty)$, so both the liminf and limsup average cost of π are equal to $\mathbb{P}^\pi(\tau = \infty)$. Since this holds for all policies π, the optimal liminf and limsup cost functions are also equal and the optimal cost satisfies Eq. (8.1).

In summary, if $\tilde{J}^*(\xi)$ is zero, then starting from the distribution ξ, the minimal expected time of reaching the set $\mathcal{S}'$ is no less than $\tilde{h}^*(\xi)$, (note that this does not imply a finite expected reaching time). □

Example 8.2 (Avoidance). Consider the opposite of the reaching problem: there is a absorbing set $\mathcal{S}'$ of undesirable destinations, and the goal is to avoid them as long as possible. Define the hitting time by

$$\tau \overset{def}{=} \min\{\, k \mid S_k \in \mathcal{S}'\},$$

and we set the objective to be maximizing the expected lifetime $E\{\tau\}$. As the following shows, we can upper-bound the probability of the event $\{\tau = \infty\}$, i.e., the probability of never being absorbed into $\mathcal{S}'$, and furthermore, in some cases when absorbing into $\mathcal{S}'$ happens with probability 1, we can upper-bound the maximal expected lifetime $E\{\tau\}$ of this problem.

Define the per-stage cost function of this POMDP as,

$$g(s,u) = 0, \quad \forall s \in \mathcal{S}', \qquad g(s,u) = -1, \quad \forall s \in \mathcal{S}\backslash\mathcal{S}', \quad \forall u \in \mathcal{U}.$$

By a similar argument as in Footnote 1 of the reaching example, one can show that the optimal liminf and limsup average cost functions are equal, and furthermore, the optimal average cost function, denoted by $J^*(\xi)$, is minus the maximal probability of never being absorbed into $\mathcal{S}'$, i.e,

$$J^*(\xi) = -\sup_{\pi\in\Pi} \mathbb{P}^\pi(\text{never reach } \mathcal{S}' \mid S_0 \sim \xi).$$

Choose a lower discretization scheme and compute the optimal average cost function $\tilde{J}^*$ for the corresponded modified problem. Since $\tilde{J}^* \leq J^*$, it follows that starting from a distribution ξ, under any policy, the probability

$$\mathbb{P}(\text{ never reach } \mathcal{S}' \mid S_0 \sim \xi) \leq -\tilde{J}^*(\xi), \tag{8.4}$$

and the inequality is non-trivial when $\tilde{J}^*(\xi) \neq -1$.

Suppose $\tilde{J}^*(\xi) = 0$. This implies that under any policy, with probability 1 the system will be absorbed into $\mathcal{S}'$. In this case we can bound the maximal expected lifetime as follows. Let $(\tilde{J}^*, \tilde{h}^*)$ be the solution pair to the 0-discount optimality equations of the modified problem. Then, since the per-stage cost is non-positive, by Theorem 7.1 of Section 7.1, $\tilde{h}^*(\xi)$ is the optimal total cost of the modified problem, and furthermore, it is a lower bound of the optimal total cost $V^*(\xi)$ of the original POMDP. Since the optimal total cost $V^*(\xi)$ of this problem is

$$V^*(\xi) = -\sup_{\pi\in\Pi} E^\pi\{\tau \mid S_0 \sim \xi\},$$

and $\tilde{h}^*(\xi) \leq V^*(\xi)$, thus the maximal expected life time is bounded:

$$\sup_{\pi\in\Pi} E^\pi\{\tau \mid S_0 \sim \xi\} \leq -\tilde{h}^*(\xi). \tag{8.5}$$

In summary, if $\tilde{J}^*(\xi)$ is zero, then starting from the distribution ξ, the maximal expected lifetime of this problem is no greater than $-\tilde{h}^*(\xi)$. □

Example 8.3 (Identification). Suppose we have a POMDP problem in which the model, not exactly known, belongs to a *finite* set Θ of possible models. Assume a prior distribution

p_0 on Θ is given. The goal is to identify the true model of the POMDP while controlling the system. We consider a simple case where identification is assumed to be the only goal and there is no concern for the cost as in adaptive control. We show that one can upper-bound the maximal identification probability of this problem as follows.

First we cast the identification problem into a new POMDP problem with augmented state space $\Theta \times \mathcal{S}$ and augmented control space $\Theta \times \mathcal{U}$. The augmented states $(i, s) \in \Theta \times \mathcal{S}$ are not observable. The observations are as in the original POMDP. Define the per-stage cost of the state $(i, s) \in \Theta \times \mathcal{S}$ and control $(j, s) \in \Theta \times \mathcal{U}$ as

$$g\big((i,s),(j,u)\big) = 0, \quad \text{if } i = j, \qquad g\big((i,s),(j,u)\big) = 1, \quad \text{if } i \neq j.$$

So, the system can be controlled with no cost, if the true model has been identified.

We now study the relation between the optimal average cost of this augmented POMDP and the probability of identification. Let the initial distribution of the augmented POMDP be the product distribution $p_0 \otimes \xi$, where ξ is the initial distribution of the state S_0 of the original POMDP. (It is not essential in assuming a product distribution.) For every sample path, we say that a policy identifies the model in k steps for that path, if the policy applies control $(\bar{i}, U_t)$ with $\bar{i}$ being the true model for all $t \geq k$. Define the probabilities of the identification events as

$$q_N^\pi(\xi) = \mathbb{P}^\pi(\text{model identified in N steps } \mid S_0 \sim \xi), \qquad q_\infty^\pi = \lim_{N \to \infty} q_N^\pi(\xi).$$

Thus q_∞^π is the probability that the model is identified in finite steps, while $1 - q_\infty^\pi$ is the non-identification probability, under the policy π. One can show that the optimal liminf average cost is less than the non-identification probability,[2] i.e., for all π,

$$J_-^*(p_0 \otimes \xi) \leq 1 - q_\infty^\pi.$$

Choose a lower discretization scheme[3] and compute the optimal average cost $\tilde{J}^*$ of the corresponded modified problem. Then for all π, $1 - q_\infty^\pi(\xi) \geq \tilde{J}^*(p_0 \otimes \xi)$, or, equivalently, under any policy

$$\mathbb{P}(\text{model identifiable in finite steps } \mid S_0 \sim \xi) \leq 1 - \tilde{J}^*(p_0 \otimes \xi), \tag{8.6}$$

and the inequality is non-trivial when $\tilde{J}^*(p_0 \otimes \xi) \neq 0$. □

We address the distinction between the identification problem considered here and the seemingly equivalent problem in reinforcement learning. One may recall that reinforce-

[2] To see this, fix N first, and consider the T-stage cost $V_T^\pi(p_0 \otimes \xi)$. For those sample paths that the model is identified in N steps, the T-stage cost is less than N, while for the rest of the sample paths, the T-stage cost is at most T. Thus

$$\begin{aligned} &\frac{1}{T} V_T^\pi(p_0 \otimes \xi) \leq q_N^\pi(\xi) \frac{N}{T} + \big(1 - q_N^\pi(\xi)\big) \frac{T}{T} = q_N^\pi(\xi) \frac{N}{T} + \big(1 - q_N^\pi(\xi)\big) \\ \Rightarrow \quad &\limsup_{T \to \infty} \frac{1}{T} V_T^\pi(p_0 \otimes \xi) \leq \lim_{T \to \infty} q_N^\pi(\xi) \frac{N}{T} + \big(1 - q_N^\pi(\xi)\big) = 1 - q_N^\pi(\xi). \end{aligned}$$

Letting $N \to \infty$, it follows that

$$\limsup_{T \to \infty} \frac{1}{T} V_T^\pi(p_0 \otimes \xi) \leq 1 - q_\infty^\pi(\xi), \quad \Rightarrow \quad J_-^*(p_0 \otimes \xi) \leq J_+^*(p_0 \otimes \xi) \leq 1 - q_\infty^\pi(\xi).$$

[3] As a reminder, we should not choose a scheme that assumes the knowledge of the true model.

ment learning methods can obtain asymptotically optimal controls while the controller is operating, without prior knowledge of the model. For this to be possible, however, the per-stage cost must be physically present while the controller is operating, while in reality the per-stage cost can be merely a design mechanism in order to keep the system in certain desired status, so that it is not physically present. The latter case is of interest in the above identification problem. (Certainly, using simulation, reinforcement learning methods for POMDP can be used to compute control policies for this problem.)

8.2 Entropy Rate of Hidden Markov Sources

8.2.1 Introduction to the Problem

Computing the entropy rate of a hidden Markov source is a classic problem in information theory. Let $\{Y_n\}_{n\geq 1}$ be the observation process of the Markov chain $\{S_n\}_{n\geq 1}$. It is for notational convenience that we let the index start with 1. Assume both the state and observation spaces are finite. We write Y^n for $(Y_1, \ldots, Y_n)$, Y_k^n for $(Y_k, \ldots, Y_n)$ and similar notation for S^n. Assume that the Markov chain $\{S_n\}$ is under an equilibrium distribution, so that $\{Y_n\}$ is also a stationary process. The entropy rate $H(\underline{Y})$ of $\{Y_n\}$ is defined by

$$H(\underline{Y}) \overset{def}{=} \lim_{n\to\infty} \frac{1}{n} H(Y^n). \tag{8.7}$$

It is equal to

$$H(\underline{Y}) = H'(\underline{Y}) \overset{def}{=} \lim_{n\to\infty} H(Y_n \mid Y^{n-1}),$$

and this equality relation and the existence of both limits in the two preceding definitions are general properties of stationary processes (Theorem 4.2.1 of [CT91]).

The method of computing the entropy rate $H(\underline{Y})$ in the textbook of information theory (Section 4 of [CT91]) is to compute the lower and upper bounds of $H(\underline{Y})$:

$$H(Y_n \mid Y^{n-1}, S_1) \leq H(\underline{Y}) \leq H(Y_n \mid Y^{n-1}). \tag{8.8}$$

The quantities $H(Y_n \mid Y^{n-1}, S_1)$ and $H(Y_n \mid Y^{n-1})$ are proved to converge to the entropy rate $H(\underline{Y})$ in the limit (Theorem 4.4.1 of [CT91]).

We have a new computation method for the entropy rate using the discretized lower approximation approach. We first describe the reduction of the problem to the form of an uncontrolled POMDP with a concave per-stage cost function, then address the application of lower approximations, as well as asymptotic convergence issues.

8.2.2 Reduction to a POMDP Problem

It is a basic fact of entropy that

$$H(Y^n) = \sum_{k=1}^{n} H(Y_k \mid Y^{k-1}). \tag{8.9}$$

As will be elaborated, this can be viewed as the n-stage cost of an uncontrolled POMDP with $H(Y_k \mid Y^{k-1})$ being the expected cost of stage k, so that the entropy rate is the average cost of that POMDP. The reduction to the control problem here, with details as follows, is a special case of the reduction result in [BMT01].

We will consider the process $\{S_n\}$ starting with an arbitrary initial distribution ξ of S_1. For this, let us first make explicit the dependence on ξ by rewriting the joint entropy of Y^n as

$$H(Y^n; \xi).$$

We use similar notation for conditional entropies. We denote an equilibrium distribution of the Markov chain $\{S_n\}$ by $\bar{\xi}$.

The entropy rate $H(\underline{Y})$ can be viewed as the average cost of the uncontrolled POMDP with initial distribution $\bar{\xi}$ and a per-stage cost function concave in the belief ξ defined as

$$g(\xi) \overset{def}{=} H(Y_1; \xi) = -E\Big\{ \log \Big(\sum_s \xi(s) p(Y_1 \mid s) \Big) \Big\}, \tag{8.10}$$

i.e., $g(\xi)$ is the entropy of Y_1 given that S_1 is distributed as ξ. The reduction of the entropy rate problem to a POMDP follows by first noticing that

$$p(Y_k \mid Y^{k-1}) = p(Y_k \mid S_k \sim \xi_k), \quad \xi_k(\cdot) = p(S_k \in \cdot \mid Y^{k-1}),$$

so that

$$\begin{aligned} & E\{\log p(Y_k \mid Y^{k-1}) \mid Y^{k-1}\} = g(\xi_k) \\ & \qquad \Rightarrow \quad H(Y_k \mid Y^{k-1}; \xi) = E\{E\{\log p(Y_k \mid Y^{k-1}) \mid Y^{k-1}\}\} = E\{g(\xi_k)\}. \end{aligned}$$

Hence $H(Y_k \mid Y^{k-1}; \xi)$ is the expected cost at stage k, with per-stage cost as defined by (8.10). The expected n-stage cost $J_n(\xi)$ in this POMDP is

$$J_n(\xi) = \sum_{k=1}^{n} H(Y_k \mid Y^{k-1}; \xi) = H(Y^n; \xi).$$

The finite-horizon Bellman equation $J_n(\xi) = g(\xi) + E\{J_{n-1}(\xi_2)\}$ is simply the chain rule of entropy

$$H(Y^n; \xi) = H(Y_1; \xi) + H(Y_2^n \mid Y_1; \xi).$$

By definition the average cost of this uncontrolled POMDP with initial distribution $\xi_1 = \bar{\xi}$ equals

$$\lim_{n \to \infty} \frac{1}{n} E\Big\{ \sum_{k=1}^{n} g(\xi_k) \Big\} = \lim_{n \to \infty} \frac{1}{n} H(Y^n; \bar{\xi}) = H(\underline{Y}),$$

the entropy rate of $\{Y_n\}$.

8.2.3 Lower Bounds of Entropy Rate

The application of lower bounds hence becomes immediate for this uncontrolled POMDP. The constructions of the fictitious process carry through and yield inequalities for the optimal finite-stage cost functions. The interpretation of the inequalities, translated to the information terminology, is simply the inequality of the conditional entropy

$$H(\tilde{Y}^1; \xi) + H(\tilde{Y}_2^n \mid \tilde{Y}^1, Q; \xi) \le H(\tilde{Y}^n; \xi) = H(Y^n; \xi), \tag{8.11}$$

where $\{Q, (\tilde{S}_n, \tilde{Y}_n)_{n\geq 1}\}$ is a process such that the marginal distribution of $(\tilde{S}_1, \tilde{Y}_1)$ and the conditional distribution of $\{(\tilde{S}_n, \tilde{Y}_n)_{n\geq 2}\}$ given $\tilde{Y}_1$ are the same as those in the original process $\{S_n, Y_n\}$, respectively, so that in particular, $\{\tilde{Y}_n\}$ and $\{Y_n\}$ have the same marginal distribution.

Apply any discretized lower approximation scheme to obtain the modified belief MDP problem; despite the non-linearity in the per-stage cost function, the solution methods for the modified problem is the same as in the POMDP problem that we considered. The average cost function $\tilde{J}^*$ of the modified problem can be computed by solving one linear system of equations, and

$$\tilde{J}^*(\bar{\xi}) \leq H(\underline{Y}), \tag{8.12}$$

where $\bar{\xi}$ is the equilibrium distribution of the Markov chain $\{S_n\}$.

8.2.4 Asymptotic Convergence

Now we consider the asymptotic convergence issue. The asymptotic convergence results we had for POMDPs carry through here, because they are valid for general MDPs. Thus in order to use them to show the asymptotic convergence of $\tilde{J}^*$ to the entropy rate, we need conditions to ensure a constant average cost and a continuous differential cost.

That the average cost of the POMDP is constant, is equivalent to the statement that the entropy rate of $\{Y_n\}$ exists and is insensitive to the initial distribution ξ. Although this seems an expected result, we have not found such a statement in the textbook of information theory. So we provide one for completeness. The proof is given in Appendix B.

Proposition 8.1. *Suppose the Markov chain $\{S_n\}$ is irreducible.*[4] *Then*

$$H(\underline{Y}) = \lim_{n\to\infty} \frac{1}{n} H(Y^n; \xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Next we claim that the conditional entropy is also insensitive to the initial distribution when the Markov chain is aperiodic. The proof is given in Appendix B.

Proposition 8.2. *Suppose the Markov chain $\{S_n\}$ is irreducible and aperiodic. Then*

$$\lim_{n\to\infty} H(Y_n \mid Y^{n-1}; \xi) = H(\underline{Y}), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

For the asymptotic convergence of $\tilde{J}^*$ to the entropy rate $H(\underline{Y})$, it is now sufficient that the differential cost is continuous. We thus make the following assumption.

Assumption 8.1. *The sequence*

$$H(Y^n; \xi) - nH(\underline{Y}) = \sum_{k=1}^{n} \Big(H(Y_k \mid Y^{k-1}; \xi) - H(\underline{Y}) \Big)$$

as functions of ξ, converges uniformly.

Under Assumption 8.1, the limiting function is continuous, and as easy to show, it is indeed the differential cost.

[4] This means that the Markov chain has only one recurrent class and a possibly non-empty set of transient states.

In view of Prop. 8.2, a sufficient condition for Assumption 8.1 to hold is when the convergence rate of the conditional entropy $H(Y_n \mid Y^{n-1}; \xi)$ is of $O(1/n^{1+\alpha})$ for some positive α. At this point, however, we are not yet clear about the convergence rate of the conditional entropies, and whether Assumption 8.1 hold for a general irreducible Markov chain $\{S_n\}$.

In summary, we have the following proposition.

Proposition 8.3. *Suppose the Markov chain $\{S_n\}$ is irreducible and Assumption 8.1 holds. Then, for every initial distribution ξ,*

$$\lim_{\epsilon \to 0} \tilde{J}_\epsilon^*(\xi) = H(\underline{Y}),$$

where $\tilde{J}_\epsilon^*(\cdot)$ is the optimal average cost function of the modified problem corresponding to either $\widetilde{\mathfrak{T}}_{D_1}$ or $\widetilde{\mathfrak{T}}_{D_2}$ of Section 4.1 associated with an ϵ-disretization scheme.

We also see that a constant upper bound on $H(\underline{Y})$ can be in principle computed from the solution $(\tilde{J}^*, \tilde{h})$ of the modified problem using the Bellman residue, as in the POMDP case. However, this upper bound may be loose, and it is also hard to compute exactly.

8.3 Summary

We have shown the use of the lower bounds in the problems of reaching, avoidance, and identification, as well as the classic problem of computing entropy rate of hidden Markov sources in information theory.

The problem of entropy rate of hidden Markov sources is technically a special case of the sequential quantization problem of hidden Markov sources considered by Borkar, Mitter and Tatikonda [BMT01]. Their work shows that one can cast the sequential quantization problem into an average cost POMDP problem with a concave per-stage cost function that measures coding cost and distortion cost. Thus, the discretized lower approximation approach can be applied to the sequential quantization problem, when the state space is finite, to obtain lower bounds and suboptimal quantizers. Furthermore, one can also formulate the sequential quantization problem as a constrained average cost POMDP problem with a concave per-stage cost function and distortion constraints. Discretized approximations can be applied to the constrained average cost POMDP as well, as will be shown in Chapter 9.

Chapter 9

Lower Bounds for Constrained Average Cost POMDPs

9.1 Introduction

Consider a finite space POMDP problem with multiple per-stage cost functions $g_0, g_1, \dots, g_n$. The objective, to be made precise later, is to minimize the average cost with respect to the per-stage cost function g_0, subject to prescribed bounds on the average costs with respect to other per-stage cost functions g_k. In practice different cost functions may correspond to consumption of different resources, say. This type of problem is also called a multi-objective problem.

More precisely, we define the constrained POMDP problem as follows. For $k = 0, \dots, n$, define $J^{\pi}_{k,+}$ as the limsup average cost of policy π with respect to the k-th per-stage cost function g_k:

$$J^{\pi}_{k,+}(\xi) \overset{def}{=} \limsup_{T \to \infty} \frac{1}{T} E^{\pi} \left\{ \sum_{t=0}^{T-1} g_k(S_t, U_t) \mid S_0 \sim \xi \right\}. \tag{9.1}$$

Let $c_1, \dots, c_n$ be given constants. The constrained average cost problem and its optimal cost function are defined as for each ξ,

$$\begin{aligned} J^*_c(\xi) &\overset{def}{=} \inf_{\pi \in \Pi} J^{\pi}_{0,+}(\xi), \\ &\quad \text{Subj.} \quad J^{\pi}_{k,+}(\xi) \le c_k, \quad k = 1, \dots, n. \end{aligned} \tag{9.2}$$

So the feasible set, denoted by Π_f, is

$$\Pi_f \overset{def}{=} \{\pi \in \Pi \mid J^{\pi}_{k,+}(\xi) \le c_k, \forall \xi \in \mathcal{P}(\mathcal{S}), k = 1, \dots, n\},$$

i.e., the set of policies of which the *limsup average cost* with respect to g_k is less than c_k for all initial distributions. We assume that Π_f is non-empty.

There are both theoretical and computational difficulties in solving this constrained average cost POMDP problem (9.2). Not much is known about the nature of the optimal or ϵ-optimal policies, e.g., their stationarity.

We show in this chapter how to compute lower bounds of the constrained optimal cost function by applying discretized lower approximation schemes proposed previously for unconstrained POMDPs. In particular, we will show:

- The constrained optimal cost of the modified belief MDP problem is a lower bound of the constrained optimal cost of the original POMDP. This is another immediate consequence of the stronger lower bound result, Theorem 3.2.
- When the modified belief MDP is unichain, its constrained optimal average cost, a constant, can be solved by one single finite-dimensional linear program, the unichain LP of a finite-state and control MDP.
- When the modified belief MDP is multichain, one can consider for *each* initial belief ξ, a finite-dimensional linear program that is related (but not identical) to the constrained modified problem. The value of this LP equals the constrained optimal of the modified problem, and is thus a lower bound of $J_c^*(\xi)$.
- Regardless of the chain structure of the modified belief MDP, a constant lower bound of $J_c^*(\xi)$ can be obtained by solving one single LP, the unichain LP.
- If any one of the LPs considered is infeasible, then the original constrained POMDP problem is infeasible.

We will also briefly comment on ways of using the policies obtained from the modified problem at the end of this chapter.

9.2 The Constrained Modified Problem and Lower Bounds of J_c^*

Consider the modified belief MDP associated with a discretized lower approximation scheme. Denote by $\bar{g}_k$ the k-th per-stage cost, i.e.,

$$\bar{g}_k(\xi, u) = \sum_{s \in \mathcal{S}} g_k(s, u)\, \xi(s), \quad k = 0, 1, \ldots, n.$$

Denote by $\widetilde{\Pi}$ the set of policies in the modified problem. The constrained modified problem can be stated as: For every initial belief state ξ,

$$\begin{aligned} \tilde{J}_c^*(\xi) \overset{def}{=} \min_{\pi \in \widetilde{\Pi}} \ & \tilde{J}_{0,+}^{\pi}(\xi) \\ \text{Subj.} \quad & \tilde{J}_{k,+}^{\pi}(\xi) \le c_k, \quad k = 1, \ldots, n, \end{aligned} \tag{9.3}$$

where for each k, $\tilde{J}_{k,+}^{\pi}$ is the limsup average cost with respect to the per-stage cost $\bar{g}_k$ in the modified problem.

