CONSTRAINED MARKOV DECISION PROCESSES

Eitan ALTMAN
INRIA
2004 Route des Lucioles, B.P.93
06902 Sophia-Antipolis Cedex
France

To Tania and Einat

Preface

In many situations in the optimization of dynamic systems, a single utility for the optimizer might not suffice to describe the real objectives involved in the sequential decision making. A natural approach for handling such cases is that of optimization of one objective with constraints on other ones. This allows in particular to understand the tradeoff between the various objectives.

In order to handle multi-objective dynamic decision making under uncertainty, we have chosen the framework of controlled Markov chains, which has already proven to be quite powerful in many applications studied in the last half century. In particular, this approach allows us to solve stochastic dynamic control problems by using some finite linear programs, in the case where the system can be described by a finite number of states and the decision maker disposes of a finite number of decision actions. This case is presented in the first part of this book.

More complex systems that cannot be described using a finite number of states or decision actions are treated in the second part of the book; we present two main approaches that allow us to handle such systems: the so called "negative dynamic programming" approach in which the costs are assumed to be bounded below, and an approach based on uniform Lyapunov function techniques.

In some cases, systems with an infinite number of states can be approximated by finite systems, which allows us to obtain a good policy for the original problem by solving a simpler control problem. This approach, as well as many other approximation issues are presented in the third part of this book.

Writing this book turned out to be a rich and interesting constrained control problem in itself. The objectives were not always easy to quantify and many evident constraints came out, such as time and page limitations. With the help of the theory developed here as well as the warm support of my wife, TANIA, we were finally able to meet the constraints and present a solution, that we hope you will enjoy reading.

Eitan Altman, August 1998

Contents

CHAPTER 1

Introduction

The aim of this monograph is to investigate a special type of situation where one controller has several objectives. Instead of introducing a single utility that is to be maximized (or a cost to be minimized) that would be some function (say, some weighted sum) of the different objectives, we consider a situation where one type of cost is to be minimized while keeping the other types of costs below some given bounds. Posed in this way, our control problem can be viewed as a constrained optimization problem over a given class of policies.

By specifying control rather than optimization problems, we have in mind models of dynamic systems, where decisions are taken sequentially. We distinguish between a control action, which is a decision taken at a given time, and a whole policy, which is a rule for selecting actions as a function of time and of the information available to the controller. In fact, for a given policy, the choice of actions at different decision epochs, may depend on the whole observed history, as well as other external 'randomization' mechanisms. A choice of a policy will determine (in some probabilistic sense) the evolution of the state of the system which we control. The trajectories of the states together with the choices of actions (or trajectories' distribution) determine the different costs.

In order to clarify the type of problems that we consider, we present in the following section a number of applications of constrained dynamic control problems. Most of the applications below are from the field of telecommunications.

1.1 Examples of constrained dynamic control problems

Telecommunications networks are designed to enable the simultaneous transmission of heterogeneous types of information: file transfers, interactive messages, computer outputs, facsimile, voice and video, etc... At the access to the network, or at nodes within the network itself, the different types of traffic typically compete for a shared resource. Typical performance measures are the transmission delay, the throughputs, probabilities of losses of packets (that stem from the fact that there are finite buffers at intermediate nodes of the network), etc... All these performance measures are determined by continuously monitoring and controlling the input flows

into the network, by controlling the admission of new calls (or sessions), by controlling the allocation of the resources to different traffic, by routing decisions. Different types of traffic differ from each other by their statistical properties, as well as by their performance requirements. For example, for interactive messages it is necessary that the average end-to-end delay be limited. Strict delay constraints are important for voice traffic; there, we hardly distinguish between different delays as long as they are lower than some limit of the order of 0.1 second. When the delay increases beyond this limit, it becomes quickly intolerable. For non-interactive file transfer, we often wish to minimize delays or to maximize throughputs.

Controllers of telecommunication systems have often been developed using heuristics and experience. However, there has been a tremendous research effort to solve such problems analytically. Here are some examples:

(1) *The maximization of the throughput of some traffic, subject to constraints on its delays.* A huge amount of research in this direction was started up by Lazar (1983) and has been pursued and developed by himself together with other researchers; some examples are Bovopoulos and Lazar (1991), Hsiao and Lazar (1991), Vakil and Lazar (1987), Korilis and Lazar (1995a, 1995b). In all these cases, limit-type optimal policies were obtained (known as window flow control). Koole (1988) and Hordijk and Spieksma (1989) considered the problem of Lazar (1983) as well as other admission control problems within the framework of Markov Decision Processes (MDPs), and discovered that for some problems, optimal policies are not of a limit-type (the so called 'thinning policies' were shown to be optimal under some conditions).

We shall study in Chapter 5 a discrete time model that extends the framework of the above problems and also includes service control. The latter control can model bandwidth assignment or control of quality of service. The flow control has the form of the control of the probability of arrivals at a time slot. The control of service is modeled by choosing the service rate, or more precisely, by assigning the probability of service completion within a time slot. A tradeoff exists between achieving high throughput, on the one hand, and low expected delays on the other. We further assume that there are costs on the service rates. The problem is formulated as a constrained MDP, where we wish to minimize the costs related to the delay subject to constrained on the throughputs and on the costs for service.

(2) *Dynamic control of access of different traffic types.* A pioneering work by Nain and Ross (1986) considered the problem where several different traffic types compete for some resource; some weighted sum of average delays of some traffic types is to be minimized, whereas for some other traffic types, a weighted sum of average delays should be bounded by some given limit. This research stimulated further investigations; for example, Altman and Shwartz (1989) who considered several constraints and Ross

and Chen (1988) who analyzed the control of a whole network. The typical optimal policies for these types of models requires some randomization or it is based on time-sharing between several fixed priority policies.

(3) *Controls of admission and routing in networks.* Feinberg and Reiman (1994) have solved the problem of optimal admission of calls of two types into a multi-channel system with finite capacity. They established the optimality of a randomized trunk reservation policy.

Other problems in telecommunications which have been solved by constrained MDPs are reported in Maglaris and Schwartz (1982), Beutler and Ross (1986) and Bui (1989). A study of a constrained control problem in a queueing model with a removable server, with possible applications in telecommunications or in production, was done by Feinberg and Kim (1996).

Constrained MDPs (CMDPs) have had an important impact in many other areas of applications:

1. In Kolesar (1970), a problem of hospital admission scheduling is considered.
2. Golabi *et al.* (1982) have used CMDPs to develop a pavement management system for the state of Arizona to produce optimal maintenance policies for a 7400-mile network of highways. A saving of 14 million dollars was reported in the first year of implementation of the system, and a saving of 101 million dollars was forecast for the following four years.
3. Winden and Dekker (1994) developed a CMDP model for determining strategic building and maintenance policies for the Dutch Government Agency (Rijksgebouwendienst), which maintains 3000 state-owned buildings with a replacement value of about 20 billion guilders and an annual budget of some 125 million guilders.

1.2 On solution approaches for CMDPs with expected costs

We focus in this section on models where all the cost objectives in the constrained problem are specified in terms of expectations of some functionals of the state and action trajectories. We describe some approaches to solve such CMDPs, briefly surveying the existing literature.

Several methods have been used in the past to solve this kind of CMDP. The first one, based on a Linear Program (LP), was introduced by Derman and Klein (1965), Derman (1970), and further developed by Derman and Veinott (1972), Kallenberg (1983), and Hordijk and Kallenberg (1984). It is based on an LP whose decision variables correspond to the occupation measure. The value of the LP is equal to the value of the CMDP, and there is a one to one correspondence between the optimal solutions of the LP and the optimal policies of the CMDP. This method is quite efficient (in terms of complexity of computations, and in the amount of decision variables, and

hence memory requirements) for calculating the value of the CMDP (for the finite state and action space) for both the discounted or total cost, as well as the average cost with unichain structure. However, for the expected average cost with general multi-chain ergodic structure, the computation of an optimal policy is very costly and, as stated by Kallenberg (1983), it 'is unattractive for practical problems. The number of calculations is prohibitive' (p. 142). An alternative efficient way (again, in terms of complexity of calculations and memory requirements) for obtaining optimal policies from the LP for the average cost was obtained by Krass (1989). In Chapters 8, 10 and 11 we present the extension of the LP approach to the case of countable state space. (This is based on Altman and Shwartz, 1991a, and Altman, 1994, 1996, 1998).

A second method was introduced by Beutler and Ross (1985, 1986) for the case of a single constraint, and is based on a Lagrangian approach. It allowed them to characterize the structure of optimal policies for the constrained problem, but it does not provide explicit computational tools. This approach was extended by Sennott (1991, 1993) to the countable state space. The use of Lagrangian techniques for several constraints is quite recent (see e.g., Arapostathis *et al.*, 1993, Piunovskiy, 1993, 1994, 1995, 1996, 1997a, 1997b, and Altman and Spieksma, 1995), and has not been much exploited.

A third method, based on an LP, was introduced in Altman and Shwartz (1989, 1993) and further studied by Ross (1989). It is based on some mixing (by a time-sharing mechanism) of stationary deterministic policies (these are policies that depend only on the current state and do not require randomization). A similar LP approach was later introduced by Feinberg (1993) for finite MDPs (finite state and action spaces), where the mixing is done in a way that is equivalent to having an initial randomization between stationary deterministic policies. These approaches require in general a huge number of decision variables. However, there are special applications where this LP can have an extremely efficient solution, and has been used even for problems with an infinite state space (see Altman and Shwartz, 1989), in the case where one can eliminate *a priori* many suboptimal stationary deterministic policies. In both the time-sharing approach in Altman and Shwartz (1989, 1993), as well as in the randomization approach described in Feinberg (1995), only mixing of finitely many policies was considered. (This is indeed sufficient in the case of finite MDPs, i.e., finite state and action spaces, since, in that case, there are only a finite number of stationary deterministic policies.)

Strong connections exist between the three solution methods. Understanding these connections enables us to obtain a unified theory for CMDPs. It also enables us to generalize the second approach to several constraints. Finally, it allows us to obtain many asymptotic results on convergence of the values and policies of some sequence of CMDPs to those of a limit

CMDP, in particular, convergence in the discount factor, in the horizon, and convergence of finite state approximations (we present these in Chapters 13–16).

An LP that computes the solution of CMDPs for all discount factors simultaneously was introduced in Altman *et al.* (1996). Although the decision variables are not the standard ones (they are the set of functions that are represented as the ratio between two polynomials with real coefficients), a solution is derived in a finite number of steps.

1.3 Other types of CMDPs

The type of cost criteria and solution approaches surveyed in the previous section are those most frequently studied. However, many other models of constrained MDPs have been investigated. These can be classified according to different types of cost criteria, according to different assumptions on the controller (one or more controllers) assumptions on the available information (the adaptive problem). We briefly describe these in this section.

A generalization of the framework introduced in the previous section is to allow different cost criteria to have different discount factors. The solution of such CMDPs is significantly more complex, requiring much more computational effort. They do not possess optimal stationary policies. The analysis and characterization of such CMDPs was presented by Feinberg and Shwartz (1995). In particular, they show that there exists an optimal policy which is ultimately stationary (i.e., it becomes stationary deterministic after some fixed time) and requires no more than K randomizations. This extends the results by Koole (1988), Ross (1989) and Borkar (1994). Another related result can be found in Feinberg and Shwartz (1996).

Ross and Varadarajan (1989, 1991) have considered problems where a constraint is imposed on the actual sample-path cost. In fact, Ross and Chen (1988) point out that the model where all costs are defined by expectations is inappropriate for some telecommunications problems, namely for problems involving voice interactive transmission: 'We remark that the model studied here would not be appropriate if real-time voice packets were also competing for the resource. This is because [the CMDP] imposes constraints on the average delay ... and not on the actual delay.' This type of constrained problem was solved by Ross and Varadarajan (1989, 1991) using again an LP approach. An interesting feature of this formulation is that ε-optimal stationary policies exist (for finite MDPs) even under the general multi-chain ergodic structure. This is in contrast to the problem where all costs are defined through expectations. Moreover, the computation of the value and the ε-optimal policy is much simpler than for the problem with expected costs. Some other results on sample-path costs (both in the constraint and in the objective function) can be found in Altman and Shwartz (1991d). Haviv (1995) raised an important criticism

on the formulation of MDPs through expected costs: they do not satisfy Bellman's principle of optimality. Haviv shows that the sample-path constrained formulation of the constrained MDP does not suffer from this drawback.

There are alternative ways to make the costs more sensitive to deviations from the expectation. One way to achieve this goal is to have some *additional cost related to the variance*. Sobel (1985) proposed to maximize the mean to variance ratio with constraints on the mean. Other approaches were proposed and analyzed in Filar and Lee (1985), Kawai (1987), Bayal-Gursoy and Ross (1992) and Filar *et al.* (1989). A unified approach which extends the above ones was presented by Huang and Kallenberg (1994) and solved using an algorithm based on parametric-linear programming. The case of infinite state space was analyzed by Altman and Shwartz (1991a). Other recent papers in this topic are Sobel (1994) and White (1994).

Another way to penalize deviations of the costs from the expectation is to introduce some constraints on the *rate of convergence*. This approach was investigated by Altman and Zeitouni (1994).

A problem with another type of constraint, namely on the probability that some conditional expected cost be bounded, was solved by White (1988).

There have been some results on extending constrained MDPs to the case of more than one controller (stochastic games). In the case of N controllers with different objectives, a set of coupled linear programs was shown in Altman and Shwartz (1995) to provide a Nash equilibrium (which is used as the concept of optimality when there is more than one controller under the assumption that the controllers are selfish and do not cooperate). It is shown that a Nash equilibrium exists among the stationary policies. This work was motivated by a problem in telecommunication that was solved in Korilis and Lazar (1995a).

The case of two controllers ('players') with constraints and with conflicting objectives was solved by Shimkin (1994), using geometric ideas based on extensions of Blackwell's approachability theory. In that setting, optimal policies turned to be non-stationary in an essential way.

An important problem in MDPs in general, and in constrained MDPs in particular, which is often encountered in applications, is of simultaneous learning and controlling. This occurs when some parameters of the problem are unknown to the decision maker. The standard cost criteria may be quite unsuitable for this type of situation. For example, the total expected discounted cost may not be well defined if we do not have any knowledge of the probability distribution. This required the introduction of new cost criteria. Schäl (1975) introduced an asymptotic discounted cost criterion for non-constrained MDPs, for which adaptive optimal policies combining estimation and control were investigated (Schäl, 1987, Hernandez-Lerma, 1989, and references therein). Altman and Shwartz (1991d) adapted

these cost criteria to CMDPs and proposed several optimal adaptive techniques (1991b, 1991d). The solutions are based on ideas on sensitivity analysis of linear programs. An alternative solution approach based on stochastic approximations can be used to solve the adaptive MDP. This approach was used by Makowski and Shwartz (1992), Ma *et al.* (1992), Ma and Makowski (1988, 1992).

1.4 Cost criteria and assumptions

We focus in this monograph on three main types of cost criteria. The first one is the total expected cost until some target set of states is reached. If the target set is empty then this criterion is merely the sum of expected instantaneous costs accumulated over an infinite horizon. The second cost criterion is the infinite horizon discounted cost. It can be obtained directly from our analysis of the first cost criterion. The third cost criterion is the limit (as the time t becomes large) of the expected total cost until time t, averaged over the time. All cost criteria are defined precisely in Chapters 2 and 6.

Many properties and results do not carry on, in general, from finite MDPs (those with finitely many states and actions) to infinite ones as many counter-examples will illustrate. If we wish to obtain the optimal value and policies for the constrained MDP using linear programming techniques when dealing with infinite MDPs, we need to restrict to one of several possible frameworks where some assumptions are made on the probabilistic structure and on the immediate costs.

When using the total cost criteria we consider one of three types of MDPs:

1. The transient MDPs, for which the total expected time spent in each state is finite under any policy. When analyzing this class of MDPs, we shall often assume that the immediate cost are bounded below.

2. The MDPs with uniform Lyapunov function, which are absorbing (the total expected 'life-time' of the system is finite under any policy). These MDPs are a subclass of the transient ones. When analyzing this class of MDPs, we shall not require that the immediate cost are bounded below, and replace this by a much weaker assumption.