Theorem 9.1. *Suppose the original constrained POMDP problem is feasible. Then the constrained modified problem is feasible, and its value satisfies*

$$\tilde{J}_c^*(\xi) \le J_c^*(\xi), \qquad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Proof: This is again an immediate consequence of the stronger lower bound result, Theorem 3.2, applied simultaneously to multiple per-stage cost functions and to the average cost case. □

To solve the constrained modified problem (9.3), we can use the linear programming method for solving constrained finite-state and control MDPs. This is the focus of the rest of this secition.

There is one minimization problem for each ξ. However, recall that in the modified problem there is a finite set $\mathcal{C}$ of supporting beliefs, and outside $\mathcal{C}$ all states are transient and lead to $\mathcal{C}$ in one step under any policy. In the unconstrained case this has allowed us to solve the modified problem by solving the restricted finite state MDP problem on $\mathcal{C}$. The constrained case is different, however. From the theory on constrained finite state and control MDP (see Puterman [Put94] or Altman [Alt99]) we know the following. If the modified belief MDP is unichain, then it is sufficient to consider only stationary randomized policies on the finite set $\mathcal{C}$, and furthermore, the constrained problem can be reduced to one minimization problem. If the modified belief MDP is multichain, then for transient states $\xi \notin \mathcal{C}$ the optimal policy on $\mathcal{C}$ can *differ* depending on ξ.[1] Hence we will consider the unichain and multichain cases separately.

9.2.1 The Unichain Case

First consider the case where the modified problem is unichain. As a standard result from the finite-state and control MDP theory, the optimal average cost $\tilde{J}_c^*$ of the constrained problem is a constant, and can be solved by one single linear program (see e.g., Section 8.9 of Puterman [Put94]):

$$\tilde{J}_c^* = \min_{q \geq 0} \sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_0(s,u)\, q(s,u) \tag{9.4}$$

$$\text{Subj.} \quad \sum_{u \in \mathcal{U}} q(s',u) - \sum_{s \in \mathcal{C}, u \in \mathcal{U}} p(s'|s,u)\, q(s,u) = 0, \quad \forall s' \in \mathcal{C} \tag{9.5}$$

$$\sum_{s \in \mathcal{C}, u \in \mathcal{U}} q(s,u) = 1, \tag{9.6}$$

$$\sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_k(s,u)\, q(s,u) \leq c_k, \quad k = 1, \ldots, n. \tag{9.7}$$

Here the transition probabilities $p(s'|s,u)$ are from the modified belief MDP model. A feasible solution $q(s,u), \forall (s,u)$, can be interpreted as the limiting probabilities of being at state s and taking action u, under the stationary randomized policy:

$$p(u|s) = \frac{q(s,u)}{\sum_{u'} q(s,u')}, \quad \forall (u,s).$$

Thus, if the modified problem is unichain, we can compute the lower bound $\tilde{J}_c^*$, a constant, by solving one single finite-dimensional LP.

Remark 9.1. We comment on the unichain condition. It is easy to check whether the MDP is communicating or weakly communicating, (see the model classification algorithm, pp. 351 of [Put94]). We do not know any efficient algorithm to check the unichain condition,

[1] Essentially, this is because a feasible policy can have its long-run average cost violate the constraint on one recurrent class, with this violation compensated by its long-run average cost on another recurrent class that is also reachable from the initial transient state.

however. Nonetheless, as the analysis for the multichain case will show, the value of the LP (9.4) is a lower bound of J_c^* regardless of the chain structure.

9.2.2 The Multichain Case

When the modified belief MDP is multichain, its constrained average cost problem cannot be solved by one single linear program, as in the unconstrained multichain case. So instead, we will consider an LP formulation for the constrained problem on the finite set of belief states $\{\xi\} \cup \mathcal{C}$ for *every* initial belief ξ. Let $\bar{s} = \xi$. Consider the following LP, which relates but does not exactly correspond to the constrained modified problem:

$$\hat{J}^*(\bar{s}) = \min_{q \geq 0, y \geq 0} \sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_0(s, u)\, q(s, u) \tag{9.8}$$

$$\text{Subj.} \quad \sum_{u \in \mathcal{U}} q(s', u) - \sum_{s \in \mathcal{C}} p(s'|s, u)\, q(s, u) = 0, \quad \forall s' \in \mathcal{C} \tag{9.9}$$

$$\sum_{u \in \mathcal{U}} q(\bar{s}, u) + \sum_{u \in \mathcal{U}} y(\bar{s}, u) - \sum_{s \in \mathcal{C}, u \in \mathcal{U}} p(\bar{s}|s, u)\, y(s, u) = 1, \tag{9.10}$$

$$\sum_{u \in \mathcal{U}} q(s', u) + \sum_{u \in \mathcal{U}} y(s', u) - \sum_{s \in \mathcal{C}, u \in \mathcal{U}} p(s'|s, u)\, y(s, u) = 0, \quad \forall s' \neq \bar{s},\ s' \in \mathcal{C} \tag{9.11}$$

$$\sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_k(s, u)\, q(s, u) \leq c_k, \quad k = 1, \ldots, n. \tag{9.12}$$

We now review certain facts on constrained finite-state MDPs to show the relation between the LP (9.8) and the constrained modified problem.

For every stationary randomized policy π (in the modified belief MDP), there is a corresponding pair (q_π, y_π) feasible for the constraints (9.9)-(9.11) in the LP (9.8). The vector $q_\pi(s, u)$ corresponds to the limiting probabilities of (s, u) under π starting from the initial state $\bar{s}$, and $q_\pi(s, u)$ are non-zero only on recurrent states reachable from $\bar{s}$.

However, the feasible set for the constraints (9.9)-(9.11) includes more points than the limiting probabilities realizable by all policies, when the MDP is multichain. For this reason, the LP (9.8) is not equivalent to the constrained modified problem. The latter can be cast into the following minimization problem:

$$\tilde{J}_c^*(\bar{s}) = \min_{q \in Q_{\bar{s}}} \sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_0(s, u)\, q(s, u) \tag{9.13}$$

$$\text{Subj.} \quad \text{constraints (9.12)},$$

where $Q_{\bar{s}}$ is the set of limits of average state-action frequencies realizable by some policy starting from the initial state $\bar{s}$. The set $Q_{\bar{s}}$ is convex and compact,[2] and it is equal to the convex hull of the limiting probabilities that are realizable by stationary randomized policies (Theorem 8.9.3 of [Put94]). Hence it is sufficient to consider policies that are convex combinations of stationary randomized policies for each initial state.

The minimization problem (9.13) is not easy to solve because of the complex constraint set $Q_{\bar{s}}$. So we will solve the LP (9.8). The optimal solution of the LP (9.8) may not be

[2] In fact in the finite space case, $Q_{\bar{s}}$ is a polyhedral set. However, we do not have to use this fact in the subsequent proofs, due to a stronger result from the minmax theory.

realizable by any policy, but the values of the two problems are actually equal, as we will address later. So, for the sake of obtaining lower bounds of the original constrained POMDP problem, the LP (9.8) is sufficient.

Proposition 9.1. *Suppose the original constrained POMDP problem is feasible. Then for each $\xi \in \mathcal{P}(\mathcal{S})$, the constrained modified problem (9.13) and the LP (9.8) are feasible, and their values are equal:*

$$\hat{J}^*(\xi) = \tilde{J}^*_c(\xi).$$

Proof: Let $\bar{s} = \xi$. Consider the Lagrangians of both problems formed by relaxing the constraints (9.12). Corresponding to multipliers $\lambda = (\lambda_1, \ldots, \lambda_n)$, the dual function for the LP (9.8) is,

$$F_1(\lambda) = \min_{q \geq 0} \sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_0(s,u)\, q(s,u) + \sum_{k=1}^{n} \lambda_k \left(\sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_k(s,u)\, q(s,u) - c_k \right) \tag{9.14}$$

$$\text{Subj.} \quad \text{constraints (9.9), (9.10) and (9.11),}$$

while the dual function for the constrained modified problem (9.13) is

$$F_2(\lambda) = \min_{q \in Q_{\bar{s}}} \sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_0(s,u)\, q(s,u) + \sum_{k=1}^{n} \lambda_k \left(\sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_k(s,u)\, q(s,u) - c_k \right). \tag{9.15}$$

For fixed λ, both problem (9.14) and problem (9.15) correspond to the unconstrained average cost multichain MDP problem (modulo a constant term, $-\sum_k \lambda_k c_k$, in the objective function) with the combined per-stage cost $\bar{g}_0 + \sum_{k=1}^{n} \lambda_k \bar{g}_k$. Therefore, $F_1(\lambda) = F_2(\lambda)$, which implies that the dual problems have the same dual optimal value:

$$\max_{\lambda \geq 0} F_1(\lambda) = \max_{\lambda \geq 0} F_2(\lambda).$$

By the strong duality of LP, there is no duality gap between the LP (9.8) and the dual problem of maximizing $F_1(\lambda)$. Since the constraint set $Q_{\bar{s}}$ is compact, by the "minmax=maxmin" equality for the convex problem (9.13) (Corollary 37.3.2 of Rockafellar [Roc70]), there is no duality gap between the constrained modified problem (9.13) and the dual problem of mazimizing $F_2(\lambda)$. Therefore, the primal optimal values are also equal. □

Thus we can bound $J^*_c(\xi)$ from below for each belief ξ by solving a finite-dimensional LP associated with ξ. Now note that any feasible variable q of the LP (9.8) is also feasible for the LP (9.4), so we have the following corollary.

Corollary 9.1. *Suppose the original constrained POMDP problem is feasible. Then regardless of the chain structure of the modified problem, the LP (9.4) is feasible and its value is a lower bound of $J^*_c(\xi)$ for all ξ.*

The next corollary is about the feasibility of the original constrained problem.

Corollary 9.2. *If any one of the LPs (9.4) and (9.8) for the constrained modified problem is infeasible, then the original constrained POMDP problem is infeasible.*

Remark 9.2. The results of Section 9.2.1 and 9.2.2 remain the same for constrained POMDP problems with per-stage cost functions that are concave in the belief. This is because the modified belief MDP has the lower approximating property for any concave per-stage cost function as well (for proofs, see the remark in Section 3.7 for using the weaker line of analysis and Section 3.8.3 for using the stronger line of analysis).

Remark 9.3. One can also prove the preceding results, for both the unichain and multichain cases, using the weaker lower bound result that depends on the DP mapping. The proof method is to use the lower bound result for the unconstrained average cost problem and the strong duality of LPs or convex programs. For reference, we include such a proof for the unichain case (the multichain case is similar) in Appendix A.2.

9.3 Summary

We have shown in this chapter how to compute lower bounds of the constrained average cost POMDP problems using discretized lower approximations. Like in unconstrained cases, these lower bounds can provide performance measure for suboptimal policies, which can be obtained by other approximation methods, e.g., the finite-state controller approach that will be addressed later.

We now discuss briefly about the policies. Consider the optimal solutions to the LPs of the previous sections. The solution corresponds to a stationary randomized policy when the modified problem is unichain, and the solution may not correspond to any policy when the modified problem is multichain. However, it is always easy to verify whether or not the solution corresponds to a stationary randomized policy and to obtain that policy when it does.

As to applying that policy in the original constrained POMDP problem, a few issues need consideration. First, the policy may not be feasible in the original POMDP. Secondly, as discussed in the unconstrained case, for the modified problem most of the belief states are transient, and hence the controls at these states do not affect the average cost of the policy in the modified problem. So it may be possible, especially when the modified problem is unichain, that the controls in the LP solution are arbitrary on these transient states. Therefore for the purpose of suboptimal control in the original problem, one should be cautious about these policies. The idea of the sensitive optimality approach may be helpful.

Chapter 10

A Function Approximation Approach to Estimation of Policy Gradient for POMDPs with Structured Policies

10.1 Introduction

In this chapter we consider reinforcement learning algorithms for POMDPs with finite spaces of states, controls and observations. We consider the average cost criterion. As we will refer often to certain key works of this area, for the convenience of comparisons, we will use notation slightly different to the previous chapters. We denote by X_t the state, Y_t the observation, and U_t the control at time t. The random per-stage cost at time t is $g_t = g(X_t, Y_t, U_t)$, where $g(\cdot)$ is the per-stage cost function depending on the state and control, as well as the observation. The state transition probability is denoted by $p(X_{t+1}|X_t, U_t)$ and the observation probability by $p(Y_{t+1}|X_{t+1}, U_t)$.

We limit the policy space to the set of finite-state controllers. A finite-state controller is like a probabilistic automaton, with the observations being its inputs and the controls its outputs. The controller has a finite number of "internal-states" that evolve in a Markovian way, and it outputs a control depending on the current internal state and the current observation.

The finite-state controller approach to POMDPs has been proposed in the work of "GPOMDP" [BB01] by Baxter and Bartlett, and "Internal-state POMDP" [AB02] by Aberdeen and Baxter. There are two distinctive features about finite-state controllers. One is that the state of a POMDP, (even though not observable), the observation, and the internal state of the controller jointly form a Markov process, so the theory of finite-state Markov decision processes (MDP) applies. In contrast, the asymptotic behavior of a POMDP under a general policy is much harder to establish. The other distinctive feature of finite-state controllers is that the gradient of the cost with respect to the policy parameters can be estimated from sample trajectories, without requiring the explicit model of a POMDP, so gradient-based methods can be used for policy improvement. This feature is appealing for both large problems in which either models are not represented explicitly, or exact inferences are intractable, and reinforcement learning problems in which the environment model is unknown and may be varying in time.

In this chapter we consider the problem of learning a finite-state controller, i.e., optimizing the parameters of the controller, assuming no knowledge of the model and the per-stage cost function. We note that if we have the model or if we can simulate the system, then there are many algorithms immediately applicable. For example, if the model is known and its size tractable, then one can directly compute the gradient and solve the optimization problem using gradient-based methods. If the model is too large, but the system can be simulated, then one can directly apply policy gradient or actor-critic algorithms for MDPs to estimate the policy gradient, because the state information that is necessary to those algorithms, is available in simulation. We consider the case beyond simulation where the controller is to be tuned in a real environment while it is operating, so that both the controller and the learning algorithm do not have access to the hidden states. The main question is then how to overcome the hidden states and estimate the policy gradient.

The gradient estimation method proposed by [BB01] and [AB02] avoids estimating the value function. The idea there is to replace the value of a state in the gradient expression by the path-dependent random cost starting from that state. To our knowledge, up to now gradient estimators that use a value function approximator have not been proposed as an alternative to GPOMDP in learning finite-state controllers for a POMDP.[1] The purpose of this chapter is to propose such an alternative.

We show that the gradient is computable by a function approximation approach. Without pre-committing to a specific estimation algorithm, our methodology starts with rewriting the gradient expression by integrating over the state variable, so that the new gradient expression involves a "value" function that does not depend on the states. This "value" function is shown to be the conditional mean of the true value function of a certain Markov chain under the equilibrium distribution, conditioned on observations, controls, and internal states. By ergodicity, asymptotically unbiased estimates of this "value" function can be obtained from sample trajectories. In particular, temporal difference (TD) methods with linear function approximation, including both β-discounted TD(λ) and average cost TD(λ), can be used, and the biases of the corresponding gradient estimators asymptotically go to zero when $\beta \to 1, \lambda \to 1$.

The computation of this value function may be viewed as the critic part of the Actor-Critic framework (e.g., Konda [Kon02]), in which the critic evaluates the policy, and the actor improves the policy based on the evaluation. Thus for POMDPs with finite-state controllers, the algorithms as well as their analysis fit in the general MDP methodology with both actor-only and actor-critic methods, and can be viewed as special cases.

The idea of estimating the conditional mean of the true value function first appeared in the work of Jaakkola, Singh and Jordan [JSJ94], which is a gradient-descent flavored method in the context of the finite memory approach to reinforcement learning in POMDPs. This earlier work does not start with gradient estimation, though it is closely related to it. It applies to a subclass of finite-state controllers that have the internal state storing a finite-length of the recent history.

Our way of using the conditional expectation in rewriting the gradient expression is to some degree new. Algorithmically, one does not have to take this additional step in order to apply the Actor-Critic framework. However, taking conditional expectations and making the conditional mean explicit in the gradient expression, we think, is a more direct

[1] Meuleau et al. [MPKK99] have used the value function to parameterize the policy and used the path-dependent random cost for gradient estimation in episodic settings. A GPOMDP/SARSA hybrid was proposed by Aberdeen and Baxter in an early work. However, the reasoning there was incorrect, because the marginal process of internal-state and observation is not a Markov chain.

approach. It allows the use of other estimation algorithms such as non-linear function approximators. Furthermore, it has facilitated our derivation of the gradient estimation algorithm for general finite-state controllers, in which case it would be less transparent to apply a projection argument normally used in MDP policy gradient methods.

Finally we show that the same function approximation approach also applies to gradient estimation in semi-Markov problems, for which a GPOMDP type algorithm was earlier proposed by Singh, Tadic and Doucet [STD02].

The organization of this chapter is as follows. In Section 10.2, we lay out our approach for reactive policies, a simple subclass of finite-state controllers, which captures all the main ideas in the analysis. We introduce some background, and present the gradient expressions, the algorithms, and an error analysis. In Section 10.3, we present the gradient estimation algorithm for finite-state controllers, and in Section 10.4, for semi-Markov problems. In Section 10.5, we provide experiments, showing that the estimates using function approximation are comparable to those from an improved GPOMDP method that uses a simple variance reduction technique, and in addition the function approximation approach provides more options in controlling bias and variance.

10.2 Gradient Estimation for POMDPs with Reactive Policies

We first present our approach for the simplest finite-state controllers – reactive policies, mainly for their notational simplicity. We will also introduce the background of policy gradient estimation. A reactive policy is a randomized stationary policy such that the probability of choosing a control is a function of the most recent observation only. The graphical model of a POMDP with a reactive policy is shown in Fig. 10-1. The process $\{(X_t, Y_t, U_t)\}$ jointly forms a Markov chain under a reactive policy, and so does the *marginal* process $\{(X_t, Y_t)\}$, (marginalized over controls U_t).

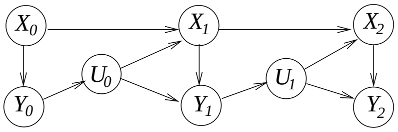

Figure 10-1: A POMDP with a reactive policy.

Let $\{\gamma_\theta \mid \theta \in \Theta\}$ be a family of reactive policies parametrized by θ. For any policy γ_θ, let

$$\mu_u(y, \theta) = p(U_t = u \mid Y_t = y; \theta)$$

be the probability of taking control u upon the observation y. The following assumptions are standard. We require that $\mu_u(y, \theta)$ is differentiable for any given u, y, and the transition probability $p(X_{t+1} = \bar{x}, Y_{t+1} = \bar{y} \mid X_t = x, Y_t = y; \theta)$ is differentiable for any given $x, y, \bar{x}, \bar{y}$. Furthermore we assume that for all $\theta \in \Theta$ the Markov chains $\{(X_t, Y_t)\}$ are defined on a common state space[2] and we make the following additional assumptions.

Assumption 10.1. *Under any policy γ_θ, the Markov chain $\{(X_t, Y_t)\}$ is irreducible and aperiodic.*

[2] Note that this is not necessarily the product of the state and the observation spaces of a POMDP.

Assumption 10.2. *There exists a constant L, such that for all $\theta \in \Theta$, $\max_{u,y} \left\| \frac{\nabla \mu_u(y,\theta)}{\mu_u(y,\theta)} \right\| \leq L$, where $0/0$ is regarded as 0.*

The first assumption ensures that the average cost is a constant and differentiable for all policies γ_θ. (A short proof of differentiability in finite-state MDPs in general is given in the Appendix C.) The second assumption of boundedness is to make it possible to compute the gradient by sampling methods.

10.2.1 Review of Gradient and Gradient Approximation

Let $\eta(\theta)$ be the average cost of the reactive policy γ_θ, and let E_0^θ denote expectation with respect to the *equilibrium distribution* of the Markov chain $\{(X_t, Y_t, U_t)\}$ under policy γ_θ. Conditional expectations are defined in the same way, i.e., with respect to the conditional distributions from the equilibrium distribution. For simplicity of notation, we will drop θ in $\eta(\theta)$ and E_0^θ, and use η and E_0 throughout the chapter.

Suppose $\theta = (\theta_1, \ldots, \theta_k) \in \mathcal{R}^k$, and let $\nabla \mu_u(y,\theta)$ be the column vector

$$\nabla \mu_u(y,\theta) = \left(\frac{\partial \mu_u(y,\theta)}{\partial \theta_1}, \ldots, \frac{\partial \mu_u(y,\theta)}{\partial \theta_k} \right)'.$$

By differentiating both sides of the optimality equation, it can be shown that the gradient $\nabla \eta$ can be expressed as

$$\nabla \eta = E_0 \left\{ \frac{\nabla \mu_U(Y,\theta)}{\mu_U(Y,\theta)} Q(X,Y,U) \right\}, \tag{10.1}$$

where the Q-function $Q(x,y,u)$ is defined by

$$Q(x,y,u) = g(x,y,u) + E\left\{ h(X_1,Y_1) \mid X_0 = x, Y_0 = y, U_0 = u \right\},$$

and $h(x,y)$ is the bias [Put94], defined by

$$h(x,y) = \lim_{T \to \infty} E \left\{ \sum_{t=0}^{T} (g_t - \eta) \mid X_0 = x, Y_0 = y \right\}.$$

In order to compute $h(x,y)$ directly one needs to pick a particular pair (x_0, y_0) as a regenerating state (Marbach and Tsitsiklis [MT01]). Since this is not possible in a POMDP,[3] one has to approximate it by other terms.