3. Contracting MDPs, which are a further subclass of MDPs with uniform Lyapunov function.

All three types of MDPs are equivalent for finite MDPs (finite state and action spaces), as was shown by Kallenberg (1983); this is however not the case in the countable state space.

For the expected average cost criteria we consider very similar frameworks:

1. The first allows for quite general probabilistic assumptions, in particular, a tightness assumption, and yet requires the immediate costs to be bounded from below; alternatively, even the tightness assumption may be relaxed and replaced by some stronger growth condition on the cost. This approach is due to Borkar (1983) and was adapted to constrained MDPs in Altman and Shwartz (1991a).
2. In the second framework we relax the boundedness on the immediate cost and require instead some tightness conditions as well as some uniform integrability ones. This framework will be shown to be equivalent to having a uniform Lyapunov function (due to Hordijk, 1977).
3. A further subclass of MDPs with uniform Lyapunov function that we shall briefly study is that of uniformly μ-recurrent MDPs, who were introduced and investigated by Dekker and Hordijk (1988), Spieksma (1990), and Dekker *et al.* (1994). This framework can be considered as the one corresponding to contracting MDPs.

We mention finally that Lyapunov functions are known to have an important role in dynamic systems and in control theory (not only in stochastic control): these are used as test functions to obtain stability properties. They are often used in the study of (non-controlled) Markov chains as a tool to establish ergodicity properties, see Meyn and Tweedie (1994).

The reasons for introducing the various frameworks and the necessity of the assumptions there will be further discussed in Section 1.7, which introduces the methodologies that we follow in this book.

1.5 The convex analytical approach and occupation measures

Our first analysis approach is based on the the properties of the set of occupation measures achievable by different classes of policies. Under some conditions, an occupation measure achievable by a policy has the property that for any given instantaneous costs, the cost criteria (i.e., the total expected cost or the expected average cost) can be expressed as the expectation of that instantaneous cost with respect to the corresponding occupation measure.

The convexity and compactness properties of these sets turn out to be essential in the study of constrained MDPs. We derive these properties for finite MDPs in the beginning of Chapters 3 and 4, and obtain the corresponding properties for infinite MDPs in the beginning of Chapter 8 for the total cost, and in the beginning of Chapter 11 for the expected average cost.

This type of analysis of occupation measure goes back to Derman (1970) who also made use of it for studying constrained MDPs (in finite state and action spaces). It was further developed by Kallenberg (1983) and Hordijk and Kallenberg (1984), and Feinberg (1995) (who considered the

semi-Markov case). The properties of occupation measures corresponding to the infinite state space were investigated by Borkar (1988, 1990), Altman and Shwartz (1988, 1991a), Altman (1994, 1996, 1998), Spieksma (1990), and Feinberg and Sonin (1993, 1995). The study of occupation measures arises also in other related areas in control. In particular, in the controlled diffusions they have already been studied by Krylov (1985) and later by Borkar and Ghosh (1990, 1993).

For the different cost criteria, the objectives turn out to be linear in the occupation measures under suitable conditions, at least for some 'good classes of policies' (such as stationary policies). An important corollary of this property is that the original control problem can be reduced to a Linear Program (LP), which we shall call the 'primal LP', where the decision variables are measures (corresponding to the occupation measures). Moreover, optimal solutions of the LP determine optimal stationary policies through induced conditional occupation measures. We present these LPs and establish their equivalence to the original control problem in Chapters 3 and 4 for finite MDPs, and obtain similar representation at the end of Chapter 8, (the total cost), and of Chapter 11, (the expected average cost) for infinite MDPs.

This approach goes back to Derman (1970) and was further developed by Derman and Veinott (1972) by Kallenberg (1983) and Hordijk and Kallenberg (1984). Its derivation for the infinite state case is due to Altman and Shwartz (1991a) and Borkar (1990) (the expected average cost) and Altman (1994, 1996, 1998) (the discounted and total cost).

In order to obtain an equivalent LP, one has first to identify classes of 'dominant' policies, i.e., classes of policies which are sufficiently rich in order to allow us to restrict ourselves to them for the search of optimal policies. Under fairly general conditions, the problem of whether a subclass of policies is dominant is related to whether this subclass is 'complete', i.e., whether any occupation measure that is achievable by some general policy can also be achieved (or outperformed, in some sense) by some policy within that subclass of policies.

This property motivates us to raise the question of whether the class of stationary policies is complete.

For the total cost, for MDPs with a uniform Lyapunov function, we show that both the stationary policies as well as the mixed stationary-deterministic are complete. Surprisingly, this result turns out not to hold for the more general transient MDPs. Indeed, counter-examples have been presented recently by Feinberg and Sonin (1995). However, we show that the set of stationary policies turns out to have the following property. For any occupation measure achievable by some policy u, there is a stationary policy that achieves an occupation measure that is smaller than or equal to the one achieved by u. These results, obtained in Altman (1996, 1998), are presented in Chapter 8.

For the expected average cost criterion there are cases and counter-examples where stationary policies do not achieve all possible occupation measures. This may occur either due to a multi-chain ergodic structure (see Hordijk and Kallenberg, 1984, for the case of finite state and actions), or, in the infinite case, due to non-tightness (see Borkar, 1990, Chapter 5, Altman and Shwartz, 1991a, and Spieksma, 1990). However, under some conditions on the ergodic structure, we show that the set of stationary policies is 'weakly complete'; by that we mean that for any occupation measure that is achievable by some policy there exists some stationary policy which achieves the same measure up to a multiplicative constant. This property, together with some growth conditions on the costs, imply that the stationary policies are dominant. These results, some of which were obtained by Borkar (1990), Altman and Shwartz (1991a), are presented in Chapter 11.

1.6 Linear Programming and Lagrangian approach for CMDPs

We begin by presenting a brief survey of the LP approach for non-constrained MDPs. The use of LPs started already in the beginning of the sixties, with the pioneering work of D'Epenoux (1960, 1963), who considered the discounted cost case, and of De Ghellinck (1960) and Manne (1960) who studied the expected average cost (with the unichain condition). The analysis via LPs, of the expected cost with the general multi-chain ergodic structure, has been presented by Denardo and Fox (1968) and Denardo (1970). Hordijk and Kallenberg (1979) presented a single LP for solving the multi-chain expected average problem. For a further survey of LP techniques for the non-constrained MDPs, see Kushner and Kleinman (1971), Heilmann (1977, 1978), Arapostathis *et al.* (1991), Puterman (1994) and Kallenberg (1994). An important contribution to generalization of the LP techniques to infinite state and action spaces is due to Lasserre (1995) who applied functional analytical tools, using the theory of infinite dimensional LPs (Anderson and Nash, 1987). Lasserre handles both the primal and dual LPs, establishes conditions for their solvability and for the absence of a duality gap, and presents conditions for the optimality of a stationary policy that is obtained using the solution to the primal LP. This work was extended in Hernández-Lerma and Lasserre (1994, 1995) and Hernández-Lerma and Hernández-Hernández (1994) to the case of non-countably infinite state and action spaces, and in Hordijk and Lasserre (1994) to the multi-chain expected average case. An alternative approach to derive the LP was obtained by Altman and Shwartz (1991a), Altman (1994, 1996) and Spieksma (1990) using probabilistic techniques, and these were obtained directly for the constrained MDPs. Finally, the LP approach, in particular, and Mathematical Programming approaches, in general have been used also in the case of more than one controller (i.e., stochastic games), see e.g., the survey by Raghavan and Filar (1991).

The problem of minimizing a single objective (the total expected cost, or the expected average cost) with no constraints can be handled by solving a system of dynamic programming equations, known as the Bellman optimality equations. These transform the problem of minimization over the class of all policies into a set of coupled minimization problems over the (much smaller) sets of actions. These dynamic programming equations may be the starting point for obtaining the LP formulation. Under suitable conditions, the value function is the *largest* 'super-harmonic function': these are functions that satisfy some optimality inequalities (obtained directly from the optimality equations) for all states and actions. This provides the LP which is dual to the one obtained using the convex analytical approach of occupation measures (which we described in the previous section). This approach is the basis of the derivation of the LPs by Kallenberg (1983) and Hordijk and Kallenberg (1979).

In the case of constrained MDPs, one can still derive directly the dual LP by using a Lagrangian approach, and then applying some minmax theorem. Indeed, the Lagrangian approach allows us to transform a constrained control problem into an equivalent minmax non-constrained control problem. If a saddle point property is shown to hold, then the problem is transformed into a maxmin problem, which can be solved using an LP. This direct derivation of the dual LP was obtained by Altman and Spieksma (1995) for the case of finite state and action spaces. In Chapters 9 and 12 we describe this approach for obtaining the dual LP (for the total cost and the expected average cost, respectively).

The Lagrangian approach turns out to be not only a tool for obtaining an LP formulation, but has its own merits. It turns out to be very useful for sensitivity analysis and for obtaining asymptotical properties of constrained MDPs; it allows us to obtain in Chapter 13 theorems for approximations of the value and policies for CMDPs, which we apply to the study of the convergence in the discount factor (especially, in the neighborhood of 1, see Chapter 14), the convergence in the horizon (Chapter 15) as well as to the study of state-truncation techniques (Chapter 16). In particular, it allows us to obtain an estimate of the approximation error. All these results are obtained for the contracting framework, and most of them are obtained also for the more general setting of uniform Lyapunov functions. An alternative approach for approximations is illustrated in Section 9.6, where state truncation is used for computing the value and optimal stationary policies of the CMDP in the case of non-negative immediate costs.

An alternative LP approach (which can also be obtained by the Lagrangian technique) is the one that corresponds to the restriction of the constrained problem to mixed stationary-deterministic policies. The fact that these policies are dominating is established in Chapters 8 and 11, so that the restriction is without loss of optimality. The decision variables here are the measures over all stationary deterministic policies. An advantage of

such formulation is that, even when the ergodic structure is general multi-chain, the same type of Linear Program applies for the expected average cost as well as the discounted cost. This fact allowed Tidball and Altman (1996b) to obtain convergence of the values and policies of discounted CMDP to those of expected average MDPs, as the discount factor tends to 1, for a general multi-chain structure. This approach extends the one by Feinberg (1993) that was derived for the case of finite state and action spaces. The LP has the same form as the one introduced by Altman and Shwartz (1993) for computing optimal time-sharing policies. We present these LPs at the end of Chapters 9 and 12.

1.7 About the methodology

We describe in this section the structure of the book, and explain the methodologies used in the future chapters. We shall illustrate some of the main ideas of the book by presenting basic results, without proof, for the discounted cost problem, for the case of finite state and action spaces. This will allow us to explain the type of assumptions needed later, and the framework for developing the theory of countable state space and compact sets of actions.

We consider discrete-time Markov chains whose transition probabilities depend on some parameters, called the actions. The state at time t as well as the action chosen at time t determine both the transition probabilities at that time as well as the value of several instantaneous costs to be paid at that time. Actions are chosen according to some decision rule, possibly randomized, which we call a *policy*. It may depend on the current state of the Markov chain, on the current time, but also on any other information available to the decision maker, such as previous states and previous chosen actions.

The basic constrained optimization problem (COP) that we study in this monograph has the form

$$\mathbf{COP}: \min_{u\in U} C(u) \;\text{ subject to }\; D^k(u) \le V_k,\; k=1,\ldots,K,$$

where $V_k, k = 1,\ldots,K$ are some given constants; $C(u)$ and $D^k(u)$, $k = 1,\ldots,K$ are some cost criteria related to a policy u (through the expected instantaneous costs they generate), and we minimize over some large class U of history-dependent policies. These costs will stand for one of the following cost criteria: the total expected cost until the state reaches some set $\mathcal{M}$ (Chapters 8 and 9), the discounted cost (Chapters 3 and 10), or the expected average cost (Chapters 4, 11 and 12). (Precise definitions of controlled Markov chains and of the cost functions will appear in Chapters 2 and 6.)

As a first step in our investigation, we shall focus on a deeper understand-

ing of policies. The space U of all history-dependent policies might be 'too large'; moreover, some of the policies that it contains may be hard to implement, e.g., if they require much memory to remember states and actions of the past. So some attention will be given to the question of identifying smaller classes of policies which are dominating, i.e., their performances (in terms of the costs they achieve) are as good as those achieved by policies in U. We shall show in Chapters 2 and 6 that *Markov policies* (in which decisions depend only on the current state and current time), and *quasi-Markov policies* (in which the decisions depend only on the current state and the number of transitions that have occurred) are dominating classes of policies. Under further conditions, we shall show later, when analyzing each cost criterion in detail, that stationary policies (in which the decisions depend only on the current state) and mixed stationary-deterministic policies (in which we choose at random between some subclass of stationary policies) are dominating.

In our analysis of policies, we shall show that one cannot improve the performance by adding extra randomizations at each step, on which decision rules may depend.

For each of the cost criteria that we study, we present three alternative approaches.

The first approach is based on occupation measures. For each given policy u, one can define a measure $f(u)$ with the property that the actual cost to be minimized can be represented as the expectation (or integral) of the immediate cost with respect to that measure. The set of all achievable measures is identified, and is shown to be a polytope. By identifying this polytope, we are then able to present an LP whose value equals the value of the control problem, and whose optimal solutions define the optimal policies (through the occupation measures that they generate).

A second approach is based on ideas of dynamic programming. Dynamic programming is an efficient tool for solving non-constrained optimal control problems, as it allows us to transform a minimization over all policies to a set of minimizations over the (much smaller) set of actions. In order to use dynamic programming techniques for constrained MDPs, we use a Lagrangian approach which transforms a *constrained minimization problem* into an inf-sup problem of the Lagrangian (the Lagrangian is the sum of the original cost to be minimized and all the other constraints, weighted by some constants $\lambda_k, k = 1, \ldots, K$ called Lagrange multipliers). The sup is then taken over all non-negative values λ of the Lagrange multipliers, and the inf is taken over the class of all control policies. By invoking a saddle-point theorem, we are able to change the order of the inf and the sup, and obtain a sup-inf problem instead. The new problem is more familiar, since it involves first minimizing with respect to the policies, and only then maximizing with respect to λ. For each fixed λ we are faced with a standard non-constrained problem of a controlled Markov chain, and we can

therefore obtain the minimization (the inf) through well-known dynamic programming (or linear programming) techniques. At this point, we show how to solve the inf-sup problem using a single LP, which turns out to be the dual of the one obtained by the first approach. We illustrate the use of this approach in Chapter 5.

The third approach is based on identifying an optimal policy among mixed stationary-deterministic policies. This is done using yet another LP, whose decision variables are the initial randomization measure over the set of stationary-deterministic policies. We introduce this method only in the second part of the monograph.

We now illustrate the first two approaches through a constrained MDP with the discounted cost criterion. We consider a finite set A of actions and a finite set $\mathbf{X}$ of states, and denote by $\mathcal{P}_{xay}$ the probability to move from state x to state y if action a is used. c and $d^k, k = 1, \ldots, K$ are some given immediate cost functions from $\mathbf{X} \times \mathsf{A}$ to $\mathbb{R}$. Each initial state x and policy u define a probability measure P_x^u over the state and action trajectories. For a given initial state x and a policy u, define the discounted costs

$$\begin{aligned} C_\alpha(x,u) &\stackrel{\text{def}}{=} (1-\alpha)\sum_{t=1}^{\infty}\alpha^{t-1}E_x^u\, c(X_t,A_t), \\ D_\alpha^k(x,u) &\stackrel{\text{def}}{=} (1-\alpha)\sum_{t=1}^{\infty}\alpha^{t-1}E_x^u\, d^k(X_t,A_t)\,, \quad k=1,\ldots,K. \end{aligned}$$

X_t and A_t are the (random) state and action at time t.

The costs can be written as

$$\begin{aligned} C_\alpha(x,u) &= \sum_{y\in\mathbf{X}}\sum_{a\in\mathsf{A}} f_\alpha(x,u;y,a)c(y,a)\ , \\ D_\alpha^k(x,u) &= \sum_{y\in\mathbf{X}}\sum_{a\in\mathsf{A}} f_\alpha(x,u;y,a)d^k(y,a),\ k=1,\ldots,K, \end{aligned}$$

where

$$f_\alpha(x,u;y,a) \stackrel{\text{def}}{=} (1-\alpha)\sum_{t=1}^{\infty}\alpha^{t-1}P_x^u(X_t=y,A_t=a).$$

The vector $f_\alpha(x,u)$ is called the occupation measure corresponding to u and to the initial state x. For any policy, it belongs to the set of measures ρ that satisfy

$$\begin{aligned} &\sum_{y\in\mathbf{X}}\sum_{a\in\mathsf{A}}\rho(y,a)(1\{v=y\}-\alpha\mathcal{P}_{yav}) = (1-\alpha)1\{x=v\}, \quad \forall v\in\mathbf{X} \\ &\sum_{y\in\mathbf{X}}\sum_{a\in\mathsf{A}}\rho(y,a)=1, \quad \rho(y,a)\geq 0, \forall y,a. \end{aligned} \tag{1.1}$$

Moreover, we show in Chapters 3 and 10 that any ρ satisfying (1.1) equals

the occupation measure corresponding to any stationary policy w that satisfies the following: for any state y for which $\sum_{a' \in \mathsf{A}} \rho(y, a') > 0$, w chooses action a at state y with probability

$$w_y(a) = \frac{\rho(y, a)}{\sum_{a' \in \mathsf{A}} \rho(y, a')}.$$

Thus, the constrained problem **COP** is equivalent to the LP:

$$\min_{\rho} \sum_{y \in \mathbf{X}} \sum_{a \in \mathsf{A}} c(y, a) \rho(y, a) \tag{1.2}$$

subject to (1.1) and

$$\sum_{y \in \mathbf{X}} \sum_{a \in \mathsf{A}} d^k(y, a) \rho(y, a) \leq V_k, \quad k = 1, \ldots, K.$$

Next, we describe the approach based on dynamic programming. For the case that $K = 0$ (i.e., no constraints), the value $C_\alpha(x) \stackrel{\text{def}}{=} \inf_{u \in U} C_\alpha(x, u)$, as a function of x, is known (see Chapter 3) to be the unique solution of the following dynamic programming equation:

$$\phi(x) = \min_{a \in \mathsf{A}} \left[(1 - \alpha) c(x, a) + \alpha \sum_{y} \mathcal{P}_{xay} \phi(y) \right] \quad \forall x \in \mathbf{X}.$$

C_α therefore satisfies the inequalities

$$\phi(v) \leq (1 - \alpha) c(v, a) + \alpha \sum_{y} \mathcal{P}_{vay} \phi(y) \quad \forall v \in \mathbf{X}, a \in \mathsf{A}. \tag{1.3}$$

The set of functions satisfying these inequalities are called *super-harmonic functions*. We shall show in later chapters that C_α is the *largest* super-harmonic function. For any x, $C_\alpha(x)$ can therefore be computed as the solution of an LP of the form:

$$\max \phi(x) \text{ subject to (1.3)} \tag{1.4}$$

(the maximization is over the vectors $\phi(v), v \in \mathbf{X}$). This is the dual to (1.2) in the case that there are no constraints. To handle the constrained case we define the Lagrangian

$$J_\alpha^\lambda(x, u) \stackrel{\text{def}}{=} C_\alpha(x, u) + \sum_{k=1}^{K} \lambda_k (D_\alpha^k(x, u) - V_k),$$

where the λ_k are *non-negative* real numbers called Lagrange multipliers. We then show that the value $C_\alpha(x)$ of the constrained problem satisfies:

$$C_\alpha(x) = \inf_{u \in U} \sup_{\lambda} J_\alpha^\lambda(x, u), \tag{1.5}$$

and that the sup and the inf are interchangeable, so that

$$C_\alpha(x) = \sup_\lambda \inf_{u\in U} J^\lambda_\alpha(x,u). \tag{1.6}$$

Since $J^\lambda_\alpha(x,u)$ can be represented as the total expected discounted cost of the policy u corresponding to the immediate cost $(c + \sum_{k=1}^K \lambda_k d^k)$, minus $\sum_{k=1}^K \lambda_k V_k$, we can now obtain $\inf_{u\in U} J^\lambda_\alpha(x,u)$ by applying (1.4), i.e., by maximizing $\phi(x) - \sum_{k=1}^K \lambda_k V_k$ over the vectors $\phi(v), v \in \mathbf{X}$, that satisfy

$$\phi(v) \le (1-\alpha)\Big(c(v,a) + \sum_{k=1}^K \lambda_k d^k(v,a)\Big) + \alpha \sum_y \mathcal{P}_{vay}\phi(y) \quad \forall v \in \mathbf{X}, a \in \mathsf{A}. \tag{1.7}$$

Finally, we add the maximization over non-negative λ to obtain the LP: $\sup_{\lambda,\phi}(\phi(x) - \sum_{k=1}^K \lambda_k V_k)$ subject to (1.7). This is the dual to (1.2).

In what follows we sketch the methods used for extending the ideas illustrated above to infinite MDPs. In particular, we come back to the necessity of the different type of assumptions, mentioned already in Section 1.4.

The sets of policies that we shall be using will turn out to be compact sets. A key issue is that of continuity or of lower semi-continuity of costs with respect to the policies; this will be necessary for the existence of an optimal policy.

In the second part of the book we shall specify two main types of frameworks that will allow us to obtain the continuity or the lower semi-continuity.

A central framework is that of 'uniform Lyapunov functions'. In controlled Markov chains, the uniform Lyapunov function condition is typically stated as follows (see Hordijk, 1977). There should exist some function $\mu : \mathbf{X} \to [1,\infty)$ that is required, among others, to decrease in expectation as long as the initial state is outside some finite set $\mathcal{M}$:

$$1 + \sum_{y\notin\mathcal{M}} \mathcal{P}_{xay}\mu(y) \le \mu(x).$$

This condition (as well as some other alternative conditions) introduced formally in Chapter 6, will be, roughly speaking, necessary and sufficient for the continuity of the costs in the policies (see e.g., Theorem 7.3 and, in particular, the equivalence between properties M1 and M5 there). In many aspects, this condition renders the problem almost equivalent to one with a finite state space.

We shall present another type of framework obtained by assuming some structure of the immediate costs (e.g., boundedness from below or growth conditions). This will be shown to yield lower semi-continuity of the costs in the policies. The importance of this lower semi-continuity can be seen from the Lagrangian approach and from the way we obtain the LP through the Lagrangian approach. An important step there is to change the order of the inf and sup in the Lagrange problem (1.5)–(1.6). To do that, we have

to make use of some saddle-point theorems, in which lower semi-continuity is necessary.

We end this section with a brief discussion on the last part of the book: the sensitivity analysis and approximations. We introduce in Chapter 13 some key theorems on stability of constrained optimization problems. We consider there a sequence of cost functions, corresponding to a sequence of constrained problems, as well as some cost functions of a limit problem. We consider both the problems of approximating a limit problem (i.e., approximating the optimal value and policy) by the sequence of approximating problems, as well as the problem of using a limit problem as an approximation for the other sequence of problems. We assume that the cost criteria for any given policy within some subset of all policies, converge to the cost of the limit problem uniformly in the subset of policies. We further assume that some saddle-point property holds for the Lagrangian corresponding to the limit problem and that a Slater condition holds. We then obtain several statements on the convergence of the values of the optimal problems. Under further lower semi-continuity and convexity-type assumptions, we further obtain statements on the convergence of optimal policies. The key theorems obtained in Chapter 13 are applied in the remaining three chapters to several convergence and approximation issues in constrained MDPs.

1.8 The structure of the book

The structure of the book is as follows. The first part, devoted to the finite MDPs (finite state and action spaces), contains Chapter 2 describing the model and then Chapters 3 and 4 that deal with the discounted and expected average costs, respectively. The theory established there is illustrated in an application to the control of flow and service in a single queue in Chapter 5.

Part II then begins with a more extensive definition and presentation of MDPs with countable state space (Chapters 6–7). We then study the total expected cost, the discounted cost and the expected average cost in Chapters 8–12.

Part III of the book, which contains Chapters 13–16, is devoted to asymptotic analysis and to approximation techniques. We first establish in Chapter 13 some key theorems for approximating the optimal value and optimal policies of **COP** by some sequence of constrained problems. We then apply these theorems in the subsequent chapters to study several applications. We first consider in Chapter 14 the convergence of discounted constrained MDPs in the discount factor and, in particular, the convergence as the discount factor approaches one. In Chapter 15 we study the convergence of finite horizon problems to those of infinite horizon. We finally consider in Chapter 16 several state-truncation techniques, which allow, in particular,

to approximate **COP** with an infinite state space by problems with finite state spaces.

Some of the sections in the monograph are marked with an asterisk. These are more technical and can be skipped at a first reading. Material from these sections is, however, used occasionally in proofs of some theorems in other sections.

PART I

Part One: Finite MDPs

CHAPTER 2

Markov decision processes

2.1 The model

Markov decision processes (MDPs), also known as controlled Markov chains, constitute a basic framework for dynamically controlling systems that evolve in a stochastic way. We focus on discrete time models: we observe the system at times $t = 1, 2, \ldots, n$. n is called the horizon, and may be either finite or infinite. A controller has an influence on both the costs and the evolution of the system, by choosing at each time unit some parameters, called actions. As is often the case in control theory, we assume that the behavior of the system at each time is determined by what is called the 'state' of the system, as well as the control action. The system moves sequentially between different states in a random way; the current state and control action fully determine the probability to move to any given state in the next time unit.

MDPs are thus a generalization of (non-controlled) Markov chains, and many useful properties of Markov chains carry over to controlled Markov chains. A key Markovian property is that conditioned on the state and action at some time t, the past states and the next one are independent.

The models that we study in this monograph are special in that more than one objective cost exists; the controller minimizes one of the objectives subject to constraints on the others. We shall call this class of MDPs Constrained MDPs, or simply CMDPs.

To make the above precise, we define a tuple $\{\mathbf{X}, \mathsf{A}, \mathcal{P}, c, d\}$ where

- $\mathbf{X}$ is a state space that contains a finite number of states. Generic notation for states will be x, y, z.
- A is a finite set of actions. We denote by $\mathsf{A}(x) \subset \mathsf{A}$ those actions that are available at state x. set $\mathcal{K} = \{(x, a) : x \in \mathbf{X}, a \in \mathsf{A}(x)\}$ to be the set of state-action pairs. A generic notation for an action will be a.
- $\mathcal{P}$ are the transition probabilities; thus, $\mathcal{P}_{xay}$ is the probability of moving from state x to y if action a is chosen.
- $c : \mathcal{K} \to \mathbb{R}$ is an immediate cost. This cost will be related to a cost function which we shall minimize.
- $d : \mathcal{K} \to \mathbb{R}^K$ is a K-dimensional vector of immediate costs, related to K constraints (which will be defined later).

A basic part of the description of a control model is to specify the mechanism by which the controller chooses actions at different time epochs. Such a mechanism is often called policy, strategy, profile, or decision rule. A first step is to specify what information is available to the decision maker.

In deterministic models, where the transition probabilities are only zero or one, the controller can fully predict the evolution of the state of the system as a result of applying a sequence of actions, if it knows the initial state. Therefore in several control models in the literature we may restrict ourselves to policies known as 'open loop', i.e., policies that do not require information on the state of the system (except for the initial state).

There are several situations, however, when the state evolution is not fully predictable by the controller, and then it becomes desirable to use policies that use more information on the system:
(i) Whenever the transition probabilities are not only zero or one,
(ii) It will turn out that in CMDPs the performance can often be improved by choosing actions using some randomization mechanisms. Knowing the outcome of the randomizations may be useful for the controller.
(iii) There are control models where some of the parameters of the system (such as transition probabilities) are unknown. The controller can estimate these and improve the control if it has information on the evolution of the system.
(iv) There are models where more than one decision maker controls the system. If there is no coordination between the controllers, then information on the evolution of the system may become crucial for controlling it (even in the case of deterministic transitions).

The above motivates us to consider different classes of 'feedback' (or 'closed loop') policies that may use information on the current state, and of previous actions and states. In order to present a general definition of policies, we define a history at time t to be a sequence of previous states and actions, as well as the current state: $h_t = (x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t)$. Let $\mathbf{H}_t$ be the set of all possible histories of length t.

A policy u is a sequence $u = (u_1, u_2, \ldots)$ (containing n elements); if the history h_t is observed at time t, then the controller chooses an action within a with probability $u_t(a|h_t)$. The class of all policies defined above is denoted by U, and is called the class of *behavioral* policies.

A case when it is desirable to actually design policies that depend on a long history is (iii) and (iv) above, which involve a learning mechanism. Quite often, it will suffice to restrict to simpler policies. We introduce the following classes of policies:

- U_M:= Markov policies; $u \in U_M$ if for any t, u_t is only a function of x_t (and not of the whole history).
- U_S:= stationary policies, which are a subset of U_M; w is stationary if w_t does not depend on t. We shall identify (with some abuse of notation)

a stationary policy with the the probability that it assigns to different actions in different states. Under any stationary policy w, the state process becomes a stationary Markov chain with transition probabilities $P_{xy}(w) = \sum_{a \in \mathsf{A}(x)} \mathcal{P}_{xay} w_x(a)$ (in case of compact actions this will be replaced by $P_{xy}(w) = \int \mathcal{P}_{xay} w_x(da)$). If a stationary probability for the Markov chain exists and is unique, it is denoted by $\pi(w)$.

- U_D:= stationary deterministic policies, which are a subset of U_S; a policy g is stationary deterministic if the action it chooses at state x is a function of x. g is thus identified with a map $g : \mathbf{X} \to \mathsf{A}$.

We shall fix an initial distribution β over the initial state; in other words, the probability that we are at state x at time 1 is given by $\beta(1)$. In particular, if β is concentrated on a single state z, then β can be written as the Dirac function $\beta(x) = \delta_z(x)$.

The initial distribution β and any given policy u determine a unique probability measure P_β^u over the space of trajectories of the states and actions. This defines the stochastic processes X_t and A_t of the states and actions. The construction of the probability space for $u \in U$ is standard, see e.g., Hinderer (1970). We denote by E_β^u the corresponding expectation operator. For the special case where $\beta = \delta_z$ is concentrated on a single state z, we shall use (with some abuse of notation) P_z^u and W_z^u instead of P_β^u and E_z^u, respectively.

2.2 Cost criteria and the constrained problem

We now define the cost criteria which are most frequently used in applications of control of CMDPs. Other cost criteria will be handled in the second part of the book. For any policy u and initial distribution β, the finite horizon cost for a horizon n is defined as

$$C^n(\beta, u) = \sum_{t=1}^{n} E_\beta^u c(X_t, A_t). \tag{2.1}$$

An alternative cost that gives less importance to the far future is the discounted cost. For a fixed discount factor α, $0 < \alpha < 1$, define

$$C_\alpha^n(\beta, u) = (1-\alpha) \sum_{t=1}^{n} \alpha^{t-1} E_\beta^u c(X_t, A_t), \tag{2.2}$$

$$C_\alpha(\beta, u) = \overline{\lim_{n \to \infty}} C_\alpha^n(\beta, u). \tag{2.3}$$

Since there are finitely many states and actions, the $\overline{\lim}$ indeed exists and

$$C_\alpha(\beta, u) = (1-\alpha) \sum_{t=1}^{\infty} \alpha^{t-1} E_\beta^u c(X_t, A_t). \tag{2.4}$$

(This need not be true for more general settings as in the second part of the book.)

Remark 2.1 (*The normalization constant*)
Quite frequently, the discounted cost is defined without the normalizing constant $(1-\alpha)$. The techniques are the same for both cases, and one could retrieve one from the other by multiplying or dividing the immediate cost by this factor. There are several advantages of using this normalization. First, we avoid the situation where, for fixed immediate cost c and d, the total discounted cost becomes very large if α is close to one. Second, with this normalization, the discounted cost will be seen to converge to the expected average cost when stationary policies are used. Finally, we shall see that the LP used to solve the discounted and the expected average costs has the same form when the normalization constant is used.