Baxter and Bartlett [BB01] proposed the following approximate gradient:

$$\nabla_\beta \eta \overset{def}{=} E_0 \left\{ \frac{\nabla \mu_U(Y,\theta)}{\mu_U(Y,\theta)} Q_\beta(X,Y,U) \right\}, \tag{10.2}$$

where Q_β is the Q-function of the discounted problem:

$$Q_\beta(x,y,u) = g(x,y,u) + \beta E\left\{ J_\beta(X_1,Y_1) \mid X_0 = x, Y_0 = y, U_0 = u \right\},$$

and J_β is the cost function of the β-discounted problem. The approximate gradient $\nabla_\beta \eta$ converges to $\nabla \eta$ when $\beta \uparrow 1$. This is due to the fact that when $\beta \approx 1$, $J_\beta - \frac{\eta}{1-\beta} \approx h$, and $E_0\{\frac{\nabla \mu_U(Y,\theta)}{\mu_U(Y,\theta)}\} = 0$, therefore $E_0\{\frac{\nabla \mu_U(Y,\theta)}{\mu_U(Y,\theta)} c\} = 0$ for any constant c. Although the state is not

[3] Note that in simulation one can apply the algorithm of [MT01] to estimate $h(x,y)$ directly, because the state process is generated by the simulator and therefore known, as discussed earlier.

observable, an estimate of $Q_\beta(X_t, Y_t, U_t)$ can be obtained by accumulating the costs along the *future* sample path starting from time t. This is the idea of the GPOMDP algorithm that estimates the approximate gradient $\nabla_\beta \eta$ by a sampling version of Eq. (10.2).

10.2.2 A New Gradient Expression for Estimation

We first write Eq. (10.1) in a different way:

$$\begin{aligned}\nabla\eta &= E_0\left\{\tfrac{\nabla\mu_U(Y,\theta)}{\mu_U(Y,\theta)} Q(X,Y,U)\right\} \\ &= E_0\left\{\tfrac{\nabla\mu_U(Y,\theta)}{\mu_U(Y,\theta)} E_0\{Q(X,Y,U) \mid Y,U\}\right\} \\ &= E_0\left\{\tfrac{\nabla\mu_U(Y,\theta)}{\mu_U(Y,\theta)} v(Y,U)\right\}, \qquad (10.3)\end{aligned}$$

where

$$v(Y,U) = E_0\left\{Q(X,Y,U) \mid Y,U\right\}, \qquad (10.4)$$

a function that depends on observation and action only. Similarly define $v_\beta(Y,U)$ to be the conditional mean of Q_β given Y, U, and the approximate gradient (Eq. (10.2)) can be written as

$$\nabla_\beta\eta = E_0\left\{\tfrac{\nabla\mu_U(Y,\theta)}{\mu_U(Y,\theta)} v_\beta(Y,U)\right\}. \qquad (10.5)$$

Thus if we can estimate $v(y,u)$ or its approximation $v_\beta(y,u)$ from sample paths, then we can estimate $\nabla\eta$ or $\nabla_\beta\eta$ using a sampling version of Eq. (10.3) or Eq. (10.5).

It turns out that by ergodicity of the Markov chain, we are able to compute v_β from a sample trajectory, and compute v with asymptotically no bias. This was first noticed in [JSJ94]. Let us reason informally why it is so for the case of $v_\beta(y,u)$. Let $\pi(x,y,u)$ be the equilibrium distribution of the Markov chain $\{(X_t, Y_t, U_t)\}$, and $\pi(y,u)$, $\pi(x|y,u)$ be the corresponding marginal and conditional distributions, respectively. For any sample trajectory $\{(y_t, u_t)\}_{t\leq T}$, by ergodicity the number of the sub-trajectories that start with (y,u), denoted by $T_{y,u}$, will be approximately $\pi(y,u)T$, as $T \to \infty$. Among these sub-trajectories the number of those that start from the state x will be approximately $\pi(x|y,u)T_{y,u}$. Thus averaging over the discounted total costs of these $T_{y,u}$ sub-trajectories, we obtain in the limit $v_\beta(y,u)$, as $T \to \infty$.

Using ergodicity, there are many ways of estimating $v(y,u)$ or $v_\beta(y,u)$ from sample paths. In what follows, we will focus on the temporal difference methods, as they have well-established convergence and approximation error analysis.

10.2.3 Computing $v_\beta(y,u)$ and $v(y,u)$ by TD Algorithms

We approximate the function $v(y,u)$ or $v_\beta(y,u)$ by a linear combination of n basis functions:

$$v(y,u) \approx \phi(y,u)'r, \quad \text{or} \quad v_\beta(y,u) \approx \phi(y,u)'r,$$

where the symbol ' denotes transpose,

$$\phi(y,u)' = [\phi_1(y,u), \ldots, \phi_n(y,u)]$$

is a vector with its entries $\phi_i(y,u)$ called the features of (y,u) and with $\phi_i(y,u)$ as a function of (y,u) being the i-th basis function, and r is a length-n column vector of linear coefficients, to be computed by TD algorithms.[4] Let Φ be the matrix

$$\Phi = \begin{bmatrix} \vdots \\ \phi(y,u)' \\ \vdots \end{bmatrix}, \tag{10.6}$$

with rows $\phi(y,u)'$. We require that the columns, i.e., the basis functions, are linearly independent, and the column space of Φ includes the set of functions

$$\left\{ \tfrac{1}{\mu_u(y,\theta)} \tfrac{\partial \mu_u(y,\theta)}{\partial_{\theta_1}}, \ldots, \tfrac{1}{\mu_u(y,\theta)} \tfrac{\partial \mu_u(y,\theta)}{\partial_{\theta_k}} \right\}$$

which we call the minimum set of basis functions.

We describe how TD algorithms are used in this case. For clarity of description, we define another identical set of features $\tilde{\phi}(x,y,u) = \phi(y,u)$. We run TD algorithms in a POMDP as if it were an MDP, with features $\tilde{\phi}$. Since $\tilde{\phi}$ does not depend on states x, we do not require state information in running the TD algorithms.

Similar to the case in an MDP [KT99], [SMSM99], it can be seen from Eq. (10.3) or Eq. (10.5) that for gradient estimation, it is not necessary to have the exact function $v(y,u)$ or $v_\beta(y,u)$. Instead, it suffices to have the projection of the function $v(Y,U)$ or $v_\beta(Y,U)$, viewed as a random variable, on a subspace that includes the minimum set of basis functions $\left\{ \tfrac{1}{\mu_U(Y,\theta)} \tfrac{\partial \mu_U(Y,\theta)}{\partial_{\theta_1}}, \ldots, \tfrac{1}{\mu_U(Y,\theta)} \tfrac{\partial \mu_U(Y,\theta)}{\partial_{\theta_k}} \right\}$, viewed as random variables, where the projection is with respect to the marginal equilibrium distribution $\pi(y,u)$.[5] Thus, our goal is to estimate the projection of the function $v(y,u)$ or $v_\beta(y,u)$ with asymptotically no bias using TD algorithms.

From established results on discounted TD(λ) (Tsitsiklis and Van Roy [TV97], see also Bertsekas and Tsitsiklis [BT96]) and average cost TD(λ) algorithms (Tsitsiklis and Van Roy [TV99]), if r^* is the limit that TD converges to, then $\phi(y,u)'r^*$ as a function of (x,y,u) is close to the projection of the function $Q(x,y,u)$ or $Q_\beta(x,y,u)$ with bias depending on the values of λ. In what follows, we combine these results with an error decomposition to show that $\phi(y,u)'r^*$ as a function of (y,u) is close to the projection of the function $v(y,u)$ or $v_\beta(y,u)$. Although the notation is for reactive policies, the analysis applies to the general case of finite-state controllers, where the variables (y,u) are replaced by an enlarged set of variables including the internal states of the controller. The question of what is the resulting value function $\phi(y,u)'r^*$, was discussed, yet not resolved, in an earlier work [SJJ94] by Singh, Jaakkola and Jordan, and our analysis thus clarifies this issue.

[4] TD algorithms include the original TD algorithms (e.g., [Sut88], [BT96], [TV99]), the least squares TD algorithms (e.g., [Boy99], [BBN03]), and many other variants. They differ in convergence rate and computation overhead, and they converge to the same limits.

[5] One can define an inner-product for the space of square-integrable real-valued random variables on the same probability space. For real-valued random variables X and Y, the inner-product is defined as $< X, Y > \overset{def}{=} E\{XY\}$, and the norm of X is $E\{X^2\}$. The projection $\hat{Y}$ of Y on the space spanned by random variables $X_1, X_2, \ldots, X_n$ is a random variable such that with $\alpha_i, \hat{\alpha}_i$ denoting some scalars,

$$\hat{Y} = \sum_i \hat{\alpha}_i X_i, \quad E\{(Y-\hat{Y})^2\} = \min_\alpha E\{(Y - \sum_i \alpha_i X_i)^2\}.$$

Estimation of v_β by Discounted TD

Consider the β-discounted TD(λ) with $\lambda = 1$. Let r_β^* be the limit of the linear coefficients that TD converges to, and $\hat{v}_\beta(y,u) = \phi(y,u)' r_\beta^*$ be the corresponding function.

Proposition 10.1. *The function $\hat{v}_\beta(y,u)$ is a projection of $v_\beta(y,u)$ on the column space of Φ, i.e.,*

$$E_0 \left\{ \left(v_\beta(Y,U) - \phi(Y,U)' r_\beta^* \right)^2 \right\} = \min_{r \in \Re^k} E_0 \left\{ \left(v_\beta(Y,U) - \phi(Y,U)' r \right)^2 \right\}.$$

Prop. 10.1 follows from results on discounted TD and the next simple lemma, which follows from the fact that $v_\beta(y,u)$ is the conditional mean of Q_β.

Lemma 10.1. *For any vector $r \in \Re^k$,*

$$\begin{aligned} E_0 \left\{ \left(Q_\beta(X,Y,U) - \phi(Y,U)' r \right)^2 \right\} &= E_0 \left\{ Q_\beta(X,Y,U)^2 - v_\beta(Y,U)^2 \right\} \\ &\quad + E_0 \left\{ \left(v_\beta(Y,U) - \phi(Y,U)' r \right)^2 \right\}. \qquad (10.7) \end{aligned}$$

Proof of Prop. 10.1: Since $\lambda = 1$, by Proposition 6.5 in [BT96] (pp. 305), the function $\tilde{\phi}(x,y,u)' r_\beta^*$ is the projection of $Q_\beta(x,y,u)$ on the feature space with respect to the equilibrium distribution, i.e., r_β^* minimizes

$$E_0 \left\{ \left(Q_\beta(X,Y,U) - \tilde{\phi}(X,Y,U)' r \right)^2 \right\} = E_0 \left\{ \left(Q_\beta(X,Y,U) - \phi(Y,U)' r \right)^2 \right\}.$$

Hence r_β^* minimizes $E_0 \left\{ \left(v_\beta(Y,U) - \phi(Y,U)' r \right)^2 \right\}$ by Lemma 10.1. □

The error analysis for the case of $\lambda < 1$, omitted here, is similar to and less complicated than the case of average cost TD(λ) as shown next.

Estimation of v by Average Cost TD

Consider the average cost TD(λ) with $\lambda < 1$.[6] Let r_λ^* be the limit of the linear coefficients that TD converges to, and $\hat{v}(y,u) = \phi(y,u)' r_\lambda^*$ be the corresponding function. The next proposition says that modulo a constant translation, $\hat{v}$ is an approximation to the projection of v on the feature space, and converges to this projection when $\lambda \uparrow 1$.

Proposition 10.2. *There exists a constant scalar $\bar{c}$ such that*

$$\begin{aligned} E_0 \left\{ \left(v(Y,U) + \bar{c} - \phi(Y,U)' r_\lambda^* \right)^2 \right\} &\leq \tfrac{\alpha_\lambda^2}{1-\alpha_\lambda^2} E_0 \left\{ Q(X,Y,U)^2 - v(Y,U)^2 \right\} \\ &\quad + \tfrac{1}{1-\alpha_\lambda^2} \inf_{r \in \Re^k} \inf_{c \in \Re} E_0 \left\{ \left(v(Y,U) + c - \phi(Y,U)' r \right)^2 \right\}, \end{aligned}$$

where $\alpha_\lambda \in [0,1)$ is a mixing factor, depending on the Markov chain, with $\lim_{\lambda \uparrow 1} \alpha_\lambda = 0$.

By Prop. 10.2, the approximation error, measured in the squared norm, is bounded by two terms. The second term is a multiple of the best approximation error possible, and is

[6] We assume that the column space of Φ does not contain the vector $[1 \dots 1]'$, to satisfy a condition in average cost TD algorithms.

zero when $v(y,u)$, modulo a constant translation, is in the feature space. The first term, vanishing as $\lambda \uparrow 1$, can be equivalently written as a multiple of the expectation of the variance of $Q(X,Y,U)$ conditioned on (Y,U):

$$\tfrac{\alpha_\lambda^2}{1-\alpha_\lambda^2}\, E_0 \left\{ Var\left\{ Q(X,Y,U) \mid Y,U \right\} \right\}.$$

It does not depend on the features, and is a penalty for not observing states X.

The proof is a straightforward combination of the results for average cost TD (Theorem 3 of [TV99]) and a decomposition of error by Lemma 10.1.

Proof: Let r_λ^* be the linear coefficients that TD converges to. Consider the Markov chain $\{(X_t, Y_t, U_t)\}$ whose equilibrium distribution we denote by $\pi(x,y,u)$. By the Bellman equation for average cost and the definition of $Q(x,y,u)$, we have $Q(x,y,u) = h(x,y,u) - \eta$, where $h(x,y,u)$ is the bias, and by definition satisfies $E_0\{h(X,Y,U)\} = 0$. For TD in the imaginary process, by Theorem 3 of [TV99],

$$E_0 \left\{ \left(h(X,Y,U) - \tilde{\psi}(X,Y,U)' r_\lambda^* \right)^2 \right\} \le \frac{1}{1-\alpha_\lambda^2} E_0 \left\{ \left(h(X,Y,U) - \tilde{\psi}(X,Y,U)' r^* \right)^2 \right\}, \tag{10.8}$$

where $\alpha_\lambda \in [0,1)$ is a mixing factor depending on the Markov chain, with $\lim_{\lambda \uparrow 1} \alpha_\lambda = 0$; $\tilde{\psi}(x,y,u)$ is defined by,

$$\tilde{\psi}(x,y,u) = \tilde{\phi}(x,y,u) - \sum_{\bar{x},\bar{y},\bar{u}} \pi(\bar{x},\bar{y},\bar{u}) \tilde{\phi}(\bar{x},\bar{y},\bar{u});$$

and r^* is defined by

$$r^* = \arg\min_{r \in \mathcal{R}^k} E_0 \left\{ \left(h(X,Y,U) - \tilde{\psi}(X,Y,U)' r \right)^2 \right\}.$$

The definition of $\tilde{\psi}$ is to make the measure of approximation error insensitive to a constant translation in the estimated function, and it satisfies that for any r

$$E_0 \left\{ \tilde{\psi}(X,Y,U)' r \right\} = c(r) + E_0 \left\{ \tilde{\phi}(X,Y,U)' r \right\} = 0,$$

where $c(r)$ is a constant depending on r. The expectation term in the right-hand side of Eq. (10.8) can be equivalently written as

$$\begin{aligned} E_0 \left\{ \left(h(X,Y,U) - \tilde{\psi}(X,Y,U)' r^* \right)^2 \right\} &= \inf_{c \in \mathcal{R}} E_0 \left\{ \left(h(X,Y,U) + c - \tilde{\phi}(X,Y,U)' r^* \right)^2 \right\} \\ &= \inf_{r \in \mathcal{R}^k} \inf_{c \in \mathcal{R}} E_0 \left\{ \left(h(X,Y,U) + c - \tilde{\phi}(X,Y,U)' r \right)^2 \right\}. \end{aligned} \tag{10.9}$$

Now we bound approximation error for $\hat{v}$. Note Lemma 10.1 holds if we replace Q_β by $Q - \eta$ and v_β by $v - \eta$, because Q and v have the same mean η. Note also that $\tilde{\psi}(x,y,u)$

does not depend on the state x. Hence define

$$\psi(y,u) = \tilde{\psi}(x,y,u),$$

apply Lemma 10.1 to both sides of Eq. (10.8) separately, and after rearranging terms, it follows that

$$\begin{aligned} E_0\left\{\left(v(Y,U) - \eta - \psi(Y,U)'r_\lambda^*\right)^2\right\} \leq \frac{1}{1-\alpha_\lambda^2}\, & E_0\left\{\left(v(Y,U) - \eta - \psi(Y,U)'r^*\right)^2\right\} \\ & + \frac{\alpha_\lambda^2}{1-\alpha_\lambda^2}\, E_0\left\{Q(X,Y,U)^2 - v(Y,U)^2\right\}, \end{aligned} \tag{10.10}$$

where we have also used the fact that

$$E_0\left\{Q(X,Y,U)^2 - v(Y,U)^2\right\} = \; E_0\left\{(Q(X,Y,U)-\eta)^2 - (v(Y,U)-\eta)^2\right\}.$$

Similarly apply Lemma 10.1 to the definition of r^* and to Eq. (10.9), and it follows that

$$r^* = \arg\min_{r \in \mathcal{R}^k} E_0\left\{\left(v(Y,U) - \eta - \psi(Y,U)'r\right)^2\right\}, \tag{10.11}$$

$$\begin{aligned} & E_0\left\{\left(v(Y,U) - \eta - \psi(Y,U)'r^*\right)^2\right\} \\ & = \inf_{c\in\mathcal{R}} E_0\left\{\left(v(Y,U) - \eta + c - \phi(Y,U)'r^*\right)^2\right\} \\ & = \inf_{r\in\mathcal{R}^k}\inf_{c\in\mathcal{R}} E_0\left\{\left(v(Y,U) - \eta + c - \phi(Y,U)'r\right)^2\right\}. \end{aligned} \tag{10.12}$$

Putting Eq. (10.10)-(10.12) together, we have the claim proved. □

10.3 Gradient Estimation for Finite-State Controllers

The graphical model of a POMDP with a finite-state controller is shown in Fig. 10-2. The controller has an internal state, denoted by Z_t, taking a finite number of values. Given the observation Y_t, the controller applies the control U_t with probability $p(U_t|Z_t,Y_t)$, and its internal state subsequently transits to Z_{t+1} with probability $p(Z_{t+1}|Z_t,Y_t,U_t)$.[7] The process $\{(X_t,Y_t,Z_t,U_t)\}$ jointly forms a Markov chain, and so does the marginal process $\{(X_t,Y_t,Z_t)\}$.

Let $\{\gamma_\theta \mid \theta \in \Theta\}$ be a parametrized family of finite-state controllers with the same internal state space. For any policy γ_θ, let

$$\mu_u(z,y,\theta) = p(U_t = u \mid Z_t = z, Y_t = y; \theta)$$

be the probability of taking control u at internal state z and observation y, and let

$$\zeta_{\bar{z}}(z,y,u,\theta) = p(Z_{t+1} = \bar{z} \mid Z_t = z, Y_t = y, U_t = u; \theta)$$

[7] One can define a finite-state controller different from the one we use here. For example, the internal state transits to Z_{t+1} with probability $p(Z_{t+1}|Z_t,U_t,Y_{t+1})$, i.e., the transition depends on Y_{t+1}, instead of Y_t. The general idea outlined in Sec. 10.2 applies in the same way. The equations will be different from the ones in this section, however.

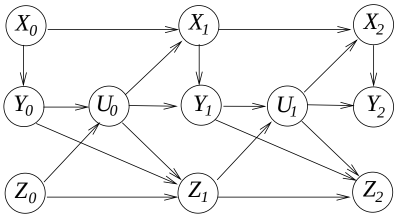

Figure 10-2: A POMDP with a finite-state controller. The states Z_t are the internal states of the controller.

be the transition probability of the internal states. We require that $\mu_u(z, y, \theta)$ and $\zeta_{\bar{z}}(z, y, u, \theta)$ are differentiable for any give $u, z, y, \bar{z}$, and the transition probability $p(X_{t+1} = \bar{x}, Y_{t+1} = \bar{y}, Z_{t+1} = \bar{z} \mid X_t = x, Y_t = y, Z_t = z; \theta)$ is differentiable for any given $x, y, z, \bar{x}, \bar{y}, \bar{z}$. Similar to the case of reactive policies, we assume that for all $\theta \in \Theta$, the Markov chains $\{(X_t, Y_t, Z_t)\}$ can be defined on a common state space, and furthermore we make the following additional assumptions.

Assumption 10.3. *Under any policy γ_θ, the Markov chain $\{(X_t, Y_t, Z_t)\}$ is irreducible and aperiodic.*

Assumption 10.4. *There exists a constant L, such that for all $\theta \in \Theta$*

$$\max_{u,y} \left\| \tfrac{\nabla \mu_u(y,\theta)}{\mu_u(y,\theta)} \right\| \leq L, \quad \max_{u,y,z,\bar{z}} \left\| \tfrac{\nabla \zeta_{\bar{z}}(z,y,u,\theta)}{\zeta_{\bar{z}}(z,y,u,\theta)} \right\| \leq L,$$

where $0/0$ is regarded as 0.

Under these assumption, for all θ, the average cost is constant and differentiable, (for a proof, see Appendix C). There is a unique equilibrium distribution for the Markov chain $\{(X_t, Y_t, Z_t, U_t)\}$. We will use the symbol E_0 to denote expectation with respect to the joint distribution of the random variables $\{(X_t, Y_t, Z_t, U_t)\}$ with the initial distribution of (X_0, Y_0, Z_0, U_0) being the equilibrium distribution.

10.3.1 Gradient Estimation

By differentiating both sides of the optimality equation, it can be shown that the gradient equals the sum of two terms:

$$\nabla \eta = E_0 \left\{ \tfrac{\nabla \mu_{U_0}(Z_0,Y_0,\theta)}{\mu_{U_0}(Z_0,Y_0,\theta)} Q(X_0, Y_0, Z_0, U_0) \right\} + E_0 \left\{ \tfrac{\nabla \zeta_{Z_1}(Z_0,Y_0,U_0,\theta)}{\zeta_{Z_1}(Z_0,Y_0,U_0,\theta)} h(X_1, Y_1, Z_1) \right\}, \tag{10.13}$$

where the Q-function $Q(x, y, u)$ is defined by

$$Q(x, y, z, u) = g(x, y, u) + E \left\{ h(X_1, Y_1, Z_1) \mid (X_0, Y_0, Z_0, U_0) = (x, y, z, u) \right\},$$

and $h(\cdot)$ is the bias function of policy γ_θ.