The expected average cost (with finite and infinite horizons, respectively) is defined as

$$C_{ea}^n(\beta,u)=\frac{\sum_{t=1}^n E_\beta^u c(X_t,A_t)}{n}, \qquad C_{ea}(\beta,u)=\overline{\lim_{n\to\infty}} C_{ea}^n(\beta,u). \tag{2.5}$$

Let $C(\beta,u)$ stand for any of the above costs. Then $C(u):\mathbf{X}\to\mathbb{R}$ will denote the function (or vector) whose x entry is $C(x,u)$. The cost functions related to the immediate costs d are defined similarly; e.g., the finite horizon cost related to d^k, $k=1,\dots,K$, is $D^{n,k}(\beta,u)=\sum_{t=1}^n E_\beta^u d^k(X_t,A_t)$.

For a fixed vector $V=(V_1,\dots,V_K)$ of real numbers, we define the constrained control problem **COP** as:

Find a policy that minimizes $C(\beta,u)$ subject to $D(\beta,u)\le V$.

$C(\beta,u)$ and $D(\beta,u)$ stand for one of the expected costs defined above, i.e., (2.1)–(2.5).

2.3 Some notation

Here, and throughout, we use the notation $q_1\le q_2$ between two vectors $q_1,q_2\in\mathbb{R}^K$ to mean componentwise ordering, i.e., $q_1(j)\le q_2(j)$, $j=1,\dots,K$. Similarly, for any two measures q_1,q_2 defined on the same measurable space $(\Omega,\mathcal{F})$, $q_1\le q_2$ means $q_1(B)\le q_2(B)$ for any $B\in\mathcal{F}$. We use the notation $\langle q_1,q_2\rangle$ between two vectors to denote their scalar product.

The set of policies satisfying the constraints is called feasible. If this set is non-empty, then the constrained problem is said to be feasible (we shall use a similar terminology for linear programs). Let $C(\beta)$ be the value of the above problem, with the obvious notation related to the different costs (2.1)–(2.5) (e.g., $C_{ea}(\beta)$ is the value of **COP** when the expected average costs are used). (If the feasible set of policies is empty, then we set $C(\beta)=\infty$.) If a feasible policy u^* achieves the minimum, i.e., $C(\beta)=C(\beta,u^*)$, then it is called optimal.

Definition 2.1 *(Uniformly optimal policy)*
A policy u is said to be uniformly optimal if it is optimal for all initial states.

Remark 2.2 (*Uniform optimal policies*)
Optimal policies are defined with respect to a given initial state. A policy that is optimal for one initial state might not even be feasible for another, so it is in general not uniformly optimal. In fact, there may be some initial states at which no policy is feasible. This is in contrast to non-constrained MDPs, in which there typically exist policies that are optimal for all initial states (or initial distributions). In the literature on non-constrained MDPs, one thus often defines optimal policies to be policies that are optimal for all initial states.

2.4 The dominance of Markov policies

An important step in solving control problems is to identify subclasses of policies which are simple to handle and to implement, and yet are good candidates to be optimal.

The class of Markov policies turns out to be rich in the following sense. For any policy in U, there exists an equivalent policy in U_M that induces the same marginal probability measure, i.e., the same probability distribution of the pairs (X_t, A_t), $t = 1, 2, \ldots$

All cost criteria that we defined in the previous section have the property that they are functions of the distribution of these pairs. We conclude that the Markov policies are sufficiently rich so that a cost that can be achieved by an arbitrary policy can also be achieved by a Markov policy. We prove this and other more general properties of Markov policies in Theorem 6.1 in the second part of the book.

Definition 2.2 (*Dominating policies*)
A class of policies $\overline{U}$ is said to be a dominating class of policies for **COP** for one of the cost criteria introduced in Section 2.1, and for a given initial distribution β, if for any policy $u \in U$ there exists a policy $\overline{u} \in \overline{U}$ such that

$$C(\beta, \overline{u}) \leq C(\beta, u) \qquad \text{and} \qquad D(\beta, \overline{u}) \leq D(\beta, u). \tag{2.6}$$

In the above definition, C, D, and **COP** stand for any one of the cost criteria previously defined. When (2.6) holds, we say that $\overline{u}$ dominates u.

We conclude

Theorem 2.1 (*Dominance of Markov policies*)
The Markov policies are dominating for any cost criterion which is a function of the marginal distribution of states and actions.

CHAPTER 3

The discounted cost

3.1 Occupation measure and the primal LP

We begin by defining the occupation measure corresponding to a policy. An occupation measure corresponding to a policy u is the total expected discounted time spent in different state–action pairs. It is thus a probability measure over the set of state–action pairs and it has the property that the discounted cost corresponding to that policy can be expressed as the *expectation* of the *immediate cost* with respect to this measure.

More precisely, define for any initial distribution β, any policy u and any pair x, a:

$$f_\alpha(\beta, u; x, a) \stackrel{\text{def}}{=} (1-\alpha)\sum_{t=1}^{\infty} \alpha^{t-1} P_\beta^u(X_t = x, A_x = a), \quad x \in \mathbf{X}, a \in \mathsf{A}(x).$$

$f_\alpha(\beta, u)$ is then the defined to be the set $\{f_\alpha(\beta, u; x, a)\}_{x,a}$. It can be considered as a probability measure, which we call the occupation measure, that assigns probability $f_\alpha^t(\beta, u; x, a)$ to the pair (x, a).

It is easy to check that the discounted costs can be expressed as

$$C_\alpha(\beta, u) = \sum_{x \in \mathbf{X}} \sum_{a \in \mathsf{A}} f_\alpha(\beta, u; x, a) c(x, a), \tag{3.1}$$

for any immediate cost c. (Thus the same representation also holds for the costs $D_\alpha^k(\beta, u)$.)

Define for any class of policies $\overline{U}$

$$\mathbf{L}_{\overline{U}}^\alpha(\beta) = \bigcup_{u \in \overline{U}} f_\alpha(\beta, u), \tag{3.2}$$

and define $\mathbf{L}^\alpha(\beta) := \mathbf{L}_U^\alpha(\beta)$.

Definition 3.1 (*Completeness for the discounted cost*)
A class of policies $\overline{U}$ is said to be complete with respect to the discounted cost problem if $\mathbf{L}^\alpha(\beta) = \mathbf{L}_{\overline{U}}^\alpha(\beta)$.

Theorem 3.1 (*Completeness of stationary policies*)
The set of stationary policies is complete.

Proof. Choose a policy $u \in U$ and let w be a stationary policy satisfying

for all y and a:

$$w_y(a) = \frac{f_\alpha(\beta, u; y, a)}{f_\alpha(\beta, u; y)},\ y \in \mathbf{X}, a \in \mathsf{A}(y) \tag{3.3}$$

whenever the denominator is non-zero. (When it is zero, $w_y(\cdot)$ is chosen arbitrarily). We show that $f_\alpha(\beta, w) = f_\alpha(\beta, u)$. For any $x \in \mathbf{X}$,

$$\begin{aligned} f_\alpha(\beta, u; x) &= \beta(x) + \sum_{t=2}^{\infty} p^u_\beta(t, x) \\ &= \beta(x)(1-\alpha) + \alpha \sum_{t=2}^{\infty} \sum_{y \in \mathbf{X}} \sum_{a \in \mathsf{A}(y)} p^u_\beta(t-1; y, a) \mathcal{P}_{yax} \\ &= \beta(x)(1-\alpha) + \alpha \sum_{y \in \mathbf{X}} \sum_{a \in \mathsf{A}(y)} f_\alpha(\beta, u; y, a) \mathcal{P}_{yax} \\ &= \beta(x)(1-\alpha) + \alpha \sum_{y \in \mathbf{X}} f_\alpha(\beta, u; y) \sum_{a \in \mathsf{A}(y)} \mathcal{P}_{yax} w_y(a) \\ &= \beta(x)(1-\alpha) + \alpha \sum_{y \in \mathbf{X}} f_\alpha(\beta, u; y) P_{yx}(w). \end{aligned} \tag{3.4}$$

This can be written in matrix notation as

$$f_\alpha(\beta, u) = (1-\alpha)\beta + \alpha f_\alpha(\beta, u) P(w).$$

The solution of this equation is

$$f_\alpha(\beta, u) = (1-\alpha)\beta(I - \alpha P(w))^{-1},$$

where I is the identity matrix. Note that all the eigenvalues of $(I - \alpha P(w))$ are non-zero, and it is therefore invertible. (This follows, for example, from Gersgorin's Theorem, see e.g., Horn and Johnson, 1985, p. 344).

Since the above holds in particular for $u = w$, we conclude that for all $x \in \mathbf{X}$, $f_\alpha(\beta, w; x) = f_\alpha(\beta, u; x)$. This implies by the definition of w that $f_\alpha(\beta, w) = f_\alpha(\beta, u)$, so that the set of stationary policies is complete. □

Define $\mathbf{Q}^\alpha(\beta)$ to be the set of vectors $\rho \in \mathbb{R}^{|\mathcal{K}|}$ satisfying

$$\begin{cases} \displaystyle\sum_{y \in \mathbf{X}} \sum_{a \in \mathcal{A}(y)} \rho(y, a)(\delta_x(y) - \alpha \mathcal{P}_{yax}) = (1-\alpha)\beta(x), \quad \forall x \in \mathbf{X} \\ \\ \rho(y, a) \geq 0, \quad \forall y, a \end{cases} \tag{3.5}$$

By summing the first constraint over x we note that $\sum_{y,a} \rho(y, a) = 1$, so that ρ satisfying the above constraints are probability measures.

Theorem 3.2 (*Properties of occupation measures*)
The set of stationary policies is complete. Moreover, $\mathbf{L}^\alpha_{U_S}(\beta)$ is closed polytope, and satisfies

$$\mathbf{L}^\alpha(\beta) = \mathbf{L}^\alpha_{U_S}(\beta) = \overline{co}\mathbf{L}^\alpha_{U_D}(\beta) = \mathbf{Q}^\alpha(\beta).$$

Proof. The first equality follows from Theorem 3.1. That $\mathbf{L}^\alpha(\beta) \subset \mathbf{Q}^\alpha(x)$ follows from (3.4). The converse relation follows from an argument similar to the one used in the proof of Theorem 3.1; for any $\rho \in \mathbf{Q}^\alpha(\beta)$, let w be the stationary policy that satisfies for all y, a:

$$w_y(a) = \frac{\rho(y,a)}{\sum_{a \in \mathsf{A}(y)} \rho(y,a)}, \ y \in \mathbf{X}, a \in \mathsf{A}(y) \tag{3.6}$$

whenever the denominator is non-zero. Define (with some abuse of notation) $\rho(y) := \sum_{a \in \mathsf{A}(y)} \rho(y,a)$. Then

$$\begin{aligned}
\rho(x) &= \beta(x)(1-\alpha) + \alpha \sum_{y \in \mathbf{X}} \sum_{a \in \mathcal{A}(y)} \rho(y,a) \mathcal{P}_{yax} \\
&= \beta(x)(1-\alpha) + \alpha \sum_{y \in \mathbf{X}} \rho(y) \sum_{a \in \mathsf{A}(y)} \frac{\rho(y,a)}{\rho(y)} \mathcal{P}_{yax} \\
&= \beta(x)(1-\alpha) + \alpha \sum_{y \in \mathbf{X}} \rho(y) \sum_{a \in \mathsf{A}(y)} w_y(a) \mathcal{P}_{yax} \\
&= \beta(x)(1-\alpha) + \alpha \sum_{y \in \mathbf{X}} \rho(y) P_{yx}(w).
\end{aligned}$$

We conclude that ρ equals $(1-\alpha)\beta(I - \alpha P(w))^{-1}$, and hence to $f_\alpha(\beta, w)$. This now implies, through the definition of w, that $\rho = f_\alpha(\beta, w)$. We conclude that $\mathbf{L}^\alpha(\beta) = \mathbf{Q}^\alpha(\beta)$.

$\mathbf{Q}^\alpha(\beta)$, and hence $\mathbf{L}^\alpha_{U_S}$ is a closed convex polytope; since $\mathbf{L}^\alpha_{U_D}(\beta) \subset \mathbf{L}^\alpha_{U_S}(\beta)$, this implies that $\overline{co}\mathbf{L}^\alpha_{U_D}(\beta) \subset \mathbf{L}^\alpha_{U_S}(\beta)$. The converse is established by showing that if f is an extreme point of $\mathbf{L}^\alpha_{U_D}$, then there exists a deterministic policy w such that $f_\alpha(\beta, w) = f$. The proof is quite technical, and we thus omit it; the interested reader may find the proof of a more general statement in the second part of the book (Corollary 10.1). □

It now follows from the definition of $\mathbf{Q}^\alpha(\beta)$, from Theorem 3.2 and from the representation of the cost in (3.1) that the value of **COP** can be obtained using the following linear programming, which we call the primal program:

$\mathbf{LP}^\alpha_{\mathbf{1}}(\beta)$: Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle \rho, c \rangle$ subject to:

$$\mathcal{D}^k(\rho) := \langle \rho, d^k \rangle \leq V_k, k = 1, \ldots, K, \qquad \rho \in \mathbf{Q}^\alpha(\beta). \tag{3.7}$$

We thus obtain the following theorem.

Theorem 3.3 (*Equivalence between COP and the LP*)
(*i*) $\mathcal{C}^* = C_\alpha(\beta)$.
(*ii*) *For any* $u \in U$, $\rho(u) := f_\alpha(\beta, u) \in \mathbf{Q}^\alpha(\beta)$, $C_\alpha(\beta, u) = \mathcal{C}(\rho(u))$

and $D_\alpha(\beta,u) = \mathcal{D}(\rho(u))$*; conversely, for any* $\rho \in \mathbf{Q}^\alpha(\beta)$*, the stationary policy* $w = w(\rho)$*, defined in* (3.6)*, satisfies* $C_\alpha(\beta, w(\rho)) \le \mathcal{C}(\rho)$ *and* $D_\alpha(\beta, w(\rho)) \le \mathcal{D}(\rho)$.
(*iii*) $\mathbf{LP}^\alpha_{\mathbf{1}}(\beta)$ *is feasible if and only if* **COP** *is. Assume that* **COP** *is feasible. Then there exists an optimal solution* ρ^* *for* $\mathbf{LP}^\alpha_{\mathbf{1}}(\beta)$*, and the stationary policy* $w(\rho^*)$ *is optimal for* **COP**.

3.2 Dynamic programming and dual LP: the unconstrained case

We describe in this section the dynamic programming method for solving non-constrained problems, i.e., $K = 0$. The following holds (see e.g., Puterman, 1994, and references therein):

Theorem 3.4 (*Dynamic programming*)
(*i*) *The value* $C_\alpha(x) \stackrel{\text{def}}{=} \inf_{u\in U} C_\alpha(x,u)$*, as a function of* x*, is the unique solution of the following dynamic programming equation:*

$$\phi(x) = \min_{a\in \mathsf{A}} \Big((1-\alpha)c(x,a) + \alpha \sum_{y\in\mathbf{X}} \mathcal{P}_{xay}\phi(y) \Big) \quad \forall x \in \mathbf{X}. \tag{3.8}$$

(*ii*) *For any state* x*, let* $\mathcal{A}(x)$ *be the set of actions that achieve the minimum of* $[(1-\alpha)c(x,a) + \alpha\sum_{y\in\mathbf{X}} \mathcal{P}_{xay} C_\alpha(y)]$*. A stationary policy* g *is uniformly optimal if and only if it chooses actions within* $\mathcal{A}(x), x \in \mathbf{X}$ *w.p.1* (*i.e., for which* $g(\mathcal{A}(x)) = 1$ *for all* $x \in \mathbf{X}$).

Proof. (i) We first show that C_α satisfies (3.8).

$$\begin{aligned}
C_\alpha(x) &= \inf_{u\in U_M} C_\alpha(x,u) \\
&= \inf_{u\in U_M} \Big((1-\alpha)c(x,u_1) + \alpha \sum_{t=2}^{\infty} E^u_x \alpha^{t-1} c(X_t, A_t) \Big) \\
&= \inf_{u\in U_M} \Big((1-\alpha)c(x,u_1) + \alpha \sum_{y\in\mathbf{X}} \mathcal{P}_{xu_1y} C_\alpha(y) \Big) \\
&= \inf_{a\in \mathsf{A}(x)} \Big((1-\alpha)c(x,a) + \alpha \sum_{y\in\mathbf{X}} \mathcal{P}_{xay} C_\alpha(y) \Big).
\end{aligned}$$

Thus C_α satisfies (3.8).