To estimate the first term of the r.h.s. of Eq. (10.13), we can write the term as

$$E_0 \left\{ \tfrac{\nabla \mu_{U_0}(Z_0,Y_0,\theta)}{\mu_{U_0}(Z_0,Y_0,\theta)} v_1(Y_0, Z_0, U_0) \right\}, \tag{10.14}$$

where

$$v_1(Y_0, Z_0, U_0) = E_0 \left\{ Q(X_0, Y_0, Z_0, U_0) \mid Y_0, Z_0, U_0 \right\}. \tag{10.15}$$

For estimating v_1, consider the Markov chain $\{(X_t, Y_t, Z_t)\}$, and apply β-discounted or average cost TD algorithms with the features $\tilde{\phi}(x, y, z, u) = \phi(y, z, u)$ not depending on x.

To estimate the second term of the r.h.s. of Eq. (10.13), we first note the relation between the bias function $h(x, y, z)$ of the Markov chain $\{(X_t, Y_t, Z_t)\}$ and the bias function $\tilde{h}(x, y, z, u, \bar{z})$ of the Markov chain $\{(X_t, Y_t, Z_t, U_t, Z_{t+1})\}$ with its per-stage cost also being $g(X_t, Y_t, Z_t)$:

$$h(x, y, z) = E\left\{ \tilde{h}(X_0, Y_0, Z_0, U_0, Z_1) \mid (X_0, Y_0, Z_0) = (x, y, z) \right\},$$

which can be verified from the optimality equations of the two Markov chains. It follows that

$$\begin{aligned} E\left\{ h(X_1, Y_1, Z_1) \mid X_0, Y_0, Z_0, U_0, Z_1 \right\} &= E\left\{ \tilde{h}(X_1, Y_1, Z_1, U_1, Z_2) \mid X_0, Y_0, Z_0, U_0, Z_1 \right\} \\ &= \tilde{h}(X_0, Y_0, Z_0, U_0, Z_1) + \eta - g(X_0, Y_0, U_0), \end{aligned}$$

where the second inequality follows from the optimality equation for the Markov chain $\{(X_t, Y_t, Z_t, U_t, Z_{t+1})\}$. The term $\eta - g(X_0, Y_0, U_0)$ is not a function of Z_1, and it can be dropped in gradient estimation, because

$$E_0\left\{ \frac{\nabla \zeta_{Z_1}(Z_0, Y_0, U_0, \theta)}{\zeta_{Z_1}(Z_0, Y_0, U_0, \theta)} \,\middle|\, X_0, Y_0, U_0, Z_0 \right\} = 0,$$

which implies

$$E_0\left\{ \frac{\nabla \zeta_{Z_1}(Z_0, Y_0, U_0, \theta)}{\zeta_{Z_1}(Z_0, Y_0, U_0, \theta)} \left(\eta - g(X_0, Y_0, U_0) \right) \right\} = 0.$$

Hence the second term in the gradient expression (10.13) equals

$$E_0\left\{ \frac{\nabla \zeta_{Z_1}(Z_0, Y_0, U_0, \theta)}{\zeta_{Z_1}(Z_0, Y_0, U_0, \theta)} \tilde{h}(X_0, Y_0, Z_0, U_0, Z_1) \right\} = E_0\left\{ \frac{\nabla \zeta_{Z_1}(Z_0, Y_0, U_0, \theta)}{\zeta_{Z_1}(Z_0, Y_0, U_0, \theta)} v_2(Y_0, Z_0, U_0, Z_1) \right\}, \tag{10.16}$$

where

$$v_2(Y_0, Z_0, U_0, Z_1) = E_0\left\{ \tilde{h}(X_0, Y_0, Z_0, U_0, Z_1) \mid Y_0, Z_0, U_0, Z_1 \right\}. \tag{10.17}$$

For estimating v_2, consider the Markov chain $\{(X_t, Y_t, Z_t, U_t, Z_{t+1})\}$, and apply the TD algorithms with the features $\tilde{\phi}(x, y, z, u, \bar{z}) = \phi(y, z, u, \bar{z})$ not depending on x. The line of error analysis in Sec. 10.2.3 applies in the same way here.

10.4 Gradient Estimation for POSMDPs with Structured Policies

Recall that POSMDPs as semi-Markov decision processes (SMDPs) with hidden states and observations generated by states. Recall that the model of an SMDP is the same as an MDP except that the time interval $\tau_{n+1} - \tau_n$, called the *sojourn time*, between transition from state X_n at time τ_n to state X_{n+1} at time τ_{n+1}, is random, and depends on X_n, X_{n+1} and the applied control U_n. The random variables $\{\tau_n\}$, called *decision epochs*, are the only times when controls can be applied. For details of SMDPs, see [Put94].

We consider the problem of a POSMDP with a *subset* of finite state controllers that take the observations, but *not* the sojourn times, as inputs. This is to preserve the SMDP structure of the joint process $\{(X_n, Y_n, Z_n, U_n)\}$ and the marginal process $\{(X_n, Y_n, Z_n)\}$. Singh, Tadic and Doucet [STD02] gave a GPOMDP type gradient estimation algorithm for this problem. We would like to point out that the function approximation approach applies as well. The details are as follows.

The average cost is defined as the limit of the expected cost up to time T divided by T, and under the irreducibility condition of the Markov chain $\{(X_n, Y_n, Z_n)\}$, by ergodicity the average cost equals to

$$\eta = \frac{E_0\{g(X_0, Y_0, U_0)\}}{E_0\{\tau_1\}},$$

where $g(x, y, u)$ is the mean of the random per-stage cost $c(x, y, \tau, u)$ that depends on the sojourn time τ. In the case of reactive policies, one can show that the gradient equals to

$$\nabla\eta = \frac{E_0\{\frac{\nabla\mu_{U_0}(Y_0,\theta)}{\mu_{U_0}(Y_0,\theta)} h(X, Y, U)\}}{E_0\{\tau_1\}},$$

where h satisfies the equation

$$h(x, y, u) = g(x, y, u) - \bar{\tau}(x, y, u)\,\eta + E\{h(X_1, Y_1, U_1) \mid (X_0, Y_0, U_0) = (x, y, u)\},$$

and $\bar{\tau}(x, y, u)$ is the expected sojourn time given $(X_0, Y_0, U_0) = (x, y, u)$.

Now notice that h is the bias function of the Markov chain $\{(X_n, Y_n, U_n)\}$ with $g(x, y, u) - \bar{\tau}(x, y, u)\,\eta$ as the expected per-stage cost, or equivalently with $c(X, Y, \tau, U) - \tau(X, Y, U)\,\eta$ as the random per-stage cost, where τ is the random sojourn time. Let $\hat{\eta}_n$ be the online estimate of η. We can thus estimate the projection of h (equivalently the conditional mean of h) by running TD algorithms (discounted or average cost version) in this MDP with per-stage cost $g_n - (\tau_{n+1} - \tau_n)\hat{\eta}_n$, and with features not depending on state x and sojourn time τ.

The general case of finite-state controllers is similar: the gradient is equal to the sum of two parts, each of which can be estimated using function approximation by considering the appropriate Markov chain – the same as in a POMDP – with per-stage cost $g_n - (\tau_{n+1} - \tau_n)\hat{\eta}_n$.

10.5 Experiments

We test GPOMDP and our method on a medium size ALOHA problem – a communication problem — with 30 states, 3 observations and 9 actions.[8] We take its model from A. R. Cassandra's POMDP data repertoire (on the web), and define per-stage costs to be the negative rewards. The true gradients and average costs in comparison are computed using the model. The family of policies we used has 3 internal states, 72 action parameters governing the randomized control probabilities $\mu_u(z, y, \theta)$, and 1 internal-transition parameter governing the transition probabilities of the internal states $\zeta_{\bar{z}}(z, y, u, \theta)$.[9] The parameters are bounded so that all the probabilities are in the interval $[0.001, 0.999]$. For experiments reported below, $\beta = 0.9, \lambda = 0.9$.

[8] In this problem, a state generates the same observation under all actions, and for each observation, the number of states that can generate it is 10.

[9] The internal-transitions are made such that the internal-state functions as a memory of the past, and the parameter is the probability of remembering the previous internal-state, with 1 minus the parameter being the probability of refreshing the internal state by the recent observation.

B-TD	**OL-TD**	**GPOMDP**
0.9678 ± 0.0089	0.875 ± 0.006	0.9680 ± 0.0088

Table 10.1: Comparison of gradient estimators. The number $\frac{\hat{\nabla}\eta' \nabla\eta}{\|\hat{\nabla}\eta\|_2 \|\nabla\eta\|_2}$ when θ is far away from a local minimum.

We demonstrate below the behavior of gradient estimators in two typical situations: when the magnitude of the true gradient is large, and when it is small. Correspondingly they can happen when the policy parameter is far away from a local minimum, and when it is close to a local minimum (or local maximum).

First we describe how the local minimum was found, which also shows that the approach of finite-state controller with policy gradient is quite effective for this problem. The initial policy has equal action probabilities for all internal-state and observation pairs, and has 0.2 as the internal-transition parameter. At each iteration, the gradient is estimated from a simulated sample trajectory of length 20000 (a moderate number for the size of this problem), without using any estimates from previous iterations. We then, denoting the estimate by $\hat{\nabla}\eta$, project $-\hat{\nabla}\eta$ to the feasible direction set, and update the policy parameter by a small constant step along the projected direction. We used GPOMDP in this procedure, (mainly because it needs less computation). The initial policy has average cost -0.234. The cost monotonically decreases, and within 4000 iterations the policy gets into the neighborhood of a local minimum, oscillating around afterwards, with average costs in the interval $[-0.366, -0.361]$ for the last 300 iterations. As a comparison, the optimal (liminf) average cost of this POMDP is bounded below by -0.460, which is computed using an approximation scheme from [YB04].

Table 10.1 lists the number $\frac{\hat{\nabla}\eta' \nabla\eta}{\|\hat{\nabla}\eta\|_2 \|\nabla\eta\|_2}$ for several gradient estimators, when the policy is far from a local minimum. The values listed are the means and standard deviations calculated from 5 sample trajectories simulated under the same policy. In the first column, the gradient estimator (B-TD) uses the batch estimate of the value function, that is, it uses the function estimated by TD at the end of a trajectory. (The use of batch data does not mean that the algorithm is offline. In fact, the estimator (B-TD) can be implemented online.) In the second column, the gradient estimator (OL-TD) uses the on-line estimates of the value function computed by TD. The TD algorithms we used are β-discounted LSPE(λ) [BBN03] and average cost LSPE(λ). The difference between the discounted and average cost TD turns out negligible in this experiment. In the third column, we use GPOMDP.[10] The estimates from B-TD and GPOMDP align well with the true gradient, while OL-TD is not as good, due to the poor estimates of TD in the early period of a trajectory.

Fig. 10-3 shows the number $\frac{\hat{\nabla}\eta' \nabla\eta}{\|\hat{\nabla}\eta\|_2 \|\nabla\eta\|_2}$ for several gradient estimators on 20 sample trajectories simulated under the same policy, when that policy is near a local minimum.[11] The horizontal axis indexes the trajectories. The blue solid line and the green dash-dot line correspond, respectively, to the gradient estimator that uses the batch estimate (B-TD) and the on-line estimate (OL-TD) of the value function, computed by β-discounted LSPE(λ).

[10] For both GPOMDP and the discounted TD algorithm, we subtracted the per-stage cost by the on-line estimate of the average cost.

[11] More precisely, the number we compute here is the inner-product of the *projections* of $-\nabla\eta$ and $-\hat{\nabla}\eta$ (on the set of feasible directions) normalized by their norms.

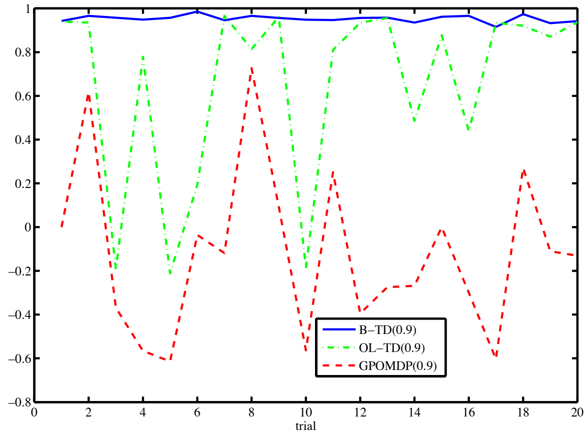

Figure 10-3: Comparison of gradient estimators. The number $\frac{\hat{\nabla}\eta'\nabla\eta}{\|\hat{\nabla}\eta\|_2\|\nabla\eta\|_2}$ when θ is near a local minimum. Linear interpolations between trials are plotted for reading convenience.

The red dash line corresponds to GPOMDP. While the estimator B-TD consistently aligns well with the true gradient, GPOMDP often points to the opposite direction.

Our experiments demonstrate that when close to a local minimum (or local maximum), where the magnitude of the gradient is small, in order to align with the gradient, the estimator needs to have much smaller bias and variance. In GPOMDP we only have one parameter β to balance the bias-variance. Hence it can be advantageous for the function approximation approach to provide more options – namely the feature space, λ and β – in controlling bias and variance in gradient estimation.

10.6 Summary

We have shown that Actor-Critic methods are alternatives to GPOMDP in learning finite-state controllers for POMDPs and POSMDPs. Actor-Critic methods provide more options in bias-variance control than GPOMDP. It is unclear, however, both theoretically or practically, which method is most efficient: actor-only, actor-critic, or their combined variants as suggested in [Kon02]. We also note that using a value function in gradient estimation can be viewed as a variance reduction technique based on Rao-Blackwellization. The control variate idea [GBB04] is a different type of variance reduction technique, and applies to both actor-only and actor-critic algorithms.

Chapter 11

Some Convergence Results on the LSPE Algorithm

In this chapter we consider finite space MDPs and prove two convergence results for a least squares policy evaluation algorithm, called LSPE, first proposed by Bertsekas and Ioffe [BI96]. LSPE(λ) is one of the iterative TD(λ) algorithms that evaluate the cost of a policy from a sample trajectory and approximate the cost function by linear function approximation. Nedić and Bertsekas [NB03] proved the convergence of LSPE with a diminishing stepsize. Recently Bertsekas et al. [BBN03] have shown that for discounted problems the LSPE algorithm with a constant stepsize converges.

In this chapter we will show the following results:

- the average cost LSPE(λ) algorithm with a constant stepsize $\gamma \leq 1$ converges to the same limit as average cost TD(λ) algorithms; and

- for both discounted and average cost cases, LSPE(λ) with any constant stepsize (under which LSPE converges) has the same convergence rate as LSTD(λ).

The LSTD algorithm is another least squares type algorithm, first proposed by Bradtke and Barto [BB96] for $\lambda = 0$ and extended by Boyan [Boy99] to $\lambda \in [0, 1]$. It is proved by Konda [Kon02] that LSTD(λ) has the optimal asymptotic convergence rate compared to other TD(λ) algorithms. Thus our result shows that LSPE(λ) with a constant stepsize has the optimal asymptotic convergence rate.

11.1 Introduction

First we establish some notation. We have a finite state and control MDP and a stationary policy, which induces a Markov chain in the MDP. For this Markov chain, let P be the state transition probability matrix with entries $P_{ij} = p(X_1 = j | X_0 = i)$. Let n be the number of states, and e the length-n column vector of all 1s.

We assume that the Markov chain is *recurrent*. Thus the Markov chain has a unique equilibrium distribution, denoted by π and satisfying

$$\sum_{i=1}^{n} \pi(i) P_{ij} = \pi(j), \quad \forall j,$$

or in matrix notation, defining $\pi = [\pi(1), \ldots, \pi(n)]$,

$$\pi P = \pi.$$

By ergodicity, the average cost η^* is a constant independent of the initial state, and

$$\eta^* = \pi \bar{g},$$

where $\bar{g}$ is a column vector of expected per-stage cost at every state. The average cost DP equation in matrix notation is

$$h = \bar{g} - \eta^* e + Ph, \tag{11.1}$$

where h, a length-n vector, is the bias function.

In least squares TD algorithms with linear function approximation, using data from sample trajectories, we approximate h by $\Phi\, r$ with a given matrix Φ, and solve various least squares problems to obtain the vector r. In particular, let

$$\phi(x)' = \big[\phi_1(x), \ldots, \phi_m(x)\big]$$

be a length-m vector with its entries called features of state x, and let Φ be the $n \times m$ matrix

$$\Phi = \begin{bmatrix} \vdots \\ \phi(x)' \\ \vdots \end{bmatrix}, \tag{11.2}$$

with rows $\phi(x)'$ and with linearly independent columns, called basis functions.

11.1.1 The Average Cost LSPE and LSTD Updates

Let $(x_0, u_0, x_1, u_1, \ldots)$ be an infinitely long sample trajectory where x_k, u_k are the state and control at time k. Let η_k be the on-line estimate of the average cost at time k:

$$\eta_t = \frac{1}{t+1} \sum_{i=0}^{t} g(x_i, u_i).$$

By ergodicity, we have $\eta_t \to \eta^*$ with probability 1, where η^* is the average cost.

The average cost LSPE(λ) algorithm with a constant stepsize γ updates the vector r_t by

$$r_{t+1} = r_t + \gamma \bar{B}_t^{-1} \left(\bar{A}_t\, r_t + \bar{b}_t \right). \tag{11.3}$$

Here,

$$\bar{B}_t = \frac{B_t}{t+1}, \quad \bar{A}_t = \frac{A_t}{t+1}, \quad \bar{b}_t = \frac{b_t}{t+1},$$

are matrices B_t, A_t and vector b_t, respectively, averaged over time, which are defined as[1]

$$B_t = \sum_{k=0}^{t} \phi(x_k)\, \phi(x_k)', \qquad A_t = \sum_{k=0}^{t} z_k \left(\phi(x_k+1)' - \phi(x_k)'\right),$$

$$b_t = \sum_{k=0}^{t} z_k \,(g(x_k, u_k) - \eta_k), \qquad z_k = \sum_{m=0}^{k} \lambda^{k-m} \phi(x_m).$$

Using the analysis of Tsitsiklis and Van Roy [TV99] on average cost TD algorithms and Nedić and Bertsekas [NB03] on discounted LSPE algorithms, it can be easily shown that with probability one

$$\bar{B}_t \to B, \quad \bar{A}_t \to A, \quad \bar{b}_t \to b,$$

where

$$B = \Phi' D \Phi, \qquad A = \Phi' D (I - \lambda P)^{-1} (P - I) \Phi,$$
$$b = \Phi' D (I - \lambda P)^{-1} (\bar{g} - \eta^* e),$$

and D is the diagonal matrix

$$D = diag(\ldots, \pi(x), \ldots)$$

with the diagonal entries $\pi(x)$.

A different least squares TD algorithm, the average cost LSTD(λ), updates the vector r_t by

$$\hat{r}_{t+1} = -\bar{A}_t^{-1}\, \bar{b}_t. \tag{11.4}$$

So with probability one $\hat{r}_t$ converges to $-A^{-1}b$, the same limit as other TD(λ) algorithms. The error of cost approximation, which depends on λ and the space spanned by the basis functions, is analyzed by [TV99].

11.1.2 The Corresponding Least Squares Problems

The LSPE(λ) and LSTD(λ) updates are related to least squares solutions, which we are now going to describe. Iteratively expanding the right-hand side of the average cost DP equation

$$h = \bar{g} - \eta^* e + Ph$$

[1] As a comparison, for the β-discounted criterion the update rule of LSPE(λ), $\lambda \in [0,1]$, is defined by Eq. (11.3) with the corresponding matrices

$$B_t = \sum_{k=0}^{t} \phi(x_k)\, \phi(x_k)', \qquad A_t = \sum_{k=0}^{t} z_k \left(\beta\phi(x_k+1)' - \phi(x_k)'\right),$$

$$b_t = \sum_{k=0}^{t} z_k \, g(x_k, u_k), \qquad z_k = \sum_{m=0}^{k} (\beta\lambda)^{k-m} \phi(x_m).$$

The matrix A_t and vector b_t converge to A and b respectively, with

$$A = \Phi' D (I - \lambda\beta P)^{-1} (\beta P - I)\Phi, \qquad b = \Phi' D (I - \lambda\beta P)^{-1} \bar{g}.$$

for m-steps, we have the multiple-step value iteration equation

$$h = \sum_{k=0}^{m} P^k(\bar{g} - \eta^* e) + P^{m+1}h. \tag{11.5}$$

For $\lambda \in [0, 1)$, define a series of scalars $(1 - \lambda)\lambda^m$, which sum to 1. Multiplying Eq. (11.5) by $(1 - \lambda)\lambda^m$ and summing over m, it follows that

$$\begin{aligned}
h &= (1-\lambda)\sum_{m=0}^{\infty}\lambda^m \left(\sum_{k=0}^{m} P^k(\bar{g} - \eta^* e) + P^{m+1}h\right) \\
&= \sum_{k=0}^{\infty} P^k(\bar{g} - \eta^* e)\left((1-\lambda)\sum_{m=k}^{\infty}\lambda^m\right) + (1-\lambda)\sum_{m=0}^{\infty}\lambda^m P^{m+1}h \\
&= \sum_{k=0}^{\infty}\lambda^k P^k(\bar{g} - \eta^* e) + \sum_{k=0}^{\infty}\lambda^k P^{k+1}h - \sum_{k=0}^{\infty}\lambda^k P^k h + h \\
&= \sum_{k=0}^{\infty}\lambda^k P^k(\bar{g} - \eta^* e) + \sum_{k=0}^{\infty}\lambda^k\left(P^{k+1}h - P^k h\right) + h \\
&= h + \sum_{k=0}^{\infty}\lambda^k\big(P^k(\bar{g} - \eta^* e) + P^{k+1}h - P^k h\big).
\end{aligned}$$

Equivalently, we can write this last equation in the expectation notation. For every state x,

$$\begin{aligned}
h(x) &= h(x) + \sum_{k=0}^{\infty}\lambda^k E\big\{g(X_k, U_k) - \eta^* + h(X_{k+1}) - h(X_k) \,|\, X_0 = x\big\} \\
&= h(x) + E\left\{\sum_{k=0}^{\infty}\lambda^k\Big(g(X_k, U_k) - \eta^* + h(X_{k+1}) - h(X_k)\Big) \,|\, X_0 = x\right\}. \qquad (11.6)
\end{aligned}$$

Equation (11.6) relates to the least squares problems associated with LSPE and LSTD. To determine the linear coefficients r_{t+1} in the approximation

$$h \approx \Phi\, r_{t+1},$$

the two methods use different approximations for the left and right-hand sides of Eq. (11.6), and correspondingly solve two different least squares problems of minimizing the difference between the approximations of the two sides.

For LSPE, the h on the left-hand side of Eq. (11.6) is approximated by $h_{t+1}(x) = \phi(x)'\bar{r}_{t+1}$, which is to be determined, while the h on the right-hand side is approximated by $h_t(x) = \phi(x)'r_t$, which is fixed. The least squares problem to determine $\bar{r}_{t+1}$ is

$$\frac{1}{t+1}\sum_{k=0}^{t}\left(h_{t+1}(x_k) - h_t(x_k) - \sum_{m=k}^{t}\lambda^{m-k}\big(g(x_m, u_m) - \eta_m + h_t(x_{m+1}) - h_t(x_m)\big)\right)^2. \tag{11.7}$$

As can be seen from the expression of (11.7), the sample value from every partial trajectory up to time t has replaced the expectation term in the right-hand side of Eq. (11.6). The squared difference between the left-hand side of Eq. (11.6) approximated by h_{t+1}, and the right-hand side approximated by this sample value and h_t, is then summed up over all partial trajectories up to time t to form the least squares problem (11.7) of LSPE(λ). The minimum of the least squares problem is

$$\bar{r}_{t+1} = \big(I + \bar{B}_t^{-1} \bar{A}_t\big)\, r_t + \bar{B}_t^{-1}\, \bar{b}_t.$$

The update rule (11.3) with a constant stepsize γ is equivalent to

$$r_{t+1} = (1-\gamma)\, r_t + \gamma\, \bar{r}_t,$$

and when $\gamma = 1$, $r_{t+1} = \bar{r}_{t+1}$.

For LSTD, the h functions of both sides of Eq. (11.6) are approximated by $\hat{h}_{t+1} = \Phi\, \hat{r}_{t+1}$, which is to be determined. The least squares problem is

$$\frac{1}{t+1} \sum_{k=0}^{t} \left(\hat{h}_{t+1}(x_k) - \hat{h}_{t+1}(x_k) - \sum_{m=k}^{t} \lambda^{m-k} \big(g(x_m, u_m) - \eta_m + \hat{h}_{t+1}(x_{m+1}) - \hat{h}_{t+1}(x_m) \big) \right)^2 . \tag{11.8}$$

As can be seen, similar to LSPE, the sample value from every partial trajectory has been used to replace the expectation term in Eq. (11.6). The minimum solution of the least squares problem (11.8) gives the update rule (11.4) of LSTD.

11.2 Convergence of Average Cost LSPE with a Constant Stepsize

We will show the convergence of the average cost LSPE(λ) algorithm with a constant stepsize, where $\lambda < 1$. (Note that all non-episodic average cost TD algorithms must have $\lambda < 1$ to ensure that the iterates are bounded.) The limit of LSPE(λ) is the same as the one of other average cost TD(λ) algorithms. This limit and its λ-dependent distance to the projection of the bias function (modulo a constant scalar shift) on the space of the basis functions are established in [TV99]. So only the convergence of LSPE need to be shown.

The convergence of the discounted LSPE(λ) ($\lambda \in [0,1]$) with a constant stepsize $\gamma \in (0,1]$ is proved by Bertsekas et al. [BBN03]. The proof in the average cost case here is a slight modification of the proof in [BBN03]. Let $\sigma(F)$ denote the spectral radius of a square matrix F, (i.e., the maximum of the moduli of the eigenvalues of F). Key to the convergence proof is to establish that

$$\sigma(I + B^{-1}A) = \sigma(I + (\Phi' D \Phi)^{-1} A) < 1.$$

To this end, we need the following Lemma 11.1, which is a counterpart to Lemma 2.1 of [BBN03] for the discounted case.

For the remainder of this section, we make the following assumptions.

Assumption 11.1. *The Markov chain is recurrent and aperiodic.*[2]

[2] We note that the aperiodicity condition is not a limitation to the algorithm. Because, as is well-known, one can always apply an aperiodicity transformation, by introducing artificial self-transitions, to construct an aperiodic chain with the same bias function as the original Markov chain. Correspondingly, TD algorithms can be modified slightly and viewed as being applied to the transformed aperiodic chain.

Condition 11.1. *The columns of matrix* $[\Phi \; e]$ *are linearly independent.*

Let $\mathcal{C}$ denote the set of complex numbers, and $\|\cdot\|_D$ denote the weighted 2-norm:

$$\|z\|_D = \left(\sum_{i=1}^n \pi(i)\, \bar{z}_i\, z_i\right)^{\frac{1}{2}}, \quad \forall\, z = (z_1, \ldots, z_n) \in \mathcal{C}^n,$$

where $\bar{x}$ denotes the conjugate of a complex number x. By Lemma 2.1 of [BBN03] it is always true that

$$\|Pz\|_D \leq \|z\|_D, \quad \forall\, z = (z_1, \ldots, z_n) \in \mathcal{C}^n.$$

For the average cost case, we need a strict inequality.

Lemma 11.1. *For all* $z \in \mathcal{C}^n, z \notin \{c\, e \mid c \in \mathcal{C}\}$, $\|P^m z\|_D < \|z\|_D$ *for some positive integer* m.

Proof: Since $\pi P = \pi$, for any positive integer m, $\pi P^m = \pi$. Let $\tilde{P} = P^m$. We have

$$\begin{aligned}
\|\tilde{P}z\|_D^2 &= \sum_{i=1}^n \pi(i) \left(\sum_{j=1}^n \tilde{p}_{ij}\, \bar{z}_j\right) \left(\sum_{j=1}^n \tilde{p}_{ij}\, z_j\right) \\
&\leq \sum_{i=1}^n \pi(i) \left(\sum_{j=1}^n \tilde{p}_{ij}\, |z_j|\right)^2 \\
&\leq \sum_{i=1}^n \pi(i) \sum_{j=1}^n \tilde{p}_{ij}\, |z_j|^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n \pi(i)\, \tilde{p}_{ij}\, |z_j|^2 = \sum_{j=1}^n \pi(i)\, |z_j|^2 = \|z\|_D^2,
\end{aligned} \tag{11.9}$$

where the first inequality follows from the fact that

$$\bar{x}y + x\bar{y} \leq 2|x||y|, \quad \forall x, y \in \mathcal{C}, \tag{11.10}$$

and the second inequality follows from Jensen's inequality applied to the convex function $(\cdot)^2$. In view of Eq. (11.10), equality holds in the first inequality of Eq. (11.9) if and only if for all j, k such that there exists some i with $P_{ij}^m > 0, P_{ik}^m > 0$,

$$z_j = c_{jk}\, z_k, \quad \text{or} \quad |z_j|\,|z_k| = 0,$$

where c_{jk} is some real number. Since $(\cdot)^2$ is a strictly convex function, equality holds in the second inequality of Eq. (11.9) if and only if for all j, k such that there exists some i with $P_{ij}^m > 0, P_{ik}^m > 0$,

$$|z_j| = |z_k|.$$

Under Assumption 11.1, there exists a positive integer m, such that $P_{ij}^m > 0$ for all i, j. For such an m, the preceding if and only if conditions are identical to

$$z_j = z_k, \quad \forall j, k,$$

i.e., $z = c\, e$ for some $c \in \mathcal{C}$. Thus, for any $z \notin \{c\, e \mid c \in \mathcal{C}\}$, strict inequality in Eq. (11.9) must hold, and hence the claim. □

We now give the convergence proof. Most parts of the analysis follow the original analysis in [BBN03] for the discounted case by setting the discount factor to 1. At certain critical places we use Lemma 11.1. In what follows, we will avoid repeating the part of analysis identical to [BBN03], and we will point out only the differences between the two analyses.

Lemma 11.2. *The matrix $I + (\Phi' D \Phi)^{-1} A$ satisfies*

$$\sigma(I + (\Phi' D \Phi)^{-1} A) < 1.$$

Proof: Let ν be an eigenvalue of $I + (\Phi' D \Phi)^{-1} A$ and let z be the corresponding eigenvector. By the proof of Lemma 2.2 of [BBN03],

$$|\nu| \|\Phi z\|_D \leq (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \|P^{m+1} \Phi z\|_D. \tag{11.11}$$

Since $[\Phi \ e]$ has linearly independent columns by Condition 11.1,

$$\Phi z \notin \{c\, e \mid c \in \mathcal{C}\}.$$

Hence by Lemma 11.1, for some m, we have the strict inequality

$$\|P^m \Phi z\|_D < \|\Phi z\|_D.$$

Since $\|P^m \Phi z\|_D \leq \|\Phi z\|_D$ for all m, it follows from Eq. (11.11) that

$$|\nu| \|\Phi z\|_D < (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \|\Phi z\|_D = \|\Phi z\|_D,$$

which implies that $|\nu| < 1$. □

Proposition 11.1. *For average cost MDPs, LSPE(λ) with $\lambda < 1$ and a constant stepsize $\gamma \leq 1$ converges with probability 1.*

Proof: The method of proof is almost identical to that for the discounted case in [BBN03], except for a slight modification, which we describe here. By the proof of Prop. 3.1 of [BBN03], if the stepsize γ is such that

$$\sigma\left(I + \gamma (\Phi' D \Phi)^{-1} A\right) < 1,$$

then LSPE(λ) converges to $r^* = -A^{-1} b$ with probability 1. Write the matrix $I + \gamma (\Phi' D \Phi)^{-1} A$ as

$$(1 - \gamma) I + \gamma (I + (\Phi' D \Phi)^{-1} A),$$

and it can be seen that

$$\sigma\left(I + \gamma (\Phi' D \Phi)^{-1} A\right) \leq |1 - \gamma| + \gamma \sigma (I + (\Phi' D \Phi)^{-1} A).$$

Using Lemma 11.2, it follows that

$$\sigma\left(I+\gamma(\Phi'D\Phi)^{-1}A\right)<1, \quad \forall\gamma,\ 0<\gamma\leq 1,$$

and hence the convergence of LSPE holds for γ in the same range. □

11.3 Rate of Convergence of LSPE

Konda [Kon02] shows that LSTD has the optimal asymptotic convergence rate compared to other TD algorithms. In this section we prove that LSPE(λ) with any constant stepsize γ (under which LSPE converges) has the same asymptotic convergence rate as LSTD(λ). The proof applies to both discounted and average cost cases and for any value of λ ($\lambda \in [0,1]$ for the discounted case and $\lambda \in [0,1)$ for the average cost case). So we will suppress the symbol λ.

The LSPE and LSTD updates are, respectively:

$$r_{t+1}=r_t+\gamma\bar{B}_t^{-1}\left(\bar{A}_t\,r_t+\bar{b}_t\right), \qquad \hat{r}_{t+1}=-\bar{A}_t^{-1}\,\bar{b}_t.$$

They converge to the same limit $r^*=-A^{-1}b$. Informally, it has been observed in [BBN03] that for various $\lambda \in [0,1]$, r_t became close to and "tracked" $\hat{r}_t$ even before the convergence to r^* took place. One intuitive explanation of this phenomenon is that as time goes by, the least squares problems solved by LSPE and LSTD differ "marginally", especially when λ is close to 1. Another explanation, given in [BBN03], is a two-time scale type of view: when t is large, $\bar{A}_t, \bar{B}_t$ and $\bar{b}_t$ change slowly so that they are essentially "frozen" at certain values, and r_t then "converges" to the unique fixed point of the linear system

$$r=r+\gamma\bar{B}_t^{-1}\left(\bar{A}_t\,r+\bar{b}_t\right),$$

which is $-\bar{A}_t^{-1}\,\bar{b}_t$, the value of $\hat{r}_t$ of LSTD.

In what follows, we will first make the above argument more precise, by showing a noisy version of a multiple-step contraction type property. It states that the distance between LSPE and LSTD shrinks geometrically outside a region, the size of which diminishes at the order of $O(1/t)$. This property will then be used to prove that LSPE has the same convergence rate as LSTD.

11.3.1 A Noisy Version of Multiple-Step Contraction

The difference between LSPE and LSTD updates can be written as

$$r_{t+1}-\hat{r}_{t+1}=\left(I+\gamma\bar{B}_t^{-1}\bar{A}_t\right)(r_t-\hat{r}_t)+\left(I+\gamma\bar{B}_t^{-1}\bar{A}_t\right)(\hat{r}_t-\hat{r}_{t+1}). \tag{11.12}$$

By a multiple-step expansion of Eq. (11.12), we have for any $k\geq 0$,

$$\begin{aligned}r_{t+k+1}-\hat{r}_{t+k+1}&=\left(I+\gamma\bar{B}_{t+k}^{-1}\bar{A}_{t+k}\right)\left(I+\gamma\bar{B}_{t+k-1}^{-1}\bar{A}_{t+k-1}\right)\cdots\left(I+\gamma\bar{B}_t^{-1}\bar{A}_t\right)(r_t-\hat{r}_t)\\&+\sum_{j=0}^{k}\left(I+\gamma\bar{B}_{t+k}^{-1}\bar{A}_{t+k}\right)\left(I+\gamma\bar{B}_{t+k-1}^{-1}\bar{A}_{t+k-1}\right)\cdots\Big(I+\gamma\bar{B}_{t+k-j}^{-1}\bar{A}_{t+k-j}\Big)(\hat{r}_{t+k-j}-\hat{r}_{t+k-j+1}).\end{aligned}$$

Taking the 2-norm of both sides, and denoting the 2-norm of the matrices by $C_{t+k,j}$:

$$C_{t+k,j} = \left\| \left(I + \gamma \bar{B}_{t+k}^{-1} \bar{A}_{t+k}\right) \left(I + \gamma \bar{B}_{t+k-1}^{-1} \bar{A}_{t+k-1}\right) \cdots \left(I + \gamma \bar{B}_{t+k-j}^{-1} \bar{A}_{t+k-j}\right) \right\|,$$

we thus have

$$\|r_{t+k+1} - \hat{r}_{t+k+1}\| \leq C_{t+k,k} \|r_t - \hat{r}_t\| + \sum_{j=0}^{k} C_{t+k,j} \|\hat{r}_{t+k-j} - \hat{r}_{t+k-j+1}\|. \tag{11.13}$$

For a fixed k, the distance terms $\|\hat{r}_{t+k-j} - \hat{r}_{t+k-j+1}\|$ in the right-hand side of Eq. (11.13) are of the order $O(1/t)$. This is stated in the following lemma, which can be easily shown.

Lemma 11.3. *Consider a sample path for which both LSTD and LSPE converge. There exist some constant C_1 and C_2 such that for all t sufficiently large,*

$$\|\bar{B}_t^{-1} \bar{A}_t\| \leq C_1, \qquad \|\hat{r}_{t+1} - \hat{r}_t\| \leq \frac{C_2}{t}.$$

Proof: Consider a sample path for which both LSTD and LSPE converge. Since $\|\bar{B}_t^{-1} \bar{A}_t\| \to \|B^{-1} A\|$, the first relation of the claim can be easily seen.

For the second relation, by definition of the LSTD updates,

$$\begin{aligned} \|\hat{r}_{t+1} - \hat{r}_t\| &= \left\|\bar{A}_t^{-1} \bar{b}_t - \bar{A}_{t-1}^{-1} \bar{b}_{t-1}\right\| \\ &\leq \left\|\bar{A}_t^{-1} - \bar{A}_{t-1}^{-1}\right\| \left\|\bar{b}_t\right\| + \left\|\bar{A}_{t-1}^{-1}\right\| \left\|\bar{b}_t - \bar{b}_{t-1}\right\|. \end{aligned} \tag{11.14}$$

Since $\left\|\bar{b}_t\right\| \to \|b\|$ and $\left\|\bar{A}_{t-1}^{-1}\right\| \to \left\|A^{-1}\right\|$, there exists some constant C and C' such that for all t sufficiently large,

$$\left\|\bar{b}_t\right\| \leq C, \qquad \left\|\bar{A}_{t-1}^{-1}\right\| \leq C'.$$

Furthermore, by the definition of $\bar{b}_t$, it can be seen that for t sufficiently large,

$$\left\|\bar{b}_t - \bar{b}_{t-1}\right\| = \frac{1}{t} \left\|z_t g(x_t, u_t)\right\| + O(1/t^2) = O(1/t),$$

(since z_t is bounded for all t). By the definition of $\bar{A}_t$ and the Sherman-Morisson formula for matrix inversion, it can be seen that

$$\begin{aligned} \left\|\bar{A}_t^{-1} - \bar{A}_{t-1}^{-1}\right\| &= \left\|(t+1) A_t^{-1} - t A_{t-1}^{-1}\right\| = \left\|A_t^{-1} + t\left(A_t^{-1} - A_{t-1}^{-1}\right)\right\| \\ &\leq \left\|A_t^{-1}\right\| + t \left\| \frac{A_{t-1}^{-1} z_t \left(\beta\phi(x_k+1)' - \phi(x_k)'\right) A_{t-1}^{-1}}{1 + \left(\beta\phi(x_k+1)' - \phi(x_k)'\right) A_{t-1}^{-1} z_t} \right\| \\ &= \frac{1}{t+1} \left\|\bar{A}_t^{-1}\right\| + \left\| \frac{\bar{A}_{t-1}^{-1} z_t \left(\beta\phi(x_k+1)' - \phi(x_k)'\right) \bar{A}_{t-1}^{-1}}{t + \left(\beta\phi(x_k+1)' - \phi(x_k)'\right) \bar{A}_{t-1}^{-1} z_t} \right\| \\ &= O(1/t) + O(1/t) = O(1/t), \end{aligned}$$

where $\beta \in [0,1]$. Plugging these relations into Eq. (11.14), it follows that $\|\hat{r}_{t+1} - \hat{r}_t\| = O(1/t)$ and the claim thus follows. □

We now state the multiple-step contraction property. First, define ρ_0 to be the spectral radius of the limiting matrix:

$$\rho_0 = \sigma(I + \gamma B^{-1}A).$$

Lemma 11.4. *For each $\rho \in (\rho_0, 1)$ and each sample path for which both LSTD and LSPE converge, there exists a positive integer k and a positive scalar sequence $\epsilon_t = C/t$, where C is some constant, such that*

1. *for t sufficiently large, whenever $\|r_t - \hat{r}_t\| \geq \epsilon_t$,*

$$\|r_{t+k} - \hat{r}_{t+k}\| \leq \rho^k \|r_t - \hat{r}_t\|;$$

2. *there exist infinitely many t with*

$$\|r_t - \hat{r}_t\| \leq \epsilon_t.$$

Proof: (Part 1): Since the spectral radius of $I + \gamma B^{-1}A$ is ρ_0, there is a matrix norm denoted by $\|\cdot\|_w$ such that $\|I + \gamma B^{-1}A\|_w = \rho_0$. Consider a sample path for which both LSTD and LSPE converge. Since $I + \gamma \bar{B}_t^{-1}\bar{A}_t \to I + \gamma B^{-1}A$, for any $\rho_1 \in (\rho_0, 1)$, there is a time N such that for all $t > N$,

$$\left\|I + \gamma \bar{B}_t^{-1}\bar{A}_t\right\|_w \leq \rho_1,$$

so that for any k

$$\left\|\prod_{i=0}^{k}(I + \gamma \bar{B}_{t+i}^{-1}\bar{A}_{t+i})\right\|_w \leq \prod_{i=0}^{k}\left\|I + \gamma \bar{B}_{t+i}^{-1}\bar{A}_{t+i}\right\|_w \leq \rho_1^{k+1}.$$

Using the fact that the matrix norm $\|\cdot\|_w$ and the matrix 2-norm are equivalent, i.e., there exists some constant C_0 such that $\|\cdot\| \leq C_0\|\cdot\|_w$, we have for all $t > N$,

$$C_{t+k,j} \leq C_0\, \rho_1^{j+1}, \quad j = 0, 1, \ldots, k.$$

Hence it follows from Eq. (11.13) that for a given k, when t is sufficiently large,

$$\begin{aligned}\|r_{t+k+1} - \hat{r}_{t+k+1}\| &\leq C_{t+k,k}\|r_t - \hat{r}_t\| + \sum_{j=0}^{k} C_{t+k,j}\, \|\hat{r}_{t+k-j} - \hat{r}_{t+k-j+1}\| \\ &\leq C_0\, \rho_1^{k+1}\|r_t - \hat{r}_t\| + C_0 \sum_{j=0}^{k} \rho_1^{j+1}\, \frac{C_1}{t} \leq C_0\, \rho_1^{k+1}\|r_t - \hat{r}_t\| + \frac{C_2}{t},\end{aligned}$$

where we have used Lemma 11.3 to bound the terms $\|\hat{r}_{t+k-j} - \hat{r}_{t+k-j+1}\|$, and C_1 and C_2 are some constants (depending on k, ρ_1, and the sample path) with $C_2 \leq \frac{C_0\, C_1 \rho_1}{1-\rho_1}$.

Now for a given $\rho \in (\rho_0, 1)$, we first choose some $\rho_1 < \rho$ in the preceding analysis, and next we choose a k such that

$$C_0\, \rho_1^{k+1} \leq \rho_2^{k+1}$$

for some $\rho_2 \in (\rho_1, \rho)$. So by the preceding analysis,

$$\|r_{t+k+1} - \hat{r}_{t+k+1}\| \leq \rho_2^{k+1}\|r_t - \hat{r}_t\| + \frac{C_2}{t}.$$

Define

$$\epsilon_t = \frac{C_2}{(\rho^{k+1} - \rho_2^{k+1})\, t}.$$

So $\epsilon_t = O(1/t) \to 0$ and for t sufficiently large, whenever $\|r_t - \hat{r}_t\| > \epsilon_t$,

$$\|r_{t+k+1} - \hat{r}_{t+k+1}\| \leq \rho^{k+1} \|r_t - \hat{r}_t\|.$$

The proof of part one is completed by redefining $(k+1)$ as k and letting $C = \frac{C_2}{\rho^{k+1} - \rho_2^{k+1}}$.

(Part 2): We prove the claim by contradiction. Suppose for all t sufficiently large, $\|r_t - \hat{r}_t\| \geq \epsilon_t$. Then by the first part of the lemma, the k-step contraction continues to happen, and we have for some fixed t_1,

$$\frac{C}{t_1 + mk} \leq \|r_{t_1+mk} - \hat{r}_{t_1+mk}\| \leq \rho^{mk} \|r_{t_1} - \hat{r}_{t_1}\|, \quad \forall m \geq 1.$$

Hence $\rho^{mk} \geq \frac{C'}{t_1+mk}, m \geq 1$, for some constant C', which is impossible when $m \to \infty$. □

Remark 11.1. (i) The spectral radius ρ_0 of the limiting matrix $I + \gamma \bar{B}^{-1} \bar{A}$ depends on λ and the stepsize γ. When $\gamma = 1$ and $\lambda \approx 1$, ρ_0 can be very small. In particular, $\rho_0 = 0$ when $\gamma = 1$ and $\lambda = 1$ (in the discounted case). Thus when $\gamma = 1$ and $\lambda \approx 1$, the differences between LSPE and LSTD are smaller than when λ is close to 0, which is consistent with what was observed in [BBN03].
(ii) Roughly speaking, contractions can start to happen when the spectral radius of the matrix $I + \gamma \bar{B}_t^{-1} \bar{A}_t$ becomes less than 1, before the spectral radius approaches ρ_0 and the matrices converge to the limiting matrix. (We provide justification for this comment: while the matrix norm in the preceding proof depends on the limiting matrix, there is an alternative proof for the lemma, albeit longer, that works with the 2-norm directly and does not rely on this matrix norm.)