Consider any solution ϕ of (3.8) and let w be a stationary policy that chooses at state x an action that achieves the minimum of

$$(1-\alpha)c(x,a) + \alpha \sum_{y\in\mathbf{X}} \mathcal{P}_{xay}\phi(y).$$

We iterate (3.8) and obtain:

$$\phi(x) = (1-\alpha)c(x,w) + \alpha \sum_{y\in\mathbf{X}} \mathcal{P}_{xwy}\phi(y) = (1-\alpha)c(x,w) + \alpha E^w_x \phi(X_2)$$

$$
\begin{aligned}
&= (1-\alpha)\left[c(x,w) + \alpha E_x^w c(X_2, A_2)\right] + \alpha^2 E_x^u (E_{X_2}^w \phi(X_3)) \\
&= (1-\alpha)\left[c(x,w) + \alpha E_x^w c(X_2, A_2)\right] + \alpha^2 E_x^w \phi(X_3) \\
&= \ \ldots \ = (1-\alpha)\sum_{t=1}^{n} E_x^w \alpha^{t-1} c(X_t, A_t) + \alpha^n E_x^w \phi(X_{n+1}) \\
&= C_\alpha^n(x,w) + \alpha^n E_x^w \phi(X_{n+1}). \qquad (3.9)
\end{aligned}
$$

Taking the limit as $n \to \infty$, we see that $\phi(x) = C_\alpha(x,w)$ for every x.

Let g be a policy as in (ii). By repeating the computation in (3.9) with $\phi = C_\alpha$ and with $w = g$, we see that $C_\alpha(x) = C_\alpha(x,g)$ for all $x \in \mathbf{X}$, so g is uniformly optimal. To obtain the converse, assume that $u \in U_S$ does not satisfy the condition in the theorem, i.e., for some $x \in \mathbf{X}$ and some $\delta > 0$,

$$
\begin{aligned}
&(1-\alpha)c(x,u) + \alpha \sum_{y \in \mathbf{X}} \mathcal{P}_{xuy} C_\alpha(y) \\
&\quad = \min_{a \in \mathsf{A}(x)} \Big((1-\alpha)c(x,a) + \alpha \sum_{x \in \mathbf{X}} \mathcal{P}_{xay} C_\alpha(y) \Big) + \delta \ .
\end{aligned}
$$

Together with (3.8), this implies

$$
\begin{aligned}
C_\alpha(x) + \delta &= (1-\alpha)c(x,u) + \alpha \sum_{y \in \mathbf{X}} \mathcal{P}_{xuy} C_\alpha(y) \\
&= (1-\alpha)c(x,u) + \alpha E_x^u C_\alpha(X_2) \\
&\le (1-\alpha)\big(c(x,u) + \alpha E_x^u c(X_2, A_2)\big) + \alpha^2 E_x^u C_\alpha(X_3) \\
&\le \cdots \le C_\alpha^n(x,u) + \alpha^n E_x^u C_\alpha(X_{n+1}).
\end{aligned}
$$

Taking the limit as $n \to \infty$, we conclude that $C_\alpha(x,u) \ge C_\alpha(x) + \delta$. Hence u is not uniformly optimal. □

We conclude from Theorem 3.4 that C_α satisfies the inequalities

$$
\phi(x) \le (1-\alpha)c(x,a) + \alpha \sum_y \mathcal{P}_{xay} \phi(y) \quad \forall x \in \mathbf{X}\,, a \in \mathsf{A}. \qquad (3.10)
$$

Definition 3.2 *(Superharmonic functions)*
The set of functions satisfying (3.10) are called super-harmonic functions.

Theorem 3.5 (*The value and super-harmonic functions*)
The value C_α is the largest super-harmonic function.

Proof. (i) From Theorem 3.4 it follows that C_α is a super-harmonic function. Choose a super-harmonic function ϕ and let g be an optimal stationary policy. Then

$$
\phi(x) \ \le \ (1-\alpha)c(x,g) + \alpha \sum_{y \in \mathbf{X}} \mathcal{P}_{xgy} \phi(y) = (1-\alpha)c(x,g) + \alpha E_x^g \phi(X_2)
$$

$$\begin{aligned} &\leq (1-\alpha)c(x,g) + \alpha E_x^g\big((1-\alpha)c(X_2,A_2) + \alpha E_{X_2}^g \phi(X_3)\big) \\ &= (1-\alpha)\left(c(x,g) + \alpha E_x^g c(X_2,A_2)\right) + \alpha^2 E_x^g \phi(X_3) \\ &\leq \ \ldots \ \leq (1-\alpha)\sum_{t=1}^{n} \alpha^{t-1} E_x^g c(X_t,A_t) + \alpha^n E_x^g \phi(X_{n+1}). \end{aligned}$$

The proof is established by taking the limit as $n \to \infty$. (A proof of a more general statement can be found in Lemma 3.6 in Feinberg and Sonin, 1983.) □

The above theorem implies that for any x, $C_\alpha(x)$ can be computed as the solution of the following LP with the decision variables $\phi(x), x \in \mathbf{X}$.

$$\begin{aligned} \mathbf{DP}^\alpha(\beta): \qquad & \text{Find } \Theta^* := \sup_\phi \langle \beta, \phi \rangle \text{ subject to} \qquad (3.11) \\ \phi(x) \ &\leq \ (1-\alpha)c(x,a) + \alpha \sum_{y \in \mathbf{X}} \mathcal{P}_{xay}\phi(y), \ x \in \mathbf{X}, a \in \mathsf{A}(x). \end{aligned}$$

This LP is dual to $\mathbf{LP}_{\mathbf{1}}^\alpha(\beta)$ in the case of no constraints.

3.3 Constrained control: Lagrangian approach

We now go back to our constrained control problem. We use a standard Lagrangian approach for convex programming to show that
(i) **COP** is equivalent to solving a non-constrained sup-inf problem;
(ii) the sup and inf can be interchanged under suitable conditions; the inf in the inf-sup problem is in fact achieved by some policy which is optimal for **COP**.
(iii) Under the Slater conditions, the sup is also obtained as max, that is, 'optimal' policies and Lagrange multipliers exist for the sup-inf problem, and they satisfy the Kuhn-Tucker conditions.

The main result is presented in the following theorem.

Theorem 3.6 (*The Lagrangian*)
(*i*) *The value function satisfies*

$$C_\alpha(\beta) = \inf_{u \in U} \sup_{\lambda \geq 0} J_\alpha^\lambda(\beta,u) = \inf_{u \in U_M} \sup_{\lambda \geq 0} J_\alpha^\lambda(\beta,u) \qquad (3.12)$$

where

$$\begin{aligned} J_\alpha^\lambda(\beta,u) \ &:= \ C_\alpha(\beta,u) + \langle \lambda, D_\alpha(\beta,u) - V \rangle \\ &= \ \sum_{t=1}^{\infty} \alpha^{t-1} E_\beta^u j^\lambda(X_t,A_t) - \langle \lambda, V \rangle \\ j^\lambda(x,a) \ &:= \ c(x,a) + \langle \lambda, d(x,a) \rangle. \qquad (3.13) \end{aligned}$$

(*ii*) *A policy* u^* *is optimal for* **COP** *if and only if* $C_\alpha(\beta) = \sup_{\lambda \geq 0} J_\alpha^\lambda(\beta,u^*)$.

(*iii*) *The value satisfies*

$$C_\alpha(\beta) = \sup_{\lambda\geq 0}\min_{u\in U} J_\alpha^\lambda(\beta,u) = \sup_{\lambda\geq 0}\min_{u\in U_D} J_\alpha^\lambda(\beta,u). \tag{3.14}$$

Moreover, there exists some $u^* \in U_S$ *such that*

$$C_\alpha(\beta) = \inf_{u\in U}\sup_{\lambda\geq 0} J_\alpha^\lambda(\beta,u) = \inf_{u\in U_S}\sup_{\lambda\geq 0} J_\alpha^\lambda(\beta,u) = \sup_{\lambda\geq 0} J_\alpha^\lambda(\beta,u^*), \tag{3.15}$$

and u^* *is optimal for* **COP**.

Proof. (i) If for some u **COP** is not feasible, then $\sup_{\lambda\geq 0} J_\alpha^\lambda(\beta,u) = \infty$. Indeed, if the jth constraint is violated, i.e., $D_\alpha^j(\beta,u) > V$, then the above supremum is obtained by choosing λ_j very large: choosing it to tend to infinity (the other ones can be chosen to be 0, for example). If **COP** is feasible for that u, then the sup is obtained by choosing $\lambda = 0$, and then $J_\alpha(\beta,u) = C_\alpha(\beta,u)$. Hence,

$$\inf_{u\in U}\sup_{\lambda\geq 0} J_\alpha^\lambda(\beta,u) = \inf_{u: D_\alpha(\beta,u)\leq V} C_\alpha^\lambda(\beta,u)$$

from which the first equality in (3.12) follows. The fact that we may restrict to Markov policies follows from Theorem 2.1. The above argument also implies (ii).

To prove (iii), we make use of the results of Section 3.1, although (iii) can be obtained independently (which is done in Section 9.4).

It follows from (3.1) that

$$C_\alpha(\beta,u) + \sum_{k=1}^K \lambda_k D_\alpha^k(\beta,u) = \sum_{y,a} f_\alpha(\beta,u;y,a) j^\lambda(y,a).$$

Thus,

$$\min_u \sup_{\lambda\geq 0} J_\alpha(\beta,u) = \min_{f\in \mathbf{L}^\alpha(\beta)} \sup_{\lambda\geq 0} \sum_{y,a} f(y,a) j^\lambda(y,a) - \sum_{k=1}^K \lambda_k V_k.$$

Since by Theorem 3.2, $\mathbf{L}^\alpha(\beta)$ is convex and compact and since the set $\lambda \geq 0$ is convex, it follows from a standard minmax theorem (Aubin, 1993, p. 126) that the min and the sup can be interchanged (this minmax theorem is stated precisely in Lemma 9.2). The restriction to U_D in (iii) follows from the fact that for any fixed λ, the minimization of $J_\alpha^\lambda(\beta,u)$ is an MDP with no constraints; hence Theorem 3.4 can be applied to show that there exists an optimal stationary deterministic policy. □

3.4 The dual LP

We now introduce the LP which is the dual to $\mathbf{LP}_{\mathbf{1}}^\alpha(\beta)$. Its decision variables are $\phi(x), x \in \mathbf{X}$ as well as the K-dimensional non-negative vector $\lambda \in \mathbb{R}_+^K$.

We have:

$$\begin{aligned} \mathbf{DP}_{\mathbf{1}}^{\alpha}(\beta): \qquad & \text{Find } \Theta^* := \max_{\phi,\lambda} \langle \beta, \phi \rangle - \sum_{k=1}^{K} \lambda_k V_k \text{ subject to} \\ \phi(x) \quad \leq \quad & (1-\alpha)(c(x,a) + \langle \lambda, d(x,a) \rangle) + \alpha \sum_{y \in \mathbf{X}_\alpha} \mathcal{P}_{xay} \phi(y), \\ & x \in \mathbf{X}_\alpha, a \in \mathsf{A}(x). \end{aligned}$$

Theorem 3.7 (**COP** *and the dual LP*)
The value of **COP** *is given by the solution of* $\mathbf{DP}_{\mathbf{1}}^{\alpha}(\beta)$.

Proof. For any fixed λ, we can use the LP (3.11) to compute $\inf_{u \in U} J_\alpha^\lambda(\beta, u)$, where $c(x,a)$ is replaced by $j^\lambda(x,a)$. Due to (3.12) in Theorem 3.6, $C_\alpha(\beta)$ is obtained by adding to this LP a further maximization over λ, which yields $\mathbf{DP}_{\mathbf{1}}^{\alpha}(\beta)$. □

Although we could have derived the dual program directly from the primal program, the above proof, using Lagrange arguments and dynamic programming, allows us to obtain further insight into this dual program.

3.5 Number of randomizations

We show in this section that when **COP** is feasible then there exists an optimal stationary policy w that requires at most K randomizations. A similar proof (for the expected average cost) was given by Koole (1988) and by Ross (1989).

Definition 3.3 (*Number of randomizations*)
We say that under a stationary policy w there are $m(y,w)$ randomizations in state y if there are exactly $m+1$ actions in $\mathsf{A}(y)$ for which $w_y(a) > 0$. We say that the total number of randomizations under w is $n(w)$ if

$$\sum_{y \in \mathbf{X}} m(y,w) = n(w).$$

In particular, if the total number of randomizations under $w \in U_S$ is $n(w)$, then there are no more than $n(w)$ states in which randomization is used.

Theorem 3.8 (*Bound on the number of randomizations*)
If **COP** *is feasible then there exists an optimal stationary policy w such that the total number $n(w)$ of randomizations that it uses is at most K (where K is the number of constraints).*

Proof. Consider any fixed initial distribution β for **COP**. Since $\mathbf{LP}_{\mathbf{1}}^{\alpha}(\beta)$ has $|\mathbf{X}| + K$ constraints, it follows that there exists an optimal solution ρ^* for this LP that has at most $|\mathbf{X}| + K$ non-zero elements. Define the stationary policy $w(\rho^*)$ as in (3.6).

Assume first that

$$\text{for each state } y \text{ there exists some } a \text{ such that } \rho^*(y,a) > 0. \tag{3.16}$$

Define

- $\mathbf{X}' \stackrel{\text{def}}{=} \{y : \text{the number of actions } a \in \mathsf{A}(y) \text{ such that } \rho^*(y,a) > 0 \text{ is greater than } 1\}$,
- $\mathcal{K}' \stackrel{\text{def}}{=} \{(y,a), y \in \mathbf{X}' \text{ such that } \rho^*(y,a) > 0\}$.

Randomizations clearly occur only in states in $\mathbf{X}'$, and the number of randomizations in a state $y \in \mathbf{X}'$ is $m(y,w)$ if and only if the number of actions for which $\rho^*(y,a) > 0$ is $m(y,w)+1$.

By (3.16), number of state–action pairs (y,a) that are not in $\mathcal{K}'$ and for which $\rho^*(y,a) > 0$ is $|\mathbf{X}| - |\mathbf{X}'|$. This implies that the number of elements in $\mathcal{K}'$ is upper-bounded by $|\mathbf{X}'| + K$. Hence, by Definition 3.3 the number of randomizations is indeed upper-bounded by K.

It remains to consider the case where (3.16) does not hold. Let $\mathcal{X}$ be the set of states for which $\sum_{a \in \mathsf{A}(y)} \rho^*(y,a) = 0$. Define the following new MDP, which is obtained from the initial MDP by eliminating some states and some actions:

- State space: $\overline{\mathbf{X}} := \mathbf{X} \setminus \mathcal{X}$;
- Action space: $\overline{\mathsf{A}}(y) := \{a : \mathcal{P}_{yax} = 0, \forall x \in \mathcal{X}\}$;

The transition probabilities and costs are unchanged, i.e.,

$$\overline{\mathcal{P}}_{xay} = \mathcal{P}_{xay}, \qquad \overline{c}(x,a) = c(x,a), \qquad \overline{d}(x,a) = d(x,a)$$

$\forall x, y \in \overline{\mathbf{X}}, a \in \mathsf{A}(x)$. Clearly, in any state $y \in \overline{\mathbf{X}}$, the original policy w did not use any action which is not in $\overline{\mathsf{A}}(y)$, otherwise $f_\alpha(\beta, w; x) > 0$ for some $x \in \mathcal{X}$, which would imply that $\sum_a \rho^*(x,a') > 0$, and this contradicts the definition of $\mathcal{X}$. Hence we may use again the same policy w in the new MDP, as for each state $y \in \overline{\mathbf{X}}$, the support of $w_y(a)$ are within $\overline{\mathsf{A}}$. It also follows that the total expected discounted costs achieved under w in the new MDP is the same as in the previous one.