11.3.2 Proof of Convergence Rate

As Konda proved in [Kon02] (Chapter 6) under certain conditions, for LSTD, $\sqrt{t}(\hat{r}_t - r^*)$ converges in distribution to $N(0, \Sigma_0)$, a Gaussian random variable with mean 0 and covariance matrix

$$\Sigma_0 = A^{-1} \Gamma (A')^{-1},$$

where Γ is the covariance matrix of the limiting Gaussian random variable to which

$$\sqrt{t}\, (\bar{A}_t r^* + \bar{b}_t)$$

converges in distribution. As Konda also proved, LSTD has the asymptotically optimal convergence rate compared to other recursive TD algorithms (the ones analyzed in [TV97] and [TV99]), in the following sense. If $\tilde{r}_t$ is computed by another recursive TD algorithm, and $\sqrt{t}(\tilde{r}_t - r^*)$ converges in distribution to $N(0, \Sigma)$, then the covariance matrix Σ of the limiting Gaussian random variable is such that

$$\Sigma - \bar{\gamma}\, \Sigma_0 : \quad \text{positive semi-definite},$$

where $\bar{\gamma}$ is some constant depending on the stepsize rule of the recursive TD algorithm.[3] The conditions assumed in Konda's analysis (see Theorem 6.1 of [Kon02]) are standard for analyzing asymptotic Gaussian approximations in stochastic approximation methods. The conditions include: boundedness of updates, geometric convergence of certain means and covariance matrices relating to the updates, and the positive definiteness of the matrix $-A-\frac{\bar{\gamma}}{2}I$. This last assumption can be satisfied by scaling the basis functions if the stepsize is chosen as $1/t$, and the rest conditions are satisfied by TD algorithms under the assumption that the Markov chain is irreducible and aperiodic (see Chapter 6 of [Kon02]).

We now show that LSPE has the same rate of convergence as LSTD, so that LSPE is also asymptotically optimal. This also implies that LSPE with a constant stepsize has the same asymptotic variance as LSTD (under the additional assumption that both r_t and $\hat{r}_t$ are bounded with probability 1, the covariance matrices of $\sqrt{t}(r_t - r^*)$ and $\sqrt{t}(\hat{r}_t - r^*)$ converge to Σ_0).

To prove this, we will use the multiple-step contraction property, Lemma 11.4, to establish first the following proposition. Since the rate of convergence is at least of the order $O(1/\sqrt{t})$ for both LSPE and LSTD, the proposition says that LSPE and LSTD "converge" to each other at a faster scale than to the limit r^*.

Proposition 11.2. *For any $\alpha \in [0,1)$, the random variable $t^\alpha(r_t - \hat{r}_t)$ converges to* 0 *w.p.*1.

To prove this proposition, we will need the following lemma to bound the difference of the successive LSPE updates.

Lemma 11.5. *Consider a sample path for which both LSTD and LSPE converge, and let C be any fixed positive number and n any fixed integer. Then there exists a constant $K(C,n)$ depending on C, n and the sample path, such that for all t sufficiently large, whenever $\|r_t - \hat{r}_t\| \leq \frac{C}{t}$, we have*

$$\|r_{t+i} - r_t\| \leq \frac{K(C,n)}{t}, \quad i = 1, \ldots, n.$$

Proof: By the definition of LSPE updates and LSTD updates,

$$\begin{aligned} \|r_{t+1} - r_t\| &= \gamma \left\|\bar{B}_t^{-1}\left(\bar{A}_t r_t + \bar{b}_t\right)\right\| \\ &= \gamma \left\|\bar{B}_t^{-1}\bar{A}_t\left(r_t + \bar{A}_t^{-1}\bar{b}_t\right)\right\| \\ &= \gamma \left\|\bar{B}_t^{-1}\bar{A}_t\left(r_t - \hat{r}_{t+1}\right)\right\|. \end{aligned} \tag{11.15}$$

By Lemma 11.3, we can choose constants C_1 and C_2 such that for t sufficiently large,

$$\gamma\,\|\bar{B}_t^{-1}\bar{A}_t\| \leq C_1, \qquad \|\hat{r}_t - \hat{r}_{t+1}\| \leq \frac{C_2}{t}.$$

Assume for a t sufficiently large, that

$$\|r_t - \hat{r}_t\| \leq \frac{C}{t}.$$

[3] More precisely, $\bar{\gamma} = \lim_{t\to\infty} \gamma_{t+1}^{-1} - \gamma_t^{-1}$, which is equal to 1 if the stepsize is chosen as $\gamma_t = 1/t$.

Define $K_0 = C$. It follows from Eq. (11.15) that

$$\begin{aligned}\|r_{t+1} - r_t\| &\leq \gamma \left\|\bar{B}_t^{-1}\bar{A}_t\right\| \|r_t - \hat{r}_{t+1}\| \\ &\leq C_1 \left(\|r_t - \hat{r}_t\| + \|\hat{r}_t - \hat{r}_{t+1}\|\right) \\ &\leq C_1 \left(\frac{K_0}{t} + \frac{C_2}{t}\right) = \frac{K_1}{t}, \end{aligned} \tag{11.16}$$

where $K_1 = C_1(K_0 + C_2)$. This in turn implies that

$$\begin{aligned}\|r_{t+1} - \hat{r}_{t+1}\| &\leq \|r_{t+1} - r_t\| + \|r_t - \hat{r}_t\| + \|\hat{r}_t - \hat{r}_{t+1}\| \\ &\leq \frac{K_1}{t} + \left(\frac{K_0}{t} + \frac{C_2}{t}\right) = \frac{K_1'}{t}, \end{aligned} \tag{11.17}$$

where $K_1' = K_1 + K_0 + C_2$. Recursively applying the argument of Eq. (11.16) and Eq. (11.17) for $t + i, i = 2, \dots n$, it can be seen that

$$\|r_{t+i} - r_{t+i-1}\| \leq \frac{K_i}{t}, \qquad \|r_{t+i} - \hat{r}_{t+i}\| \leq \frac{K_i'}{t}, \quad i \leq n \tag{11.18}$$

where K_i and K_i' are recursively defined by

$$K_i = C_1(K_{i-1}' + C_2), \qquad K_i' = K_i + K_{i-1}' + C_2,$$

and they depend on i and constants C, C_1 and C_2 only. By the triangle inequality,

$$\|r_{t+i} - r_t\| \leq \frac{K_i + K_{i-1} + \cdots + K_1}{t} \leq \frac{\sum_{i=1}^n K_i}{t}, \quad i \leq n.$$

Hence the constant $K(C, n)$ in the claim can be defined by $K(C, n) = \sum_{i=1}^n K_i$, and the proof is complete. □

Proof of Prop. 11.2: We will prove the proposition by contradiction. Consider a sample path for which both LSTD and LSPE converge. Fix any $\delta > 0$. Assume that there exists a subsequence of $r_t - \hat{r}_t$ indexed by t_i such that

$$t_i^\alpha \|r_{t_i} - \hat{r}_{t_i}\| \geq \delta.$$

Equivalently, for $C_1 = \delta$,

$$\|r_{t_i} - \hat{r}_{t_i}\| \geq C_1 \, t_i^{-\alpha}. \tag{11.19}$$

Let ρ, k and $\epsilon_t = \frac{C_2}{t}$ be as defined in Lemma 11.4. By part two of Lemma 11.4, there exist infinitely many t such that

$$\|r_t - \hat{r}_t\| \leq \epsilon_t = \frac{C_2}{t}.$$

Since eventually $C_1 t^{-\alpha} \geq \epsilon_t = \frac{C_2}{t}$ for sufficiently large t, there are infinitely many time intervals during which $\|r_t - \hat{r}_t\|$ lies between the two.

Consider one such interval. Let a be the start of the interval, and T_a the interval length, defined such that

$$\|r_{a-1} - \hat{r}_{a-1}\| \leq \epsilon_{a-1}, \tag{11.20}$$

$$\epsilon_{a+i} < \ \|r_{a+i} - \hat{r}_{a+i}\| < \frac{C_1}{(a+i)^\alpha}, \quad 0 \leq i < T_a, \tag{11.21}$$

$$\|r_{a+T_a} - \hat{r}_{a+T_a}\| \geq \frac{C_1}{(a+T_a)^\alpha}. \tag{11.22}$$

Write T_a as $T_a = lk + m$ with $0 \leq m < k$. By Eq. (11.21) and Lemma 11.4, within the time interval $[a, a+T_a]$ the k-step contractions continue to happen, so that

$$\begin{aligned}\|r_{a+T_a} - \hat{r}_{a+T_a}\| &\leq \rho^{T_a - m}\|r_{a+m} - \hat{r}_{a+m}\| \\ &\leq \rho^{T_a - m}\left(\|r_{a-1} - \hat{r}_{a-1}\| + \|r_{a-1} - r_{a+m}\| + \|\hat{r}_{a-1} - \hat{r}_{a+m}\|\right) \\ &\leq \rho^{T_a - m}\left(\epsilon_{a-1} + \|r_{a-1} - r_{a+m}\| + \|\hat{r}_{a-1} - \hat{r}_{a+m}\|\right).\end{aligned}$$

By Eq. (11.20) and Lemma 11.5, for a given k, when the time a is sufficiently large, $\|r_{a-1} - r_{a+m}\|$, where $m+1 \leq k$, is at most $C'/(a-1)$ for some constant C' (depending on k). By Lemma 11.3, $\|\hat{r}_{a-1} - \hat{r}_{a+m}\| \leq Ck/a$ for some constant C. Therefore for some constant C,

$$\|r_{a+T_a} - \hat{r}_{a+T_a}\| \leq \rho^{T_a - m}\left(\epsilon_{a-1} + C/a + C/a\right) \leq \rho^{T_a - m}\frac{C_3}{a},$$

where C_3 is some constant. Combining this with Eq. (11.22), we have

$$\rho^{T_a - m}\frac{C_3}{a} \geq \frac{C_1}{(a+T_a)^\alpha},$$

and equivalently, redefining the constants, we have for some constant C

$$\rho^{T_a} \geq \frac{Ca}{(a+T_a)^\alpha}.$$

Clearly for a sufficiently large, in order to satisfy this inequality, T_a must be at least of the order of $a^{1/\alpha}$, and therefore the preceding inequality implies that

$$\rho^{T_a} \geq \frac{Ca}{(2T_a)^\alpha}$$

for a sufficiently large. However, it is impossible for T_a to satisfy the inequality. Thus we have reached a contradiction, and the claimed convergence statement follows. □

Proposition 11.3. *Assume that with probability* 1 *the random variable* $\sqrt{t}\,(\hat{r}_t - r^*)$ *of LSTD converges to a Gaussian random variable* $N(0, \Sigma_0)$. *Then with probability* 1 *the random variable* $\sqrt{t}\,(r_t - r^*)$ *of LSPE converges to the same random variable.*

Proof: To prove the claim, we write

$$\sqrt{t+1}(r_{t+1} - r^*) = \sqrt{t+1}\left(I + \gamma \bar{B}_t^{-1}\bar{A}_t\right)(r_t - \hat{r}_{t+1}) + \sqrt{t+1}(\hat{r}_{t+1} - r^*),$$

and thus it suffices to show that $\sqrt{t+1}\left(I + \gamma \bar{B}_t^{-1}\bar{A}_t\right)(r_t - \hat{r}_{t+1}) \to 0$ w.p.1.

Consider a sample path for which both LSTD and LSPE converge. When t is sufficiently large, $\|I + \gamma \bar{B}_t^{-1} \bar{A}_t\| \le C$ for some constant C, so that

$$\left\| \sqrt{t+1}\left(I + \gamma \bar{B}_t^{-1} \bar{A}_t\right)(r_t - \hat{r}_{t+1})\right\| \le C\sqrt{t+1}\left(\|r_t - \hat{r}_t\| + \|\hat{r}_t - \hat{r}_{t+1}\|\right).$$

Since $\|\hat{r}_t - \hat{r}_{t+1}\| \le \frac{C'}{t}$ for some constant C' (Lemma 11.3), the second term $C\sqrt{t+1}\|\hat{r}_t - \hat{r}_{t+1}\|$ converges to 0. By Prop. 11.2, the first term, $C\sqrt{t+1}\|r_t - \hat{r}_t\|$, also converges to 0. The proof is thus complete. □

Remark 11.2. As has been proved, LSPE with *any* constant stepsize (under which LSPE converges) has the same asymptotic optimal convergence rate as LSTD, i.e., the convergence rate of LSPE does not depend on the constant stepsize. As the proof of the Lemma 11.4 and the discussion after it show, the stepsize affects how closely (as reflected in the constant in ϵ_t) and how fast (as reflected in ρ of the multiple-step contraction) r_t tracks the solution $-\bar{A}_t^{-1}\bar{b}_t$ of the linear system. These, however, happen at the scale of $O(t)$, while the convergence of r_t to r^* is at the scale of $O(\sqrt{t})$. This explains why the constant stepsize does not affect the asymptotic convergence rate of LSPE.

11.4 Summary

In this chapter we first proved the convergence of the average cost LSPE with a constant stepsize by a slight modification of the proof for its discounted counterpart. We then proved the optimal convergence rate of LSPE with a constant stepsize for both discounted and average cost cases. The analysis also shows that LSTD and LSPE with a constant stepsize converge to each other at a faster scale than they to the common limit.

Chapter 12

Conclusions

In this thesis, we have addressed three main topics: lower cost approximations of the optimal cost function for POMDP and POSMDP with various cost criteria and constraints, reinforcement learning-based approximation algorithms for POMDP and MDP, and properties of the optimal solution to the average cost POMDP problem.

For lower cost approximation, using the special structure of hidden states in POMDP, we establish a lower bound result which holds for POMDP and POSMDP in general. In particular, this result holds regardless of the existence of solutions to the optimality equations in the average cost case, or the analytical difficulty in characterizing the optimal solutions in the constrained case. We also address computational and convergence issues.

For reinforcement learning in POMDP, we propose and analyze a function approximation approach to estimation of the policy gradient for POMDPs with finite-state controllers. As an actor-critic type of method, our algorithm is an extension of the existing actor-only policy gradient methods in POMDPs; and it also clarifies the view that reinforcement learning methods for POMDPs are special cases of those for MDPs. For reinforcement learning in MDP, we prove two convergence results for LSPE, a least squares TD algorithm for policy evaluation. These are the convergence of the average cost LSPE(λ) with a constant stepsize, and the asymptotically optimal convergence rate of LSPE with a constant stepsize for both discounted and average cost cases.

For the average cost POMDP problem, besides constructing examples with non-constant optimal cost functions, we give also a new necessary condition for the optimal liminf cost to be constant. The result leads us further to prove for a finite space POMDP the near-optimality of the class of finite-state controllers under the assumption of a constant optimal liminf cost function. This provides a theoretical guarantee for the finite-state controller approach.

As future work for POMDPs with the expected cost criteria, we consider two questions that have central importance. One is the existence of solutions to the average cost optimality equations of POMDPs, the understanding of which will greatly help algorithm design. The other is incorporating model learning and approximate inference into decision making, a necessary step for tackling large scale problems in which exact inference of the hidden states is intractable.

Finally, we note that efficient computational methods for large scale POMDPs are active research fields, and there are many important approaches and issues that we have not been able to address in this thesis. In particular, belief compression, approximate linear programming, approximate value iteration, MDP-based model approximation, and kernel-based learning are promising methods for large scale problems, perhaps more efficient than discretization-based methods. There are also problems with different cost criteria to be considered, such as multi-objective and robust control in POMDPs.

Appendix A

Analysis Based on the Weaker Lower Bound Result

A.1 Average Cost POSMDPs

In Chapter 6, using the stronger lower bound result, Theorem 6.1, we have proved that the optimal cost function of the modified SMDP problem associated with a lower cost approximation scheme is a lower bound of the optimal cost function of the original POSMDP problem. A question of our interest is what lower bound statement can we prove using the weaker line of analysis based on the inequalities for the optimal cost functions.

Recall that for average cost POMDPs with finite space models, the weaker and stronger lines of analysis give the same conclusion (see discussions preceding Theorem 5.1 in Section 5.3), where the weaker one is based essentially on the inequalities for the optimal finite-stage cost functions. For average cost POSMDPs with finite space models, however, the random time length τ_N of a N-stage problem varies with the policy and initial distribution, therefore we cannot use the finite-stage inequalities to deduce the average cost lower bound analogous to that for POMDPs.

Yet we still have the inequalities for discounted problems for all discounting rate $\alpha > 0$, so we can apply a vanishing discounting argument. This line of analysis, however, only establishes at the end that the optimal average cost function of the modified problem is a lower bound of the optimal *limsup* cost function J^*_{C+}, while the stronger result says that the same lower bound is a lower bound of J^*_{C-}. Thus it seems to us that there is a technical limitation in the weaker line of analysis.

Nonetheless, we provide in this appendix the proof of the weaker result for reference and comparison. The notation and assumptions for POSMDPs are as in Chapter 6. We assume that τ_n takes continuous values – the discrete case is easy to prove. The technique of the proof is to discretize the time and then use Tauberian theorem.

The Weaker Statement and its Proof

First we clarify a subtlety. Recall that the per-stage cost $g(s, u)$ depends on the discounting rate α. However, it can be seen that the inequalities hold for any discounting rate $\alpha > 0$ and $g(s, u)$ defined at a *different* discounting rate α', provided that g is bounded. So, in particular, we can use the inequalities for discounted problems with the per-stage cost defined as the per-stage cost in the average cost problem.

Now, recall that in the modified belief SMDP associated with a discretized lower approximation scheme, $\tilde{J}^*(\xi) = \tilde{J}^*_{C-}(\xi)$. Furthermore, on the finite-state and control SMDP, (similar and however different to the finite-state and control MDP), under certain conditions, there exists a Blackwell optimal policy (Denardo [Den71]). Extend it to the entire belief space and denote it by $\tilde{\pi}^*$. It is Blackwell optimal in the sense that for any ξ, there exists α_ξ such that $\tilde{\pi}^*$ is optimal at ξ for discounted problems with $\alpha \in (0, \alpha_\xi)$.

Indeed for proving the following lower bound, instead of the existence of a Blackwell optimal policy, it suffices that for each initial distribution ξ there exists an average cost optimal policy $\tilde{\pi}^*$ that is also optimal for a sequence of α_k-discounted problems with $\alpha_k \downarrow 0$. It can be seen that there exists such a policy and such a sequence $\{\alpha_k\}$ for each ξ, because there is only a finite number of deterministic and stationary policies on SMDP with state space $\{\xi\} \cup \mathcal{C}$.

Proposition A.1. *For every initial distribution ξ, the optimal average cost of the modified belief SMDP is less than the optimal limsup average cost of the original POSMDP:*

$$\tilde{J}^*(\xi) \leq J^*_{C+}(\xi).$$

Proof: We first prove the case where all the per-stage costs are *non-negative*, (the proof of the non-positive case is identical.) We then prove the case where the costs are arbitrarily signed.

(i) Assume the costs $g(s, u)$ are non-negative. Denote by $\widetilde{E}^{\tilde{\pi}}_\xi$ the expectation and X the belief states in the modified SMDP with policy $\tilde{\pi}$ and initial distribution ξ. Denote by E^π_ξ the expectation in the original POSMDP with policy π and initial distribution ξ. Let $\tilde{\pi}^*$ be an average cost optimal policy for the modified problem that is either Blackwell optimal or optimal for a sequence of discounted problems as in the discussion preceding the proposition. Without loss of generality we assume it is Blackwell optimal, i.e., optimal for all α-discounted problems with $\alpha < \alpha_\xi$, (in terms of the proof it is the same in the case of taking a sequence of α_k). By the inequality for the optimal discounted cost, Eq. (6.3) of Chapter 6, we have for any $\alpha < \alpha_\xi$ and any π,

$$\widetilde{E}^{\tilde{\pi}^*}_\xi \left\{ \sum_{n=0}^{\infty} e^{-\alpha \tau_n} \bar{g}(X_n, \tilde{U}_n) \right\} \leq E^\pi_\xi \left\{ \sum_{n=0}^{\infty} e^{-\alpha \tau_n} g(S_n, U_n) \right\}. \tag{A.1}$$

We are going to discretize the time and use Tauberian theorem. According to Assumption 6.1, there exist $a > 0$ and $\delta \in (0, 1)$ such that

$$\sup_{s,u} P(\tau_1 \leq a \mid S_0 = s, U_0 = u) < \delta.$$

Suppose we discretize time into length $\rho < a$ intervals. Define random variables $\tilde{c}_k$ and c_k by

$$\tilde{c}_k = \sum_{n: \tilde{\tau}_n \in [k\rho, (k+1)\rho)} \bar{g}(X_n, \tilde{U}_n), \qquad c_k = \sum_{n: \tau_n \in [k\rho, (k+1)\rho)} g(S_n, U_n). \tag{A.2}$$

For any interval, let events $A_m = \{m \text{ decision epochs happened in the interval}\}$. The event A_m is contained in the event that $\tau_j - \tau_{j-1} < \rho$ for $(m-1)$ consecutive decision epochs. Hence we have

$$P(A_m) \leq \delta^{m-1}$$

and

$$E_\xi^\pi \left\{c_k\right\} \le \sum_{m=0}^{\infty} P(A_m) m L \le \frac{L}{(1-\delta)^2}.$$

The same bound holds for $\widetilde{E}_\xi^{\tilde{\pi}} \left\{\tilde{c}_k\right\}$.