The new MDP can be seen as being obtained from the original one by adding the constraints of not using some actions. Therefore, if any other policy in the new MDP performed better than w, then it would also perform better than w in the original MDP, which contradicts the fact that w is optimal for the original MDP. We conclude that w is optimal in the new MDP. By Theorem 3.3, the corresponding ρ is optimal for the equivalent LP.

Finally, assumption (3.16) holds for the new LP, and therefore w does not use more than K randomizations in the new MDP. This clearly holds also for the original MDP, which concludes the proof. □

CHAPTER 4

The expected average cost

The next cost criterion that we study is the expected average cost. We assume throughout this chapter that the MDP is unichain, which is defined in the following.

Definition 4.1 (*Unichain MDP*)
An MDP is said to be unichain if under any $w \in U_D$, the corresponding Markov chain contains a single (aperiodic) ergodic class.

4.1 Occupation measure and the primal LP

As we did in the discounted case, we define occupation measures. Here too, they will allow us to obtain an LP for solving **COP**.

For any given initial distribution β and policy u, and any state–action pair x, a, define

$$f^t_{ea}(\beta, u; x, a) = \frac{1}{t} \sum_{s=1}^{t} P^u_\beta(X_s = x, S_s = a), \qquad a \in \mathsf{A}(x). \tag{4.1}$$

The finite-horizon state–action frequencies (or occupation measure) $f^t_{ea}(\beta, u)$ are the sets $\{f^t_{ea}(\beta, u; x, a)\}_{x,a}$. $f^t_{ea}(\beta, u)$ can be considered as a probability measure that assigns probability $f^t_{ea}(\beta, u; x, a)$ to the pair (x, a).

With some abuse of notation, we define $f^t_{ea}(\beta, u; x) = f^t_{ea}(\beta, u; x, \mathsf{A}(x))$. The subscript ea stands for *expected average*.

We denote by $F_{ea}(\beta, u)$ the non-empty compact set obtained as all the accumulation points of $\{f^t_{ea}(\beta, u)\}_t$. $F_{ea}(\beta, u)$ are called the sets of occupation measures for the expected average cost. Thus, unlike the case in the discounted framework, to a given initial distribution and a given policy there may correspond an infinite set of occupation measures.

The motivation for introducing these sets is that, again, the cost is related to the occupation measure. Indeed, we have the following:

Observation 4.1 *(Representation of the cost)*
For any β, $u \in U$ and $f \in F_{ea}(\beta, u)$,

$$C_{ea}(\beta, u) \geq \langle f, c \rangle := \sum_{y \in \mathbf{X}} \sum_{a \in \mathsf{A}(x)} f(y, a) \tag{4.2}$$

with equality holding for some $f \in F_{ea}(\beta, u)$. If $u \in U_S$ then (4.2) holds with equality.

Below, we shall show that one may restrict to stationary policies without loss of optimality. For $u \in U_S$, it turns out that $F_{ea}(\beta, u)$ are all singletons. Thus, the cost has again the same linear representation as we had in the discounted case, as long as we use U_S, or other 'nice' classes of policies. We shall make this precise below.

Define,

$$\mathcal{L}_{\overline{U}}(\beta) = \bigcup_{u \in \overline{U}} \{F_{ea}(\beta, u)\} \text{ for any set of policies } \overline{U},$$

$$\mathbf{Q}_{ea}(\beta) = \left\{ \begin{array}{l} \rho(y,a), y \in \mathbf{X}, a \in \mathsf{A}(y) : \\ \\ \quad \sum_{y \in \mathbf{X}} \sum_{a \in \mathsf{A}(x)} \rho(y, da)(\delta_x(y) - \mathcal{P}_{yax}) = 0, \quad x \in \mathbf{X} \\ \\ \quad \sum_{y \in \mathbf{X}} \sum_{a \in \mathsf{A}(y)} \rho(y,a) = 1, \qquad \rho(y,a) \geq 0 \ \ \forall y, a \end{array} \right\} \tag{4.3}$$

where δ_x is the Dirac probability measure concentrated on x. We set $\mathcal{L}(\beta) = \mathcal{L}_U(\beta)$. $\mathcal{L}_{\overline{U}}(\beta)$ is called the set of expected occupation measures achievable by $\overline{U}$.

Definition 4.2 (*Completeness for the expected average cost*)
A class of policies $\overline{U}$ is called complete for the expected average cost criterion (for a given initial distribution β) if

$$\mathcal{L}(\beta) = \mathcal{L}_{\overline{U}}(\beta) \text{ and } \forall u \in \overline{U}, \ F_{ea}(\beta, u) \text{ is a singleton .}$$

Thus a complete class of policies $\overline{U}$ has the property that the achievable expected occupation measures under $\overline{U}$ are the same as under all policies.

The following lemma, which follows immediately from (4.2), motivates the definition of complete classes of policies.

Lemma 4.1 (*Sufficiency of complete classes of policies*)
Any complete class of policies is dominant.

Theorem 4.1 (*Completeness of stationary policies*)
The stationary policies are complete and hence dominant.

Proof. Choose a policy $u \in U$. Let t_n be some increasing sequence of times along which $f^t_{ea}(\beta, u)$ converges to some limit $f \in F_{ea}(\beta, u)$. Define the stationary policy w as follows:

$$w_y(a) = \frac{f(y,a)}{\sum_{a' \in \mathsf{A}(y)} f(y,a')}, \ a \in \mathsf{A}(y)$$

whenever the denominator is non-zero. When it is zero, $w_y(\cdot)$ is chosen to

be an arbitrary probability measure over $\mathsf{A}(y)$. It follows from the unichain assumption that the Markov chain with transition probabilities $P(w)$ has a unique invariant probability measure $\pi(w)$, independent of the initial distribution β, that satisfies

$$\pi_y(w) = \lim_{t\to\infty} f^t_{ea}(\beta, w; y),$$

and hence, $F_{ea}(\beta, w) = \{f^w\}$ is a singleton and it satisfies

$$f^w(y, a) = w_y(a)\pi_w(y), \ \forall y \in \mathbf{X}, a \in \mathsf{A}(y). \tag{4.4}$$

We show that $f^w = f$. Since clearly for any $x \in \mathbf{X}$,

$$P^u_\beta(X_{t+1} = x) = \sum_{y\in\mathbf{X}} \sum_{a\in\mathsf{A}(y)} P^u_\beta(X_t = y, A_t = a)\mathcal{P}_{yax},$$

we have

$$\begin{aligned} f^t_{ea}(\beta, u; x) - \frac{\beta(x)}{t} &= \sum_{y\in\mathbf{X}} \sum_{a\in\mathsf{A}(x)} f^t_{ea}(\beta, u; y, a)\mathcal{P}_{yax} \\ &\quad - \frac{\sum_{y\in\mathbf{X}} \sum_{a\in\mathsf{A}(x)} P^u_\beta(X_t = y, A_t = a)\mathcal{P}_{yax}}{t} \end{aligned} \tag{4.5}$$

Hence

$$\begin{aligned} f(x) &= \lim_{n\to\infty} f^{t_n}_{ea}(\beta, u; x) = \lim_{n\to\infty} \sum_{y\in\mathbf{X}} \sum_{a\in\mathsf{A}(x)} f^{t_n}_{ea}(\beta, u; y, a)\mathcal{P}_{yax} \\ &= \sum_{y\in\mathbf{X}} \sum_{a\in\mathsf{A}(x)} f(y, a)\mathcal{P}_{yax}. \end{aligned} \tag{4.6}$$

By definition of w and of $P_{xy}(w)$,

$$\begin{aligned} \sum_{y\in\mathbf{X}} \sum_{a\in\mathsf{A}(x)} f(y, a)\mathcal{P}_{yax} &= \sum_y f(y) \sum_{a\in\mathsf{A}(y)} w_y(a)\mathcal{P}_{yax} \\ &= \sum_y f(y)P_{yx}(w). \end{aligned} \tag{4.7}$$

This, together with (4.6), leads to

$$f(x) = \sum_y f(y)P_{yx}(w), \tag{4.8}$$

where $f(y) := \sum_{a\in\mathsf{A}(t)} f(y, a)$. The steady-state probability of the Markov chain obtained when the stationary policy w is applied, $\pi(w)$, is known to be the unique probability measure over $\mathbf{X}$ satisfying (4.8). This, together with the definition of w, implies that $\{f\} = \{f^w\}$, which establishes the proof. □

Theorem 4.2 (*Characterizing the achievable occupation measures*) *$\mathcal{L}_{U_S}(\beta)$ is a closed convex polytope and*

$$\mathcal{L}(\beta) = \mathcal{L}_{U_S}(\beta) = \overline{co}\mathcal{L}_{U_D}(\beta) = \mathbf{Q}_{ea}(\beta).$$

Proof. The completeness of U_S was established in Theorem 4.1, and thus $\mathcal{L}(\beta) = \mathcal{L}_{U_S}(\beta)$. (4.6) shows that $\mathcal{L}(\beta) \subset \mathbf{Q}_{ea}(\beta)$. We next show the converse. For any $\rho \in \mathbf{Q}_{ea}(\beta)$, define again the stationary policy w by

$$w_y(a) = \frac{\rho(y,a)}{\rho(y)},\ a \in \mathsf{A}(y)$$

where $\rho(y) := \sum_{a' \in \mathsf{A}(y)} \rho(y,a')$ whenever the denominator is non-zero. When it is zero, $w_y(\cdot)$ is chosen to be an arbitrary probability measure over $\mathsf{A}(y)$. It follows from the definition of $\mathbf{Q}_{ea}(\beta)$ and of w that for all $x \in \mathbf{X}$,

$$\rho(x) = \sum_{y \in \mathbf{X}} \rho(y) \sum_{a \in \mathsf{A}(y)} w_y(a) \mathcal{P}_{yax} = \sum_{y \in \mathbf{X}} \rho(y) P_{yx}(w).$$

Since $\pi_y(w) = f_{ea}(\beta, w; y),\ y \in \mathbf{X}$ is the unique solution to $\pi = \pi P(w)$ that satisfies $\pi(\mathbf{X}) = 1, \pi \geq 0$; it follows that $\rho(x) = f_{ea}(\beta, u; x)$ for all $x \in \mathbf{X}$, and by the definition of w, $\rho = f_{ea}(\beta, u)$. This establishes $\mathcal{L}_{U_S}(\beta) = \mathbf{Q}_{ea}(\beta)$.

Since $\mathcal{L}_{U_D}(\beta) \subset \mathcal{L}_{U_S}(\beta)$ and since $\mathcal{L}_{U_S}(\beta)$ is a closed polytope, clearly $\overline{co}\mathcal{L}_{U_D}(\beta) \subset \mathcal{L}_{U_S}(\beta)$. To establish the converse, assume that f is an extreme point of $\mathcal{L}_{U_S}(\beta)$. Let w be a stationary policy such that

$$w_y(a) = \frac{f(y,a)}{\sum_{a' \in \mathsf{A}(y)} f(y,a')},\ a \in \mathsf{A}(y)$$

whenever the denominator is non-zero; when it is zero we set $w_y = \delta_a(\cdot)$ where a is some arbitrary action in $\mathsf{A}(y)$. We shall show that $w \in U_D$.

To establish the last point, we first recall a key property of Markov chains. Consider a Markov chain over with finite state space $\mathbf{X}$ containing a single recurrent class, and let y be a recurrent state. Let $N(x)$ be the number of visits to a state x between two consecutive visits to state y, and let T be the time between two consecutive visits of state y. Then

$$f(x) := \lim_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} P(X_s = x) = \frac{EN(x)}{ET}.$$

In particular, $f(y) = (E[T])^{-1}$. For our controlled Markov chains, this implies that for any stationary policy u, state x and action a,

$$f^u(x,a) = \frac{EN(x,a)}{ET}, \tag{4.9}$$

where $N(x,a)$ is the number of times that the state action pair (x,a) are

visited between two consecutive visits to state y, and f^u is the single limit of f^t_{ea} as $t \to \infty$.

Going back to our problem, assume that $w \notin U_D$. We have by the proof of Theorem 4.1 that $f_{ea}(\beta, w) = f$. Now, there exists some state y such that $f(y) > 0$, and in which w uses two actions, a and b, with positive probabilities: $w_y(a) > 0$, $w_y(b) > 0$. Define the following policies in U_S:

$$v^1_x(a') = \begin{cases} w_x(a) + w_x(b) & \text{if } x = y, a' = a, \\ 0 & \text{if } x = y, a' = b, \\ w_y(a') & \text{otherwise}, \end{cases}$$

$$v^2_x(a') = \begin{cases} 0 & \text{if } x = y, a' = a, \\ w_x(a) + w_x(b) & \text{if } x = y, a' = b, \\ w_y(a') & \text{otherwise}, \end{cases}$$

Define

$$q_1 = \frac{w_y(a)}{w_y(a) + w_y(b)} \qquad q_2 = \frac{w_y(b)}{w_y(a) + w_y(b)}.$$

The policy w behaves in the same way as if at each time we return to state y we toss a coin, and with probability q_i it will use the fixed stationary policy v^i, $i = 1, 2$, until the next visit to state y.

We have

$$E^w T = q_1 E^{v^1}[T] + q_2 E^{v^2}[T],$$

and

$$E^w N(x, a) = q_1 E^{v^1}[N(x, a)] + q_2 E^{v^2}[N(x, a)].$$

This, together with (4.9), implies that for any $x \in \mathbf{X}$,

$$\begin{aligned} f^w(x, a) &= \frac{q_1 E^{v^1}[N(x, a)] + q_2 E^{v^2}[N(x, a)]}{q_1 E^{v^1}[T] + q_2 E^{v^2}[T]} \\ &= \frac{q_1 f^{v^1}(x, a) E^{v^1}[T] + q_2 f^{v^2}(x, a) E^{v^2}[T]}{q_1 E^{v^1}[T] + q_2 E^{v^2}[T]} \\ &= p_1 f^{v^1}(x, a) + p_2 f^{v^1}(x, a), \end{aligned}$$

where

$$p_1 = \frac{q_1 / E^{v^2}[T]}{q_1 / E^{v^2}[T] + q_2 / E^{v^1}[T]}, \quad p_2 = \frac{q_2 / E^{v^1}[T]}{q_1 / E^{v^2}[T] + q_2 / E^{v^1}[T]}.$$

We conclude that f^w is not an extreme point of $\mathcal{L}_{U_S}$, which concludes the proof. □

4.2 Equivalent Linear Program

We now obtain an LP formulation similar to the one we obtained for the discounted cost; we show again that the **COP** is equivalent to an LP with

a countable set of decision variables and a countable set of constraints. Consider the following LP:
LP$_3$(β): Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle \rho, c \rangle$ subject to:

$$\mathcal{D}^k(\rho) := \langle \rho, d^k \rangle \leq V_k, k = 1, \ldots, K, \qquad \rho \in \mathbf{Q}_{ea}(\beta),$$

where $\mathbf{Q}_{ea}(\beta)$ was defined in (4.3).

Define $w(\rho)$ to be any stationary policy such that

$$w_y(a) = \rho(y,a) \Big(\sum_{a \in \mathsf{A}(y)} \rho(y,a) \Big)^{-1}$$

whenever the denominator is non-zero. We show that there is a one to one correspondence between feasible (and optimal) solutions of the LP, and the feasible (and optimal) solutions of **COP**.

Theorem 4.3 (*Equivalence between COP and* **LP$_3$**(β))
(*i*) $\mathcal{C}^* = C_{ea}(\beta)$.
(*ii*) *For any* $u' \in U$, *there exists a dominating stationary policy* $u \in U_S$ *such that* $\rho(u) := f_{ea}(\beta, u) \in \mathbf{Q}_{ea}(\beta)$, $C_{ea}(\beta, u) = \mathcal{C}(\rho(u))$ *and* $D_{ea}(\beta, u) = \mathcal{D}(\rho(u))$; *conversely, for any* $\rho \in \mathbf{Q}_{ea}(\beta)$, *the stationary policy* $w(\rho)$ *satisfies* $C_{ea}(\beta, w(\rho)) = \mathcal{C}(\rho)$ *and* $D_{ea}(\beta, w(\rho)) = \mathcal{D}(\rho)$.
(*iii*) **LP$_3$**(β) *is feasible if and only if* **COP** *is feasible. Assume that* **COP** *is feasible. Then there exists an optimal solution* ρ^* *for* **LP$_3$**(β), *and the stationary policy* $w(\rho^*)$ *is optimal for* **COP**.

Proof. We start from (ii). The dominance of stationary policies was proved in Theorem 4.1. The other claims follow by combining the linear representation of the cost (4.2) with the proofs of Theorems 4.1, 4.2. This establishes (ii). This then implies (i) and (iii). □

4.3 The Dual Program

Next, we present the formal Dual Program DP for the LP above. The decision variables are $\psi \in \mathbb{R}$, $\phi : \mathbf{X} \to \mathbb{R}$ and the K-dimensional non-negative vectors $\lambda \in \mathbb{R}^K_+$.

$$\begin{array}{lll} \textbf{DP}_3(\beta): & \text{Find } \Theta^*(\beta) := \sup_{\psi,\phi,\lambda} \psi - \langle \lambda, V \rangle \text{ subject to} & \\ \phi(x) + \psi & \leq & c(x,a) + \langle \lambda, d(x,a) \rangle + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y), \quad x \in \mathbf{X}, a \in \mathsf{A}(x). \end{array}$$

A probabilistic interpretation for this LP can be obtained in the same way as we obtained the dual LP for the discounted cost criterion, see Chapter 3. In particular, we can obtain the corresponding results for the Lagrangian. This is done in detail for the infinite MDPs in Chapter 12.

4.4 Number of randomizations

As we did for the discounted cost, we show that when **COP** is feasible, then there exists an optimal stationary policy w that requires at most K randomizations (where we use the same Definition 3.3 for the number of constraints). The result below is due to Koole (1988) and Ross (1989).

Theorem 4.4 (*Bound on the number of randomizations*)
Consider the expected average cost with the unichain assumption.
If **COP** *is feasible, then there exists an optimal stationary policy w such that the total number $n(w)$ of randomizations that it uses is at most K (where K is the number of constraints).*

Proof. The proof is very similar to the discounted cost. We indicate below the changes in the proof compared to that of the discounted case (Theorem 3.8).

In the discounted case, $\mathbf{LP}_{\mathbf{1}}^{\alpha}(\beta)$ had $|\mathbf{X}| + K$ constraints, whereas in the expected average cost $\mathbf{LP}_{\mathbf{3}}(\beta)$has $|\mathbf{X}| + K + 1$ constraints. The additional constraint which appears here explicitly is $\sum_{y,a} \rho(y,a) = 1$. (This is in fact an implicit constraint in the discounted cost as well, which can be seen by summing all the equality constraints in the definition of $\mathbf{Q}^{\alpha}$.)

However, unlike the discounted case, if we sum the first $|\mathbf{X}|$ equality constraints in the definition of $\mathbf{Q}_{ea}(\beta)$, we get "0=0"; in other words, we see that these are dependent constraints. The number of independent constraints in $\mathbf{LP}_{\mathbf{3}}(\beta)$ is thus upper-bounded by $|\mathbf{X}| + K$, as is the case for the discounted cost.

From this point, the rest of the proof is exactly the same as that of Theorem 3.8 (the discounted cost). □

CHAPTER 5

Flow and service control in a single-server queue

We consider below a problem of flow and service control in a single queue. Although there are two controllers, they can be viewed as one control entity since they have a common cost to minimize. Formulating this problem as a constrained MDP and using an equivalent Lagrangian approach, we show that there exist optimal stationary policies with a special monotone structure. In particular, when any one of the controllers has only two available actions, then it has an optimal randomized threshold policy. We show that if the thresholds are different, then the randomizations can be performed independently.

We then pose the question of whether in CMDP with several controllers, one can always restrict to decentralized policies, i.e., policies for which the controllers perform their randomizations independently. We illustrate via a simple example that the answer is negative.

5.1 The model

We consider a discrete-time single-server queue with a buffer of finite size L. We assume that at most one customer may join the system in a time slot. This possible arrival is assumed to occur at the beginning of the time slot. The state corresponds to the number of customers in the queue at the beginning of a time slot.

Let a_{min} and a_{max} be two real numbers satisfying $0 < a_{min} \leq a_{max} < 1$. At the end of the slot, if the queue is non-empty and if the action of the server is a, then a service of a customer is successfully completed with probability $a \in \mathsf{A}$ where A is a finite subset of $[a_{min}, a_{max}]$. If the service fails, the customer remains in the queue; and if it succeeds, then the customer leaves the system.

Let b_{min}, b_{max} be two real numbers satisfying $0 \leq b_{min} \leq b_{max} < 1$. At the beginning of each time slot, if the state is x, then the flow controller chooses an action b from a finite set $\mathsf{B}(x) \subset [b_{min}, b_{max}]$. In this case, the probability of having one arrival during this time slot is equal to b. We assume that $0 \in \mathsf{B}(x)$ for all x; moreover, when the buffer is full, no arrivals are possible ($\mathsf{B}(L) = \{0\}$). In all states other than L, we assume that the

available actions for the flow controller are the same, and we denote them by $\mathsf{B}(x) = \mathsf{B}$.

Our control thus consists of two components, and the set of actions is $\mathsf{A} \times \mathsf{B}$. If the service control a is fixed, then the service time distribution is Bernoulli with parameter a, and the expected service time is $1/a$. a can thus be interpreted as the allocated bandwidth, or quality of service.

We assume that a customer who enters an empty system may leave the system (with probability a, when action a is used) at the end of this same time slot.

The transition law $\mathcal{P}$ is:

$$P_{xaby} := \begin{cases} \bar{b}a, & \text{if } L \geq x \geq 1,\ y = x-1; \\ ba + \bar{b}\bar{a}, & \text{if } L \geq x \geq 1,\ y = x; \\ b\bar{a}, & \text{if } L > x \geq 0,\ y = x+1; \\ 1 - b\bar{a}, & \text{if } y = x = 0; \end{cases}$$

(for any number $\chi \in [0,1]$, $\bar{\chi} := 1 - \chi$).

We assume that there are three components for the cost: a holding cost c, a cost d^1 corresponding to the actions of the service controller, and a cost d^2 related to the actions of the flow controller.

The immediate cost c is related to the component that we wish to minimize (i.e., to the delays) and is assumed to be a function only of the state. We assume that $c(x)$ is a real-valued *increasing convex* function on $\mathbf{X}$. c can be interpreted as a holding cost. We know from Little's law that the expected queue length is proportional to the expected waiting time, and therefore if c is linear, it is directly related to the delay.

The immediate cost d^1 corresponding to the service rate is assumed to be a function only of a. It can be interpreted as a cost function per quality of service or per bandwidth allocation. The cost d^2 corresponding to the throughput is assumed to be a function only of b. It is natural to assume that d^1 is increasing in a and $d^1 \geq 0$, whereas d^2 is decreasing in b.

We consider the discounted costs $C_\alpha(\beta, u)$ and $D^k_\alpha(\beta, u) \leq V_k$, $k = 1, 2$, all defined as in (2.4), for some given discount factor α. We wish to minimize $C_\alpha(\beta, u)$ under the constraints $D^k_\alpha(\beta, u) \leq V_k$, $k = 1, 2$; V_k are some given constants.

By using Theorem 3.6 (iii), we know that the expected discounted cost satisfies

$$C(\beta) = \min_{u \in U_S} \sup_{\lambda \geq 0} J^\lambda(\beta, u); \tag{5.1}$$

$J^\lambda(\beta)$ is the difference between the discounted cost corresponding to the immediate cost

$$j^\lambda(x, a, b) = c(x) + \lambda_1 d^1(a) + \lambda_2 d^2(b), \tag{5.2}$$

and $\sum_{k=1}^2 \lambda_k V_k$, see (9.24).

Remark 5.1 (*Other cost criteria*)
The above also holds for the expected average cost or for the so-called total expected cost criteria, as will be shown in Corollaries 9.2 and 12.2. Note that for the expected average cost, the Lagrangian is defined directly through (12.29), but since one may restrict to stationary policies (without loss of optimality), (5.1) can be used with the definition using (5.2).

We shall restrict our analysis to the discounted cost. Standard methods can be used to show that the structure of optimal policies carries over also to the expected average cost. In particular, one may use the results of Section 14.2 to establish this. We thus consider the problem of

$$\min_{u\in U} C_\alpha(\beta,u) \;\text{ subject to }\; D^k_\alpha(\beta,u)\le V_k,\quad k=1,2, \tag{5.3}$$

where β is the fixed initial distribution, V_k are some given constants, and C and D are defined with respect to the immediate costs defined above.

5.2 The Lagrangian

We consider in this section the equivalent non-constrained problem of minimizing the Lagrangian

$$J^\lambda_\alpha(\beta,u)=\sum_{t=1}^{\infty}\alpha^{t-1}E^u_\beta j^\lambda(X_t,A_t).$$

(We ignore the term $\sum_k \lambda_k V_k$ which has to be further subtracted, as it does not depend on the policy; in particular, it has no influence on the policy that minimizes the Lagrangian.) Below, we shall omit λ from the notation.

It follows from Theorem 3.6 that there exists an optimal stationary policy for the Lagrangian which is optimal for the original constrained problem.

We shall therefore restrict, without loss of optimality, to stationary policies.

Definition 5.1 *For a stationary policy v, define the projection $v^a_x(a) \stackrel{\text{def}}{=} v_x(a,\mathsf{B})$. Define similarly $v^b_x(b) \stackrel{\text{def}}{=} v_x(\mathsf{A},b)$.*

Remark 5.2 Note that for deterministic policies, the projections fully determine the original policy.

We now describe the type of monotonicity of optimal policies that will occur in our problem. Let $u:\mathbf{X}\to M_1(\mathsf{A})$. Denote $a^{sup}_x(u) :=$ the greatest a in the support of u_x, i.e., the greatest $a\in\mathsf{A}$ such that $u_x(a)>0$. Denote $a^{inf}_x(u) :=$ the smallest a in the support of u_x.

We say that u is **strongly monotone decreasing** in its a-component if for any $x\in\mathbf{X}$ and y with $y<x$, $a^{inf}_y(u)\ge a^{sup}_x(u)$.

We say that u is **strongly monotone increasing** in its a-component if for any $x\in\mathbf{X}$ and y with $y<x$, $a^{sup}_y(u_t)\le a^{inf}_x(u_t)$.

The analogous definitions hold naturally for actions b. As a direct consequence of the definition of strongly monotone policies, we have

Lemma 5.1 *If u is strongly monotone in its a-component, then u chooses a singleton among A in at least m states, where $m = |\mathbf{X}| - |\mathsf{A}| + 1$. A similar statement holds for the b-component.*

Proof. Assume that u is strongly monotone increasing. Let x be a state at which u_x randomizes between more than one action. Then for all $y > x$, an action a with $a > a_x^{inf}(u)$ is not used, i.e., $u_y(a) = 0$. The set of actions $\mathcal{A}(x,u) \stackrel{\text{def}}{=} \{a : \exists y > x \text{ such that } u_y(a) > 0\}$ is monotone decreasing, and in particular, it decreases by at least one action at each x as above: $|\mathcal{A}(x-1,u)| - 1 \geq |\mathcal{A}(x)|$. This implies the proof. □

Theorem 5.1 (*Structure of optimal policies*)
Fix some initial distribution β. Assume that the holding cost c is convex non-decreasing, and either $c(1) > c(0)$ or $c(2) - c(1) > c(1) - c(0)$. Then any optimal stationary policy is strongly monotone in both a- and b-components. If $\mathsf{A} = \{a_1, a_2\}$ with $a_1 < a_2$, then the a-projection of any optimal policy is of a randomized threshold type, i.e.,

$$u_x^a(a_2) = \begin{cases} 1 & \text{if } x > m_a \\ q_a & \text{if } x = m_a \\ 0 & \text{if } x < m_a \end{cases} \tag{5.4}$$

where m_a is an integer, and $q_a \in [0,1]$ some real number.
If $\mathsf{B} = \{b_1, b_2\}$ with $b_1 < b_2$, then the b-projection of any optimal policy is of a randomized threshold type, i.e.,

$$u_x^b(b_2) = \begin{cases} 0 & \text{if } x > m_b \\ q_b & \text{if } x = m_b \\ 1 & \text{if } x < m_b \end{cases} \tag{5.5}$$

where m_b is an integer, and $q_b \in [0,1]$ some real number.

Before proving the theorem, we need some definitions and some auxiliary results, which we establish in three lemmas.

Let $\mathcal{N}$ be the set of real-valued functions on $\mathbf{X}$. Define the operator $R : \mathbf{X} \times \mathsf{A} \times \mathsf{B} \times \mathcal{N} \to \mathbb{R}$ as

$$R(x,a,b,f) := E\left[f(X_{t+1}) | X_t = x, A_t = a, B_t = b\right].$$

We get:

$$R(x,a,b,f) = \begin{cases} (1 - b\bar{a})\, f(x) + b\bar{a} f(x+1) & x = 0 \\ \bar{b}a f(x-1) + (ba + \bar{b}\bar{a}) f(x) + b\bar{a} f(x+1) & 1 \leq x \leq L \end{cases} \tag{5.6}$$

(in the above equation we shall understand $0 \cdot f(L+1) := 0$.) Let $R(x,f)$ denote the vector whose entries are $R(x,a,b,f)$ (it has one entry for each pair (a,b)).

Define the operator $S : \mathbf{X} \times \mathsf{A} \times \mathsf{B} \times \mathcal{N} \to \mathbb{R}$ as

$$S(x, a, b, f) := j(x, a, b) + \alpha R(x, a, b, f),$$

and let $S(x, f)$ denote the vector whose entries are $\{S(x, a, b, f)\}$ (it has one entry for each pair (a, b)).

In order to simplify the analysis of the boundary (at $x = 0$), we shall extend functions of the form $f : \mathbf{X} \to \mathbb{R}$ to $\mathbf{X} \cup \{-1\} \to \mathbb{R}$, and set $f(-1) = f(0)$. With this definition we have *for all* $0 \le x \le L$, a and b:

$$R(x, a, b, f) = \bar{b}af(x - 1) + (ba + \bar{b}\bar{a})f(x) + b\bar{a}f(x + 1).$$

Let $T_\alpha : \mathcal{N} \to \mathcal{N}$ be the dynamic programming operator associated with our minimization problem:

$$T_\alpha f(x) \stackrel{\text{def}}{=} \min_{a,b} S(x, a, b, f), \qquad x \in \mathbf{X}.$$

We shall extend the image of T_α to functions over $\mathbf{X} \cup \{-1\}$ and define $T_\alpha f(-1) \stackrel{\text{def}}{=} T_\alpha f(0)$. We similarly define $J_\alpha(-1) = J_\alpha(0)$.

We shall use the following:

Lemma 5.2 *(i) J_α satisfies*

$$J_\alpha(x) = T_\alpha J_\alpha(x).$$

(ii) A policy u is uniformly optimal if and only if it achieves the argmin in $T_\alpha J_\alpha(x)$.
(iii) For any $f \in \mathcal{N}$, $\lim_{n\to\infty} T_\alpha^n f = J_\alpha$.

Proof. (i) and (ii) follow from Theorem 3.4. (iii) follows from a well-known value iteration theorem, see e.g., Wessels (1977) (for more details, see Chapter 15). □

Remark 5.3 (*The finite horizon problem*)
$T_\alpha^n f$ is in fact the optimal value of an n-step horizon discounted problem with immediate cost j^λ and final cost f. Lemma 5.2 (iii) shows that the optimal finite horizon cost converges to the infinite cost. In Chapter 15 we prove this convergence in a more general setting: that of constrained MDPs.