For any $\epsilon > 0$, we pick some $\rho < a$. There exists $\bar{\alpha}_\xi$ such that for all $\alpha < \bar{\alpha}_\xi$

$$1 - e^{-\alpha\rho} < \frac{\epsilon\rho}{2\frac{L}{(1-\delta)^2}}, \qquad e^{\alpha\rho} - 1 < \frac{\epsilon\rho}{2\frac{L}{(1-\delta)^2}},$$

thus, for all k and $\alpha \in (0, \bar{\alpha}_\xi)$

$$(1 - e^{-\alpha\rho}) \widetilde{E}_\xi^{\tilde{\pi}} \left\{\tilde{c}_k\right\} < \frac{\epsilon\rho}{2}, \qquad (e^{\alpha\rho} - 1) E_\xi^\pi \left\{c_k\right\} < \frac{\epsilon\rho}{2}.$$

We have,

$$\widetilde{E}_\xi^{\tilde{\pi}^*} \left\{ \sum_{n=0}^{\infty} e^{-\alpha\tilde{\tau}_n} \bar{g}(X_n, \tilde{U}_n) \right\} \ge \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{ \sum_{k=0}^{\infty} e^{-\alpha\rho k} \tilde{c}_k e^{-\alpha\rho} \right\} = \sum_{k=0}^{\infty} e^{-\alpha\rho k} \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{ \tilde{c}_k e^{-\alpha\rho} \right\},$$

where in the equality the interchange of expectation and summation is justified by monotone convergence theorem. Since

$$\begin{aligned} \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{ \tilde{c}_k e^{-\alpha\rho} \right\} &= \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{\tilde{c}_k\right\} - (1 - e^{-\alpha\rho}) \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{\tilde{c}_k\right\} \\ &\ge \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{\tilde{c}_k\right\} - \frac{\epsilon\rho}{2}, \end{aligned}$$

we have

$$\widetilde{E}_\xi^{\tilde{\pi}^*} \left\{ \sum_{n=0}^{\infty} e^{-\alpha\tilde{\tau}_n} g(X_n, \tilde{U}_n) \right\} \ge \sum_{k=0}^{\infty} e^{-\alpha\rho k} \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{\tilde{c}_k\right\} - (1 - e^{-\alpha\rho})^{-1} \frac{\epsilon\rho}{2}.$$

Similarly we have

$$E_\xi^\pi \left\{ \sum_{n=0}^{\infty} e^{-\alpha\tau_n} g(S_n, U_n) \right\} \le \sum_{k=0}^{\infty} e^{-\alpha\rho k} E_\xi^\pi \left\{c_k\right\} + (1 - e^{-\alpha\rho})^{-1} \frac{\epsilon\rho}{2}.$$

It then follows from Eq. (A.1) that for all $\alpha \in (0, \bar{\alpha}_\xi)$

$$(1 - e^{-\alpha\rho}) \sum_{k=0}^{\infty} e^{-\alpha\rho k} \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{\tilde{c}_k\right\} \le (1 - e^{-\alpha\rho}) \sum_{k=0}^{\infty} e^{-\alpha\rho k} E_\xi^\pi \left\{c_k\right\} + \epsilon\rho, \tag{A.3}$$

By a Tauberian theorem we have, from the left-hand side of Eq. (A.3),

$$\liminf_{\alpha \to 0} (1 - e^{-\alpha\rho}) \sum_{k=0}^{\infty} e^{-\alpha\rho k} \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{\tilde{c}_k\right\} \ge \liminf_{K \to \infty} \frac{1}{K+1} \widetilde{E}_\xi^{\tilde{\pi}^*} \left\{ \sum_{k=0}^{K} \tilde{c}_k \right\}$$

and from the right-hand side of Eq. (A.3),

$$\liminf_{\alpha\to 0}\,(1-e^{-\alpha\rho})\sum_{k=0}^{\infty} e^{-\alpha\rho k} E_\xi^\pi\left\{c_k\right\} \le \limsup_{K\to\infty}\ \frac{1}{K+1} E_\xi^\pi\left\{\sum_{k=0}^{K} c_k\right\}.$$

Since $|\bar g(X,u)| < L$,

$$\frac{1}{\rho(K+1)}\tilde J_C^{\tilde\pi^*}\left(\xi,\rho(K+1)\right) \le \frac{1}{\rho(K+1)}\left(\widetilde E_\xi^{\tilde\pi^*}\left\{\sum_{k=0}^{K}\tilde c_k\right\} + L\right)$$

$$\Rightarrow\qquad \rho\,\tilde J^*(\xi) = \rho\,\liminf_{T\to\infty}\ \frac{1}{T}\tilde J_C^{\tilde\pi^*}(\xi,T) \le \liminf_{K\to\infty}\ \frac{1}{K+1}\widetilde E_\xi^{\tilde\pi^*}\left\{\sum_{k=0}^{K}\tilde c_k\right\},$$

and similarly,

$$\frac{1}{\rho(K+1)} E_\xi^\pi\left\{\sum_{k=0}^{K} c_k\right\} \le \frac{1}{\rho(K+1)} J_C^\pi\left(\xi,\rho(K+1)\right)$$

$$\Rightarrow\qquad \limsup_{K\to\infty}\ \frac{1}{K+1} E_\xi^\pi\left\{\sum_{k=0}^{K} c_k\right\} \le \rho\,\limsup_{T\to\infty}\ \frac{1}{T} J_C^\pi(\xi,T),$$

it follows that

$$\rho\,\tilde J^*(\xi) \le \rho\,\limsup_{T\to\infty}\ \frac{1}{T}J_C^\pi(\xi,T) + \epsilon\rho \qquad\Rightarrow\qquad \tilde J^*(\xi) \le \inf_{\pi\in\Pi}\,\limsup_{T\to\infty}\ \frac{1}{T}J_C^\pi(\xi,T) + \epsilon.$$

Because ϵ is arbitrarily small, we conclude that $\tilde J^*(\xi) \le J_{C+}^*(\xi)$.

(ii) Now we consider arbitrary signed bounded costs. We follow the same steps as in the proof above. First we define random variables:

$$g_n^+ = \max\left\{\, g(S_n,U_n)\,,0\,\right\},\qquad g_n^- = -\min\left\{\, g(S_n,U_n)\,,0\,\right\}.$$

For almost surely all sample paths the infinite summation

$$\sum_{n=0}^{\infty} e^{-\alpha\tau_n} g(S_n,U_n) = \sum_{n=0}^{\infty} e^{-\alpha\tau_n} g_n^+ - \sum_{n=0}^{\infty} e^{-\alpha\tau_n} g_n^-$$

is well-defined and bounded. We have

$$E_\xi^\pi\left\{\sum_{n=0}^{\infty} e^{-\alpha\tau_n} g(S_n,U_n)\right\} = E_\xi^\pi\left\{\sum_{n=0}^{\infty} e^{-\alpha\tau_n} g_n^+\right\} - E_\xi^\pi\left\{\sum_{n=0}^{\infty} e^{-\alpha\tau_n} g_n^-\right\},$$

and similarly defining $\tilde g_n^+$ and $\tilde g_n^-$ in the modified problem, we have

$$\widetilde E_\xi^{\tilde\pi^*}\left\{\sum_{n=0}^{\infty} e^{-\alpha\tilde\tau_n} g(X_n,\tilde U_n)\right\} = \widetilde E_\xi^{\tilde\pi^*}\left\{\sum_{n=0}^{\infty} e^{-\alpha\tau_n}\tilde g_n^+\right\} - \widetilde E_\xi^{\tilde\pi^*}\left\{\sum_{n=0}^{\infty} e^{-\alpha\tau_n}\tilde g_n^-\right\}.$$

We then discretize time into length ρ intervals, define random variables c_k^+, c_k^-, $\tilde c_k^+$, $\tilde c_k^-$ for the corresponding sequences $\{g_n^+\}$ $\{g_n^-\}$ $\{\tilde g_n^+\}$ $\{\tilde g_n^-\}$, respectively. For α sufficiently small, we thus have

$$(1-e^{-\alpha\rho})\left(\sum_{k=0}^{\infty} e^{-\alpha\rho k}\widetilde{E}_{\xi}^{\tilde{\pi}^*}\left\{\tilde{c}_k^+\right\} - \sum_{k=0}^{\infty} e^{-\alpha\rho k}\widetilde{E}_{\xi}^{\tilde{\pi}^*}\left\{\tilde{c}_k^-\right\}\right)$$

$$\leq (1-e^{-\alpha\rho})\left(\sum_{t=0}^{\infty} e^{-\alpha\rho k}E_{\xi}^{\pi}\left\{c_k^+\right\} - \sum_{t=0}^{\infty} e^{-\alpha\rho k}E_{\xi}^{\pi}\left\{c_k^-\right\}\right) + 2\epsilon\rho,$$

$$\Rightarrow \qquad (1-e^{-\alpha\rho})\sum_{k=0}^{\infty} e^{-\alpha\rho k}\widetilde{E}_{\xi}^{\tilde{\pi}^*}\left\{\tilde{c}_k^+ - \tilde{c}_k^-\right\} \leq (1-e^{-\alpha\rho})\sum_{t=0}^{\infty} e^{-\alpha\rho k}E_{\xi}^{\pi}\left\{c_k^+ - c_k^-\right\} + 2\epsilon\rho,$$

where, to derive the second equation, we have used the fact that $E_{\xi}^{\pi}\left\{c_k^{\pm}\right\}, \widetilde{E}_{\xi}^{\tilde{\pi}^*}\left\{\tilde{c}_k^{\pm}\right\}$ are bounded. Due to this same fact we can apply Tauberian theorem to both sides of the equation to have, defining $c_k = c_k^+ - c_k^-$ and $\tilde{c}_k = \tilde{c}_k^+ - \tilde{c}_k^-$:

$$\liminf_{K\to\infty} \frac{1}{K+1}\widetilde{E}_{\xi}^{\tilde{\pi}^*}\left\{\sum_{k=0}^{K}\tilde{c}_k\right\} \leq \limsup_{K\to\infty} \frac{1}{K+1}E_{\xi}^{\pi}\left\{\sum_{k=0}^{K} c_k\right\} + 2\epsilon\rho.$$

Following the same line of argument as in (i), we then conclude that $\tilde{J}^*(\xi) \leq J_{C+}^*(\xi)$. □

A.2 Constrained Average Cost POMDPs

In Chapter 9, using the stronger lower bound result, Theorem 3.2, we obtain as an immediate consequence that the constrained optimal cost function of the modified belief MDP is a lower bound of the constrained optimal cost of the original POMDP. For finite-space POMDPs and modified MDPs that are essentially finite, the same lower bound claim can also be proved using the weaker lower bound result that depends on the DP mapping of the modified problem. As a reference, we provide such a proof for the case where the modified problem is unichain (compare with Section 9.2.1), and the multichain case is similar.

The notation is the same as in Section 9.2.1. The unichain LP for the modified problem is:

$$\tilde{J}_c^* = \min_{q\geq 0} \sum_{s\in\mathcal{C},u\in\mathcal{U}} \bar{g}_0(s,u)\,q(s,u) \tag{A.4}$$

$$\text{Subj.} \quad \sum_{u\in\mathcal{U}} q(s',u) - \sum_{s\in\mathcal{C},u\in\mathcal{U}} p(s'|s,u)\,q(s,u) = 0, \quad \forall s'\in\mathcal{C} \tag{A.5}$$

$$\sum_{s\in\mathcal{C},u\in\mathcal{U}} q(s,u) = 1, \tag{A.6}$$

$$\sum_{s\in\mathcal{C},u\in\mathcal{U}} \bar{g}_k(s,u)\,q(s,u) \leq c_k, \quad k=1,\ldots,n. \tag{A.7}$$

Proposition A.2. *Suppose the original constrained POMDP problem is feasible. If the modified belief MDP is unichain, then the constrained modified problem is feasible and its optimal average cost $\tilde{J}_c^*$, a constant, satisfies*

$$\tilde{J}_c^* \leq J_c^*(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Proof: First we prove that if the original problem is feasible, then the constrained modified problem is feasible. We prove it by showing that a maximization problem dual to the LP (A.4) has bounded value. Consider the Lagrangian formed by relaxing the constraints (A.7) in the LP (A.4), and the corresponding dual function is

$$F(\lambda) = \min_{q \geq 0} \sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_0(s,u)\, q(s,u) + \sum_{k=1}^{n} \lambda_k \left(\sum_{s \in \mathcal{C}, u \in \mathcal{U}} \bar{g}_k(s,u)\, q(s,u) - c_k \right) \tag{A.8}$$
$$\text{Subj.} \quad \text{constraints (A.5) and (A.6)}$$

for any non-negative $\lambda = (\lambda_1, \ldots, \lambda_n)$. It is the LP (modulo a constant term, $-\sum_k \lambda_k c_k$, in the objective) of an unconstrained average cost MDP with the combined per-stage cost $\bar{g}_0 + \sum_{k=1}^{n} \lambda_k \bar{g}_k$. Denote the optimal average cost of this MDP problem, (a constant, due to the unichain condition), by $\tilde{J}^*(\lambda)$. Thus

$$F(\lambda) = \tilde{J}^*(\lambda) - \sum_{k=1}^{n} \lambda_k c_k.$$

Since the original constrained POMDP problem is feasible, there exists a policy π such that $J_{k,+}^{\pi} \leq c_k, k = 1, \ldots, n$. Due to the fact

$$\limsup_{i \to \infty} (a_i + b_i) \leq \limsup_{i \to \infty} a_i + \limsup_{i \to \infty} b_i$$

and the non-negativity of λ, it follows that for the combined per-stage cost, the limsup average cost of π, denoted by $J_{+}^{\pi}(\lambda, \xi)$, satisfies

$$J_{+}^{\pi}(\lambda, \xi) \leq J_{0,+}^{\pi}(\xi) + \sum_{k=1}^{n} \lambda_k J_{k,+}^{\pi}(\xi) \leq J_{0,+}^{\pi}(\xi) + \sum_{k=1}^{n} \lambda_k c_k, \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

By the lower bound property of the unconstrained POMDP (Theorem 5.1),

$$\tilde{J}^*(\lambda) \leq J_{+}^{\pi}(\lambda, \xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Hence for all λ,

$$\begin{aligned} F(\lambda) &\leq J_{0,+}^{\pi}(\xi) \leq \max_{s \in \mathcal{S}, u \in \mathcal{U}} g_0(s,u), \\ \Rightarrow \quad \max_{\lambda \geq 0} F(\lambda) &\leq J_{0,+}^{\pi}(\xi) \leq \max_{s \in \mathcal{S}, u \in \mathcal{U}} g_0(s,u). \end{aligned} \tag{A.9}$$

Thus the dual problem $\max_{\lambda \geq 0} F(\lambda)$ has finite value, and it follows from the strong duality of LP that the primal LP (A.4) is feasible, and furthermore, $\tilde{J}_c^* = \max_{\lambda \geq 0} F(\lambda)$.

To show $\tilde{J}_c^* \leq J_c^*$, notice that the inequality (A.9) holds for any policy π in the feasible set Π_f of the original problem. Thus

$$\tilde{J}_c^* = \max_{\lambda \geq 0} F(\lambda) \leq \inf_{\pi \in \Pi_f} J_{0,+}^{\pi}(\xi) = J_c^*(\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

This completes the proof. □

Appendix B

Entropy Rate of Non-Stationary Hidden Markov Sources

In this appendix, we provide proofs for two propositions in Section 8.2, which state the convergence of the entropy rate and the convergence of the conditional entropy of a non-stationary hidden Markov source. We will restate the propositions for convenience.

Proposition B.1. *Suppose the Markov chain $\{S_n\}$ is irreducible. Then*

$$H(\underline{Y}) = \lim_{n\to\infty} \frac{1}{n} H(Y^n;\xi), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Proof: First we prove the case where the Markov chain is aperiodic. For every initial distribution ξ, by the chain rule of entropy,

$$H(Y^n;\xi) = H(S^n, Y^n;\xi) - H(S^n \mid Y^n;\xi).$$

Because $\{(S_n, Y_n)\}$ is jointly a Markov chain, it is easy to show that under the aperiodicity condition the term $\frac{1}{n}H(S^n, Y^n;\xi)$ converges to the same limit for all initial distribution ξ. So we only need to show that $\frac{1}{n}H(S^n \mid Y^n;\xi)$ converges to the same limit for all ξ.

We have,

$$\begin{aligned} H(S^n \mid Y^n;\xi) &= H(S_1 \mid Y^n;\xi) + \sum_{k=2}^{n} H(S_k \mid Y^n, S_1 \dots S_{k-1};\xi) \\ &= H(S_1 \mid Y^n;\xi) + \sum_{k=2}^{n} H(S_k \mid Y_k, \dots, Y_n, S_{k-1};\xi), \end{aligned}$$

where the first equality is due to the chain rule of entropy, and the second equality is due to the conditional independence of S_k and (Y^{k-1}, S^{k-2}) given S_{k-1}. The conditional probability $p(S_k \mid Y_k, \dots, Y_n, S_{k-1})$ does not depend on ξ, therefore we can write for $k \geq 2$,

$$\begin{aligned} H(S_k \mid Y_k, \dots, Y_n, S_{k-1};\xi) &= -E\left\{E\{\log p(S_k \mid Y_k, \dots, Y_n, S_{k-1}) \mid S_{k-1}\} \mid \xi\right\} \\ &= E\{f_{k-1}(S_{k-1}) \mid \xi\} = \sum_s \xi_{k-1}(s) f_{k-1}(s), \end{aligned}$$

where

$$f_{k-1}(s) = -E\{\log p(S_k \mid Y_k, \dots, Y_n, S_{k-1}) \mid S_{k-1} = s\}$$

and ξ_{k-1} is the marginal distribution of S_{k-1} when the initial distribution is ξ. The function $f_k(s)$ is non-negative and bounded by some constant C_1 for all s and k – in fact it is bounded by the entropy of S_k given $S_{k-1} = s$.

Since $\{S_n\}$ is irreducible and aperiodic, there exists a constant C_2 and a positive number $\beta < 1$ such that

$$\|\xi_k - \bar{\xi}\|_\infty \leq C_2 \beta^k,$$

where $\bar{\xi}$ is the equilibrium distribution. Therefore, using the relation

$$|H(S_1 \mid Y^n; \xi) - H(S_1 \mid Y^n; \bar{\xi})| \leq 2 \max_{\tilde{\xi} \in \mathcal{P}(\mathcal{S})} H(S_1; \tilde{\xi}) \leq C_3,$$

for some constant C_3, it follows that

$$\begin{aligned}
\left| \frac{1}{n} H(S^n \mid Y^n; \xi) - \frac{1}{n} H(S^n \mid Y^n; \bar{\xi}) \right| &= \frac{1}{n} \sum_{k=2}^{n} \sum_{s} |\xi_k(s) - \bar{\xi}(s)| f_k(s) \\
&\quad + \frac{1}{n} |H(S_1 \mid Y^n; \xi) - H(S_1 \mid Y^n; \bar{\xi})| \\
&\leq \frac{1}{n} \sum_{k=2}^{n} C_1 C_2 \beta^k + \frac{1}{n} C_3 \\
&\leq \frac{1}{n} \frac{C_1 C_2}{1-\beta} + \frac{1}{n} C_3 \to 0,
\end{aligned}$$

i.e., $\frac{1}{n} H(S^n \mid Y^n; \xi)$ converges to the same limit for all initial distribution ξ.

We now prove the case where the Markov chain is periodic. Let the period be d. Applying the common technique in dealing with periodic chains, we consider a block of d consecutive states as one state random variable $\bar{S}$ and their observations as one observation random variable $\bar{Y}$. The Markov chain $\{\bar{S}_m\}$ is now aperiodic. By the preceding proof, we have $\frac{1}{m} H(\bar{Y}^m; \xi) \to H(\underline{\bar{Y}})$,[1] the entropy rate of $\{\bar{Y}_m\}$. Since

$$H(\bar{Y}^{\lfloor n/d \rfloor}; \xi) \leq H(Y^n; \xi) \leq H(\bar{Y}^{\lfloor n/d+1 \rfloor}; \xi),$$

and

$$\frac{1}{n} H(\bar{Y}^{\lfloor n/d \rfloor}; \xi) \to \frac{1}{d} H(\underline{\bar{Y}}), \qquad \frac{1}{n} H(\bar{Y}^{\lfloor n/d+1 \rfloor}; \xi) \to \frac{1}{d} H(\underline{\bar{Y}}),$$

we have $\frac{1}{n} H(Y^n; \xi) \to \frac{1}{d} H(\underline{\bar{Y}})$, i.e., $\frac{1}{n} H(Y^n; \xi)$ converges to the same limit for all ξ. The proof is complete. □

Proposition B.2. *Suppose the Markov chain $\{S_n\}$ is irreducible and aperiodic. Then*

$$\lim_{n \to \infty} H(Y_n \mid Y^{n-1}; \xi) = H(\underline{Y}), \quad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Proof: First we show that $\limsup_{n \to \infty} H(Y_n \mid Y^{n-1}; \xi) \leq H(\underline{Y})$. By the property of conditional entropy, we have for any $k < n$,

$$H(Y_n \mid Y^{n-1}; \xi) \leq H(Y_n \mid Y_{k+1}, \ldots, Y_{n-1}; \xi) = H(Y_m \mid Y^{m-1}; \xi_k),$$

[1] Note that the initial distribution of $\bar{S}_1$ is determined by ξ, the initial distribution of S_1. Therefore we can still write the distribution parameter in $H(\bar{Y}^m; \xi)$ as ξ.

where $m = n - k$ and $\xi_k(\cdot) = P(S_k \in \cdot | S_1 \sim \xi)$. As $k \to \infty$, $\xi_k \to \bar{\xi}$ the equilibrium distribution. Hence, if we fix m and define $k(n) = n - m$, then we have

$$\limsup_{n\to\infty} H(Y_n \mid Y^{n-1}; \xi) \leq \lim_{n\to\infty} H(Y_m \mid Y^{m-1}; \xi_{k(n)}) = H(Y_m \mid Y^{m-1}; \bar{\xi}).$$

Since m is arbitrary, it follows then

$$\limsup_{n\to\infty} H(Y_n \mid Y^{n-1}; \xi) \leq \lim_{m\to\infty} H(Y_m \mid Y^{m-1}; \bar{\xi}) = H(\underline{Y}),$$

where the last equality is a known fact of stationary hidden Markov sources (Theorem 4.2.1 of [CT91]).

Next we show that $\liminf_{n\to\infty} H(Y_n \mid Y^{n-1}; \xi) \geq H(\underline{Y})$. By the property of conditional entropy, we have for any $k < n$,

$$H(Y_n \mid Y^{n-1}; \xi) \geq H(Y_n \mid Y^{n-1}, S_k; \xi) = H(Y_n \mid Y_{k+1}, \ldots, Y_{n-1}, S_k; \xi),$$

where the equality follows from the Markovian property. Let $\xi_k(\cdot) = P(S_k \in \cdot | S_1 \sim \xi)$. Then, the right hand side of the preceding equation can be written as

$$H(Y_n \mid Y_{k+1}, \ldots, Y_{n-1}, S_k; \xi) = H(Y_m \mid Y^{m-1}, S_1; \xi_k),$$

where $m = n - k$. Fix m and define $k(n) = n - m$. Since when $k \to \infty$, $\xi_k \to \bar{\xi}$, the equilibrium distribution, we have

$$\lim_{n\to\infty} H(Y_m \mid Y^{m-1}, S_1; \xi_{k(n)}) = H(Y_m \mid Y^{m-1}, S_1; \bar{\xi}),$$

which implies for any fixed m,

$$\liminf_{n\to\infty} H(Y_n \mid Y^{n-1}; \xi) \geq H(Y_m \mid Y^{m-1}, S_1; \bar{\xi}).$$

Since m is arbitrary, it follows that

$$\liminf_{n\to\infty} H(Y_n \mid Y^{n-1}; \xi) \geq \lim_{m\to\infty} H(Y_m \mid Y^{m-1}, S_1; \bar{\xi}) = H(\underline{Y}),$$

where the last equality is a known fact for the stationary hidden Markov source (Theorem 4.4.1 of [CT91]). □

Appendix C

Differentiability of Average Cost

Consider a family of finite-state Markov chains parameterized by $\theta \in \Theta$ with a common state space $\mathcal{X} = \{1, 2, \ldots, k\}$, and denote a chain with parameter θ by $\{X_t^\theta\}$. Denote its transition probability matrix by P_θ, and its stationary distribution by π_θ, which satisfies $\pi'_\theta P_\theta = \pi'_\theta$. The average cost $\eta_\theta = \sum_{i \in \mathcal{X}} \pi_\theta(i) g_\theta(i)$, where $g_\theta(i)$ is the expected cost per-stage at state i. Thus the question of differentiability of η_θ with respect to θ becomes, assuming g_θ is differentiable, the question of differentiability of the stationary distribution π_θ. The following result is well-known. We provide a proof which is quite short.

Assumption C.1. *For all $\theta \in \Theta$, the Markov chain $\{X_t^\theta\}$ is recurrent and aperiodic.*

Proposition C.1. *Under Assumption C.1, if the transition probability matrix P_θ is differentiable, then the stationary distribution π_θ is differentiable.*

Proof: We use the following result from non-negative matrices (see Seneta [Sen73] pp. 5, proof of Theorem 1.1 (f)): Under Assumption C.1, each row of the matrix $Adj(I - P_\theta)$ is a left eigenvector of P_θ corresponding to the eigenvalue 1.

Thus, denoting the first row of $Adj(I - P_\theta)$ by $q_\theta = (q_\theta(1), \ldots, q_\theta(k))$, we have $\pi_\theta(i) = \frac{q_\theta(i)}{\sum_j q_\theta(j)}$. Since $Adj(I - P_\theta)$ is differentiable when P_θ is differentiable, consequently π_θ is differentiable. □

The preceding proof also shows that the bias function h_θ is differentiable under the same assumption. This is because under the recurrence and aperiodicity condition of the Markov chain, the bias function is equal to, in matrix notation,

$$h_\theta = (I - P_\theta + P_\theta^*)^{-1}(I - P_\theta^*)\, g_\theta,$$

where P_θ^* is the matrix with every row identical to π_θ (see Appendix A of Puterman [Put94]). Using these facts, one can thus derive gradient expressions simply by differentiating both sides of the optimality equation $\eta_\theta + h_\theta = g_\theta + P_\theta h_\theta$.

Appendix D

On Near-Optimality of the Set of Finite-State Controllers

Proposition 2.3 in Section 2.4.3 states that a necessary condition for a constant optimal liminf average cost function is the existence of history dependent near-liminf optimal policies. We now use Prop. 2.3 and the same line of analysis in its proof to prove that for a finite space POMDP, the set of finite-state controllers contains a near-optimal policy.

First we prove a lemma that will allow us to truncate at some finite stage a certain history dependent randomized policy that is near-liminf optimal, and form another sufficiently good policy that only uses the finite stage control rules of the former policy.

Recall that Π is the set of history dependent randomized policies. Each $\pi \in \Pi$ is a collection of conditional control probabilities $\{\mu_t\}_{t \geq 0}$, where μ_t maps the history h_t consisting of past observations and controls up to time t, to a measure on the control space. Recall also that the k-stage cost $J_k^\pi(\xi)$ of policy π is a linear function of ξ.

We define a set of state distributions. For all $s \in \mathcal{S}$, let $e_s \in \mathcal{P}(\mathcal{S})$ be the distribution that assigns probability 1 to state s, i.e., $e_s(\{s\}) = 1$. Abusing notation, we use J_-^* to denote the constant function value of $J_-^*(\cdot)$, when the latter is constant.

Lemma D.1. *Assume a finite state space $\mathcal{S}$. If $J_-^*(\cdot)$ is a constant function, then for any $\epsilon > 0$, there exists a policy $\pi_0 \in \Pi$ and an integer k_0 (depending on π_0) such that*

$$|\tfrac{1}{k_0} J_{k_0}^{\pi_0}(\xi) - J_-^*| \leq \epsilon, \qquad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Proof: Pick an arbitrary $\hat{\xi}$ in the relative interior of $\mathcal{P}(\mathcal{S})$, and let $C_{\hat{\xi}}$ be defined as in Lemma 2.1.

For any $\epsilon > 0$, let $\delta = \frac{\epsilon}{3C_{\hat{\xi}}} < \epsilon/3$. By Prop. 2.3, we can choose a policy $\pi \in \Pi$ such that

$$J_-^\pi(\xi) \leq J_-^* + \delta, \qquad \forall \xi \in \mathcal{P}(\mathcal{S}). \tag{D.1}$$

Since $\mathcal{S}$ is finite and by definition $J_-^\pi(\cdot)$ is the pointwise liminf of the functions $\{\frac{1}{k} J_k^\pi | k \geq 1\}$, there exists K_1 such that

$$\tfrac{1}{k} J_k^\pi(e_s) \geq J_-^\pi(e_s) - \delta \geq J_-^* - \delta, \qquad \forall s \in \mathcal{S}, \quad k \geq K_1.$$

Since J_k^π is a linear function, it follows that

$$\tfrac{1}{k} J_k^\pi(\xi) \geq J_-^* - \delta, \qquad \forall \xi \in \mathcal{P}(\mathcal{S}), \quad k \geq K_1. \tag{D.2}$$

For the $\hat{\xi}$ that we picked at the beginning, by the definition of $J_-^\pi(\cdot)$ and Eq. (D.1), there exists $K_2 > K_1$, and a $k_0 \geq K_2$ such that

$$\tfrac{1}{k_0} J_{k_0}^\pi(\hat{\xi}) \leq J_-^\pi(\hat{\xi}) + \delta \leq J_-^* + 2\delta. \tag{D.3}$$

By our choice of k_0 and Eq. (D.2), $\frac{1}{k_0} J_{k_0}^\pi(\xi) - J_-^* + \delta$ is a concave (since it is linear) and non-negative function. Therefore, applying Lemma 2.1, we have

$$\tfrac{1}{k_0} J_{k_0}^\pi(\xi) - J_-^* + \delta \leq C_{\hat{\xi}}\big(\tfrac{1}{k_0} J_{k_0}^\pi(\hat{\xi}) - J_-^* + \delta\big) \leq 3\, C_{\hat{\xi}} \delta \leq \epsilon, \qquad \forall \xi \in \mathcal{P}(\mathcal{S}),$$

where the second inequality is due to Eq. (D.3) and the third inequality due to our choice of δ. Combining the above relation with Eq. (D.2), we thus have

$$J_-^* - \epsilon \leq \tfrac{1}{k_0} J_{k_0}^\pi(\xi) \leq J_-^* + \epsilon, \qquad \forall \xi \in \mathcal{P}(\mathcal{S}).$$

Take the policy π_0 in the claim to be π, and the proof is complete. □

Suppose $\pi_0 = \{\mu_t\}_{t\geq 0}$. Let us form another policy $\pi_1 = \{\mu_t'\}_{t\geq 0}$ by repeating the control rules of π_0 from the start for every k_0-stage interval as follows. For any t, define $\bar{k}(t) = \mathrm{mod}(t, k_0)$, and define

$$\mu_t'(h_t, \cdot) = \mu_{\bar{k}(t)}(\delta_{\bar{k}(t)}(h_t), \cdot),$$

where $\delta_{\bar{k}(t)} : \mathcal{H}_t \to \mathcal{H}_{\bar{k}(t)}$ maps a length-t history h_t to a length-$\bar{k}(t)$ history by extracting the last length-$\bar{k}(t)$ segment of h_t.

By Lemma D.1, the k_0-stage average cost of π_0 is uniformly "close" to the optimal J_-^*. Hence, the liminf average cost $J_-^{\pi_1}$ of π_1 is also, evidently, uniformly close to the optimal J_-^*.

Corollary D.1. *Assume a finite state space $\mathcal{S}$. If $J_-^*(\cdot)$ is a constant function, then for any $\epsilon > 0$, there exists an integer k_0 (depending on ϵ), and an ϵ-liminf optimal policy $\pi_1 \in \Pi$ such that the control rule of π_1 at each stage depends functionally only on the history of the most recent k_0 stages.*

The above conclusions hold for finite state space models. We now consider finite space models, i.e., POMDPs with finite state, observation and control spaces. The controller π_1 has a finite-length of history window. So for finite space POMDPs, π_1 is equivalent to a finite-state controller with its internal state memorizing the current stage number modulo k_0 and the most recent length-k sample path with $k \leq k_0$. In a finite space POMDP governed by a finite-state controller, it is easy to see that the state and observation of the POMDP, and the internal state of the controller jointly form a time-homogeneous Markov chain. Thus by the MDP theory, for any initial distribution e_s, the liminf average cost and the limsup average cost are equal: $J_-^{\pi_1}(e_s) = J_+^{\pi_1}(e_s)$.

The function $J_+^{\pi_1}(\cdot)$ is convex (see Remark 2.2).[1] Hence for any initial distribution ξ,

$$J_+^{\pi_1}(\xi) \leq \sum_{s\in\mathcal{S}} \xi(s) J_+^{\pi_1}(e_s) = \sum_{s\in\mathcal{S}} \xi(s) J_-^{\pi_1}(e_s) \leq J_-^* + \epsilon.$$

[1] In fact in this case $J_+^{\pi_1}(\cdot) = J_-^{\pi_1}(\cdot)$ and both are equal to a linear function, as can be easily shown, either by the convexity of $J_+^{\pi_1}$ and concavity of $J_-^{\pi_1}$, or by the MDP theory.

Thus we have established that

$$J_{+}^{\pi_1}(\xi) \leq J_{-}^{*} + \epsilon, \qquad \forall \xi \in \mathcal{P}(\mathcal{S}),$$

which not only implies that π_1 is both ϵ-liminf and ϵ-limsup optimal, but also implies that the optimal limsup cost function $J_{+}^{*}(\cdot)$ is constant and $J_{+}^{*} = J_{-}^{*}$. As a summary, we have the following theorem.

Theorem D.1. *Assume a finite space POMDP and that J_{-}^{*} is constant. Then $J_{+}^{*} = J_{-}^{*}$, and for any $\epsilon > 0$ there exists a finite-state controller that is both ϵ-liminf and ϵ-limsup optimal.*

Remark D.1. From the preceding discussion, one can see that it is also possible to have conclusions analogous to Theorem D.1 for the more general case of a finite state space, and possibly infinite observation and control spaces, provided that one can establish that $J_{-}^{\pi_1}(e_s) = J_{+}^{\pi_1}(e_s)$ for the infinite-state Markov chain induced by the controller π_1 that has infinite number of internal states.

Remark D.2. When the constant average cost DP equation admits a bounded solution, or when it admits unbounded solutions in the sense defined and analyzed by [FGAM91], it is known that J_{-}^{*} and J_{+}^{*} are constant and equal. Then, by Theorem D.1, the class of finite-state controllers contains near-optimal policies. The conclusion of Theorem D.1 is stronger than this. Platzman [Pla80] showed an example (see Example 2.3 of Section 2.4.3), in which there is no solution to the DP equation while the optimal average cost is constant, and the optimal policy that is deterministic, is non-stationary. One can demonstrate trivially that there is an ϵ-optimal finite-state controller for Platzman's example.[2]

[2] Consider the following finite-state controller for Platzman's example. During the first N stages, it applies action 3 and accounts the number of times that observations 1 and 2 occurred. At time N, if more '1's than '2's have been observed, then it applies action 1 for the rest of the time; otherwise, it applies action 2 for the rest of the time. By the law of large number, clearly for any $\epsilon > 0$, there exists an N sufficiently large such that the policy is ϵ-optimal.

Bibliography

[AB02] D. Aberdeen and J. Baxter, *Internal-state policy-gradient algorithms for infinite-horizon POMDPs*, Tech. report, RSISE, Australian National University, 2002.

[ABFG$^+$93] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: a survey*, SIAM J. Control and Optimization **31** (1993), no. 2, 282–344.

[Alt99] E. Altman, *constrained Markov decision processes*, Chapman & Hall/CRC, 1999.

[Åst65] K. J. Åström, *Optimal control of Markov processes with incomplete state information*, Journal of Mathematical Analysis and Applications **10** (1965), 174–205.

[Åst69] ———, *Optimal control of Markov processes with incomplete state information, ii. the convexity of the loss function*, Journal of Mathematical Analysis and Applications **26** (1969), 403–406.

[BB96] S. J. Bradtke and A. G. Barto, *Linear least-squares algorithms for temporal difference learning*, Machine Learning **22** (1996), no. 2, 33–57.

[BB01] J. Baxter and P. L. Bartlett, *Infinite-horizon policy-gradient estimation*, J. Artificial Intelligence Research **15** (2001), 319–350.

[BBN03] D. P. Bertsekas, V. S. Borkar, and A. Nedić, *Improved temporal difference methods with linear function approximation*, Tech. Report LIDS 2573, MIT, 2003, also appear in "Learning and Approximate Dynamic Programming," IEEE Press, 2004.

[Ber75] D. P. Bertsekas, *Convergence of discretization procedures in dynamic programming*, IEEE Trans. Automatic Control **AC-20** (1975), 415–419.

[Ber01] ———, *Dynamic programming and optimal control*, second ed., vol. 2, Athena Scientific, 2001.

[BI96] D. P. Bertsekas and S. Ioffe, *Temporal differences-based policy iteration and applications in neuro-dynamic programming*, Tech. Report LIDS-P-2349, MIT, 1996.

[BMT01] V. S. Borkar, S. K. Mitter, and S. Tatikonda, *Optimal sequential vector quantization of Markov sources*, SIAM J. Control Optim. **40** (2001), no. 1, 135–148.

[Bon02] B. Bonet, *An epsilon-optimal grid-based algorithm for partially observable Markov decision processes*, The 19th Int. Conf. on Machine Learning, 2002.

[Bor00] V. S. Borkar, *Average cost dynamic programming equations for controlled Markov chains with partial observations*, SIAM J. Control Optim. **39** (2000), no. 3, 673–681.

[Boy99] J. A. Boyan, *Least-squares temporal difference learning*, The 16th Conf. on Machine Learning, 1999.

[BS78] D. P. Bertsekas and S. Shreve, *Stochastic optimal control: The discrete time case*, Academic Press, 1978.

[BT96] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*, Athena Scientific, 1996.

[Cas98] A. R. Cassandra, *Exact and approximate algorithms for partially observable Markov decision problems*, Ph.D. thesis, Dept. Computer Science, Brown University, 1998.

[CLZ97] A. R. Cassandra, M. L. Littman, and N. L. Zhang, *Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes*, The 13th Conf. UAI, 1997.

[CT91] T. M. Cover and J. A. Thomas, *Elements of information theory*, John Wiley & Sons, Inc., 1991.

[CW98] X.-R. Cao and Y.-W. Wan, *Algorithms for sensitivity analysis of Markov chains through potentials and perturbation realization*, IEEE Trans. Control Systems Technology **6** (1998), 482492.

[Den71] E. V. Denardo, *Markov renewal programs with small interest rates*, Annals of Mathematical Statistics **42** (1971), no. 2, 477–496.

[DM68] E. V. Denardo and B. L. Miller, *An optimality condition for discrete dynamic programming with no discounting*, Annals of Mathematical Statistics **39** (1968), no. 4, 1220–1227.

[Dud89] R. M. Dudley, *Real analysis and probability*, Chapman and Hall Mathematics Series, Chapman & Hall, 1989.

[DY79] E. B. Dynkin and A. A. Yushkevich, *Controlled Markov processes*, Springer-Verlag, 1979.

[FGAM91] E. Fernández-Gaucherand, A. Arapostathis, and S. I. Marcus, *On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes*, Annals of Operations Research **29** (1991), 439–470.

[GBB04] E. Greensmith, P. L. Bartlett, and J. Baxter, *Variance reduction techniques for gradient estimates in reinforcement learning*, J. Machine Learning Research **5** (2004), 1471–1530.

[GL95] P. W. Glynn and P. L'Ecuyer, *Likelihood ratio gradient estimation for regenerative stochastic recursions*, Advances in Applied Probability **27** (1995), no. 4, 1019–1053.

[HCA05] S-P. Hsu, D-M. Chuang, and A. Arapostathis, *On the existence of stationary optimal policies for partially observed MDPs under the long-run average cost criterion*, Systems and Control Letters (2005), to appear.

[HG01] S. G. Henderson and P. W. Glynn, *Approximating martingales for variance reduction in Markov process simulation*, Mathematics of Operations Research **27** (2001), 253–271.

[JSJ94] T. S. Jaakkola, S. P. Singh, and M. I. Jordan, *Reinforcement learning algorithm for partially observable Markov decision problems*, NIPS, 1994.

[Kon02] V. R. Konda, *Actor-critic algorithms*, Ph.D. thesis, MIT, Cambridge, MA, 2002.

[KT99] V. R. Konda and J. Tsitsiklis, *Actor-critic algorithms*, NIPS, 1999.

[Lau96] S. L. Lauritzen, *Graphical models*, Oxford University Press, 1996.

[LCK95] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, *Learning policies for partially observable environments: Scaling up*, Int. Conf. Machine Learning, 1995.

[LCK96] ———, *Efficient dynamic-programming updates in partially observable Markov decision processes*, Tech. Report CS-95-19, Brown University, 1996.

[Lov91] W. S. Lovejoy, *Computationally feasible bounds for partially observed Markov decision processes*, Operations Research **39** (1991), no. 1, 162–175.

[LR02] B. Lincoln and A. Rantzer, *Suboptimal dynamic programming with error bounds*, The 41st Conf. on Decision and Control, 2002.

[MPKK99] N. Meuleau, L. Peshkin, K.-E. Kim, and L. P. Kaelbling, *Learning finite-state controllers for partially observable environment*, The 15th Conf. UAI, 1999.

[MT01] P. Marbach and J. N. Tsitsiklis, *Simulation-based optimization of Markov reward processes*, IEEE Trans. Automatic Control **46** (2001), no. 2, 191–209.

[NB03] A. Nedić and D. P. Bertsekas, *Least squares policy evaluation algorithms with linear function approximation*, Discrete Event Dynamic Systems: Theory and Applications **13** (2003), 79–110.

[OG02] D. Ormoneit and P. Glynn, *Kernel-based reinforcement learning in average-cost problems*, IEEE Trans. Automatic Control **47** (2002), no. 10, 1624–1636.

[Pla80] L. K. Platzman, *Optimal infinite-horizon undiscounted control of finite probabilistic systems*, SIAM J. Control and Optimization **18** (1980), no. 4, 362–380.

[Put94] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, Inc., 1994.

[Ran05] A. Rantzer, *On approximate dynamic programming in switching systems*, The 44th Conf. on Decision and Control and European Control Conference 2005 (Seville), December 2005.

[Roc70] R. T. Rockafellar, *Convex analysis*, Princeton Univ. Press, 1970.

[Ros68] S. M. Ross, *Arbitrary state Markovian decision processes*, Annals of Mathematical Statistics **39** (1968), no. 6, 2118–2122.

[Ros70] ———, *Average cost semi-Markov decision processes*, Journal of Applied Probability **7** (1970), 649–656.

[RS94] W. J. Runggaldier and L. Stettner, *Approximations of discrete time partially observable control problems*, No. 6 in Applied Mathematics Monographs, Giardini Editori E Stampatori in Pisa, 1994.

[SB98] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, MIT Press, 1998.

[Sen73] E. Seneta, *Non-negative matrices*, John Wiley & Sons, 1973.

[Sen99] L. I. Sennott, *Stochastic dynamic programming and the control of queueing systems*, A Wiley-Interscience Publication, John Wiley & Sons, 1999.

[SJJ94] S. P. Singh, T. S. Jaakkola, and M. I. Jordan, *Learning without state-estimation in partially observable Markovian decision processes*, The 11th Conf. Machine Learning, 1994.

[SMSM99] R. S. Sutton, D. McAllester, S. P. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation*, NIPS, 1999.

[Son78] E. J. Sondik, *The optimal control of partially observable Markov decision problems over the infinite horizon: Discounted costs*, Operations Research **26** (1978), 282–304.

[SPS99] R. S. Sutton, D. Precup, and S. Singh, *Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning*, Artificial Intelligence **112** (1999), 181–211.

[STD02] S. Singh, V. Tadic, and A. Doucet, *A policy gradient method for SMDPs with application to call admission control*, Control, Automation, Robotics and Vision, 2002.

[Sut88] R. S. Sutton, *Learning to predict by the methods of temporal differences*, Machine Learning **3** (1988), 9–44.

[TK03] G. Theocharous and L. P. Kaelbling, *Approximate planning in POMDPs with macro-actions*, NIPS, 2003.

[TV97] J. N. Tsitsiklis and B. Van Roy, *An analysis of temporal-difference learning with function approximation*, IEEE Trans. Automatic Control **42** (1997), no. 5, 674–690.

[TV99] ———, *Average cost temporal-difference learning*, Automatica **35** (1999), no. 11, 1799–1808.

[YB04] H. Yu and D. P. Bertsekas, *Discretized approximations for POMDP with average cost*, The 20th Conf. UAI, 2004.

[Yu05] H. Yu, *A function approximation approach to estimation of policy gradient for pomdp with structured polices*, The 21st Conf. UAI, 2005.

[ZH01] R. Zhou and E. A. Hansen, *An improved grid-based approximation algorithm for POMDPs*, IJCAI, 2001.

[ZL97] N. L. Zhang and W. Liu, *A model approximation scheme for planning in partially observable stochastic domains*, J. Artificial Intelligence Research **7** (1997), 199–230.