Lemma 5.3 *Let $h : \mathbf{X} \cup \{-1\} \to \mathbb{R}$ be a non-decreasing function with $h(-1) = h(0)$. Let $\zeta_1, \zeta_2 \in [0, 1]$. Then, for all $0 \le x < L$,*

$$F(x) := \zeta_2 h(x + 1) + \bar{\zeta}_2 h(x) - \zeta_1 h(x) - \bar{\zeta}_1 h(x - 1) \ge 0 \qquad (5.7)$$

Moreover, if (i) $h(x + 1) > h(x)$ and $\zeta_2 \ne 0$, or (ii) if $h(x) > h(x - 1)$ and $\zeta_1 \ne 1$, then $F(x) > 0$.
Proof.

$$F(x) \ge h(x) - \zeta_1 h(x) - \bar{\zeta}_1 h(x - 1) = \bar{\zeta}_1[h(x) - h(x - 1)] \ge 0,$$

and the second claim follows similarly. □

We shall say that $f \in \mathcal{N}$ satisfies assumption:

- **WC** (weakly convex) if for all $0 \leq x < L-1$,

$$f(x+2) - f(x+1) \geq f(x+1) - f(x). \tag{5.8}$$

- **SC(x)** (strongly convex) if for x given,

$$f(x+2) - f(x+1) > f(x+1) - f(x). \tag{5.9}$$

- **MI** if $f(x)$ is monotone increasing in x, i.e., for any $0 \leq x < L$,

$$f(x+1) \geq f(x). \tag{5.10}$$

Lemma 5.4 *Assume that the holding cost c satisfies* **WC** *and* **MI**.
(i) Assume that f satisfies **WC** *and* **MI**. *Then $T_\alpha f$ satisfies* **WC** *and* **MI**.
(ii) The value function J_α satisfies **WC** *and* **MI**.
(iii) If J_α satisfies **SC(x)** *in one state $x < L-1$, then it satisfies* **SC(y)** *for all $y \geq x$. If $J_\alpha(1) - J_\alpha(0) > 0$, then J_α satisfies* **SC(y)** *for all states y, $0 \leq y < L-1$.*
Finally, assume that the holding cost c satisfies **WC**, **MI** *and either $c(1) > c(0)$ or* **SC(0)**. *Then,*
(iv) J_α satisfies **SC(y)** *for all states y, $0 \leq y < L-1$.*

Proof. Let $\overline{U}(f)$ be the set of stationary policies that achieve the argmin of $T_\alpha f$. Choose any $0 \leq z \leq L-1$. Choose some $u \in \overline{U}(f)$, and for all x, select some (a_x, b_x) in the support of u_x (i.e., $u_x(a_x, b_x) > 0$). Hence, $(a_x, b_x) \in T_\alpha f(x)$.

We begin by establishing **MI**. Defining $a = a_{x+1}, b = b_{x+1}$, we have

$$\begin{aligned} T_\alpha f(x+1) - T_\alpha(x) &= S(x+1, a_{x+1}, b_{x+1}, f) - S(x, a_x, b_x, f) \\ &\geq S(x+1, a_{x+1}, b_{x+1}, f) - S(x, a_{x+1}, b_{x+1}, f) \\ &= c(x+1) - c(x) \qquad (5.11) \\ &\quad +\alpha \Big\{ a\bar{b}[f(x) - f(x-1)] + (ab + \bar{a}\bar{b})[f(x+1) - f(x)] \\ &\quad +\bar{a}b[f(x+2) - f(x+1)] \Big\} \\ &\geq c(1) - c(0) \geq 0. \qquad (5.12) \end{aligned}$$

(The equation above holds indeed for $x = L-1$, too, since we then have $b = 0$; in that case, we shall understand $bf(x+2) := 0$).

Next we check **WC**. Choose any $0 \leq x \leq L-2$; denote

$$F(x) \stackrel{\text{def}}{=} \min_u S(x+2, f) - \min_u S(x+1, f) - [\min_u S(x+1, f) - \min_u S(x, f)].$$

Define $a_2 \stackrel{\text{def}}{=} a_{x+2}, b_2 \stackrel{\text{def}}{=} b_{x+2}, a_1 \stackrel{\text{def}}{=} a_x, b_1 \stackrel{\text{def}}{=} b_x$. We have

$$\begin{aligned} F(x) \quad \geq \quad & S(x+2, a_{x+2}, b_{x+2}, f) - S(x+1, a_{x+2}, b_{x+2}, f) \\ & -[S(x+1, a_x, b_x, f) - S(x, a_x, b_x, f)] \\ = \quad & c(x+2) - c(x+1) - c(x+1) + c(x) + \\ & \alpha \Big\{ a_2 \left[b_2(f(x+2) - f(x+1)) + \overline{b}_2(f(x+1) - f(x)) \right] \\ & \quad + \bar{a}_2 \left[b_2(f(x+3) - f(x+2)) + \bar{b}_2(f(x+1) - f(x)) \right] \\ & - a_1 \left[b_1(f(x+1) - f(x)) + \bar{b}_1(f(x) - f(x-1)) \right] \\ & \quad - \bar{a}_1 \left[b_1(f(x+2) - f(x+1)) + \bar{b}_1(f(x+1) - f(x)) \right] \Big\} \end{aligned} \tag{5.13}$$

$$\geq 0 \tag{5.14}$$

which follows by applying Lemma 5.3 with

$$\zeta_2 = \bar{a}_2, \ \zeta_1 = \bar{a}_1, \quad h(x) = b_1(f(x+2) - f(x+1)) + \bar{b}_1(f(x+1) - f(x)).$$

Since f satisfies **WC**, h is indeed increasing. The equation above holds for $x = L-2$, too, since in that case $g_2 = 0$; we shall then understand $g_2 f(x+3) := 0$.

(ii) Choose $f(x) = 0, \forall x \in \mathbf{X}$. By repeated application of Lemma 5.4 (i), it follows that $T_\alpha^n f$ satisfies **MI** and **WC** for $n = 1, 2, \ldots$; moreover, $\lim_{n\to\infty} T_\alpha^n f$ satisfies **MI** and **WC**. Hence by Proposition 5.2 (iii), J_α satisfies **MI** and **WC**.

(iii) Suppose that J_α satisfies **SC(x-1)** for some fixed $0 < x < L-1$. By substituting J_α instead of f in (5.13) and again applying Lemma 5.3 (this time we apply the second part of the Lemma; indeed condition (ii) there holds since b_1 cannot be equal to one, and $h(x) = J_\alpha(x+1) - J_\alpha(x)$ satisfies $h(x) > h(x-1)$ by the assumption), we thus get strict inequality in (5.14). Hence J_α satisfies **SC(x)** as well, and we conclude similarly that it satisfies **SC(y)** for any $y \geq x$.

To prove the second claim, we again substitute J_α instead of f in (5.13) and consider $x = 0$. Again we have the case of the strict inequality in Lemma 5.3 since $h(x) := J_\alpha(x+1) - J_\alpha(x)$ satisfies indeed $h(x) - h(x-1) = J_\alpha(1) - J_\alpha(0) > 0$ (recall that $h(0) := 0$ since $J_\alpha(-1) := J_\alpha(0)$). We thus get again strict inequality in (5.14). It follows that $J_\alpha(0)$ satisfies **SC(0)**, and hence by the first claim, it satisfies **SC(y)** for all $0 \leq y < L-1$.

(iv) Fix $x = 0$. Assume $c(1) > c(0)$. It follows that (5.12) holds with strict inequality for any f satisfying **MI** and in particular for $f = J_\alpha$. Hence J_α, which by Proposition 5.2 (i) is equal to $\min_u S(x, J_\alpha)$, satisfies

$$J_\alpha(1) - J_\alpha(0) \geq c(1) - c(0) > 0.$$

The proof is then established by applying the second part of (iii).

Next assume that c satisfies **SC(0)**. Substituting J_α into (5.13) and considering $x = 0$ we get $F(x) > 0$ since we have a strict inequality in (5.14). Hence J_α satisfies **SC(0)**. The proof is then established by applying the first part of (iii). □

Proof of Theorem 5.1: (i) Let $\overline{U}$ be the set of stationary policies with the property that each $u \in \overline{U}$ achieves the argmin of $T_\alpha J_\alpha$. It follows from Theorem 3.4 that a stationary policy is optimal if and only if it is in $\overline{U}$. (The theorem is stated for the total expected cost of which the discounted cost problem can be viewed as a special case, as discussed at the end of Section 10.4. Note that for any stationary policy u and state x, $f_\alpha(\beta, u; x) > 0$ since under any u, every state can be reached from any other state. Hence the condition in Theorem 9.2 holds.)

Fix some $x < L$. Assume that for any a_1 and a_2, and for any b_2 and b_1 satisfying $b_2 > b_1$,

$$\begin{aligned}\Delta^b(x) \quad &\stackrel{\text{def}}{=} \quad S(x+1, a_1, b_2, J_\alpha) - S(x+1, a_1, b_1, J_\alpha) \\ &\quad -[S(x, a_2, b_2, J_\alpha) - S(x, a_2, b_1, J_\alpha)] > 0.\end{aligned} \tag{5.15}$$

Assume the $u^b_x(b_1) > 0$. This implies, by definition, that there exists some a' such that $(a', b_1) \in argmin T_\alpha J_\alpha(x)$. This implies that $[S(x, a', b_2, J_\alpha) \geq S(x, a', b_1, J_\alpha)]$. (5.15) then implies that for all $y > x$ and all a,

$$S(y, a, b_2, J_\alpha) > S(y, a, b_1, J_\alpha),$$

so that $(a, b_2) \notin argmin T_\alpha J_\alpha(y)$. Hence $u^b_y(b_2) = 0$, which shows that u^b is strongly monotone decreasing.

We similarly obtain the monotonicity for u^a. Fix some $x < L$. Assume that for any b_1 and b_2, and for any a_2 and a_1 satisfying $a_2 > a_1$,

$$\begin{aligned}\Delta^a(x) \quad &\stackrel{\text{def}}{=} \quad S(x+1, a_2, b_1, J_\alpha) - S(x+1, a_1, b_1, J_\alpha) \\ &\quad -[S(x, a_2, b_2, J_\alpha) - S(x, a_1, b_1, J_\alpha)] < 0.\end{aligned} \tag{5.16}$$

Assume the $u^a_x(a_2) > 0$. This implies, by definition, that there exists some b' such that $(a_2, b') \in argmin T_\alpha J_\alpha(x)$. This implies that $[S(x, a_2, b', J_\alpha) \leq S(x, a_1, b', J_\alpha)]$. (5.16) then implies that for all $y > x$ and all b,

$$S(y, a_2, b, J_\alpha) < S(y, a_1, b_1, J_\alpha),$$

so that $(a_1, b) \notin argmin T_\alpha J_\alpha(y)$. Hence $u^b_y(a_1) = 0$, which shows that u^b is strongly monotone increasing.

It remains to show that (5.15) and (5.16) indeed hold.

$$\begin{aligned}\Delta^a(x) \quad &= \quad \alpha(b_2 - b_1) \times \\ &\qquad \Big(a_1[J_\alpha(x+1) - J_\alpha(x)] + \bar{a}_1[J_\alpha(x+2) - J_\alpha(x+1)]\end{aligned}$$

$$- (a_2[J_\alpha(x) - J_\alpha(x-1)] + \bar{a}_2[J_\alpha(x+1) - J_\alpha(x)])\Big)$$
$$> \quad 0 \tag{5.17}$$

where the last inequality follows from Lemma 5.3 with $\zeta_2 = \bar{a}_1$, $\zeta_1 = \bar{a}_2$ and $h(x) = J_\alpha(x) - J_\alpha(x-1)$, and since, by Lemma 5.4 (iv), J_α satisfies **SC(x)** for all x. This establishes the monotonicity of u^a. The monotonicity of u^b follows from similar arguments. □

5.3 The original constrained problem

We now go back to the original constrained MDP.

Theorem 5.2 (*Optimality of strongly monotone policies*)
Choose some β, and consider the discounted constrained problem (5.3). Assume that the CMDP is feasible. Then,
(*i*) *There exists an optimal stationary policy.*
(*ii*) *If the holding cost c is convex non-decreasing, and either $c(1) > c(0)$ or $c(2) - c(1) > c(1) - c(0)$, then any optimal stationary policy is strongly monotone increasing in the a-component and strongly monotone decreasing in the b-component.*
(*iii*) *In particular, if* A *has only two actions, then the a-projection of any optimal policy is of a randomized threshold type given in Theorem* 5.1. *The same statement holds for the b-projection.*

Proof. Fix some initial distribution β. According to Theorem 5.1, for any value of the Lagrange multipliers, all stationary policies that are optimal for the Lagrange problem are strongly monotone in both components. According to Theorem 3.6 among all the stationary optimal policies that are optimal for the Lagrangian problem, there exists one that is optimal for the constrained one. Hence this one is indeed strongly monotone in both components, which establishes the proof. □

5.4 Structure of randomization and implementation issues

We assume in this section that there are only two actions in A and two actions in B. Using the results of the previous section, we conclude that if an optimal stationary policy satisfies $m_a \neq m_b$ (defined in Theorem 5.1), then the randomizations occur at different states for each component; for all x other than these two thresholds, deterministic actions are taken and there is no randomization. At m_a only one component uses randomization, the one corresponding to the service control; at m_b only the flow controller randomizes.

At no state is there a need to jointly randomize over all components of $\mathsf{A} \times \mathsf{B}$. The optimal policy is fully described by its projections and can be

implemented and performed in a decentralized way, without coordinating between the flow and service controller to do a joint randomization.

This is not the case if $m_a = m_b$; in this case, the optimal stationary policy does not use randomization at any other state; but at the particular state m_a, it might randomize between 3 pairs of actions of the form (a_i, b_j). In general, this cannot be performed in a decentralized way. For example, a policy that randomizes between (a_1, b_1) and (a_2, b_2) cannot be implemented by performing randomization independently on the A and B components, since independent randomizations will give rise also to pairs of the form (a_1, b_2) and (a_2, b_1).

Our structural results are compatible with the fact that we know that there exist optimal stationary policies for CMDPs that use at most K randomizations, where K are the number of constraints (this was proved in Sections 3.5 and 4.4). Now suppose that instead of having one holding cost $c(x)$, we had two holding costs, $c(x)$ and $d^3(x)$, (in addition to the costs d^1 and d^2 related to the actions). $c(x)$ may be the cost related to the delay, and $d^3(x)$ may be an actual holding cost. Consider now the discounted cost with an additional constraint of the form $D^3(\beta, u) \leq V_3$. Assuming that d^3 has the same convexity and monotonicity properties as assumed in Theorem 5.2 (ii), we may repeat the arguments based on the Lagrangian, and show that here again, any optimal stationary policy has a strongly monotone structure for each component. This implies, in the case that $m_a \neq m_b$, that only two randomization are required. This is less than three randomizations that might be needed in a general CMDP.

5.5 On coordination between controllers

The following example illustrates the need for coordination in order to perform a joint randomization when there is more than one controller. We consider a control of an i.i.d. process, i.e., a CMDP with a single state in which the system always remains.

Example 5.1 *(The need for coordination)*
Consider $\mathbf{X} = \{x\}$ in which the system always remains (under any action). Since $\mathbf{X}$ is a singleton, we shall omit it from the notations of the costs. There are two controllers, whose actions are $\mathsf{A} = \{a_1, a_2\}$ and $\mathsf{B} = \{b_1, b_2\}$, respectively. Consider the costs:

$$\begin{aligned} c(a_1, b_1) &= 0 \quad d(a_1, b_1) = 4 \\ c(a_1, b_2) &= N \quad d(a_1, b_2) = N \\ c(a_2, b_1) &= N \quad d(a_2, b_1) = N \\ c(a_2, b_2) &= 4 \quad d(a_2, b_2) = 2. \end{aligned}$$

N is some large number, say $N \geq 100$.

For any stationary policy u, which chooses an action pair (a_i, b_j) with

probability $u(a_i, b_j)$, the expected average cost as well as the discounted cost for any discount factor $\alpha \in [0, 1)$ have the same value, given by

$$\begin{aligned} C(u) &= Nu(a_1, b_2) + Nu(a_2, b_1) + 4u(a_2, b_2) \\ D(u) &= 4u(a_1, b_1) + Nu(a_1, b_2) + Nu(a_2, b_1) + 2u(a_2, b_2). \end{aligned}$$

With $V := 3$, **COP** is

$$\text{minimize } C(u) \text{ subject to } D(u) \leq V.$$

For all N large enough, there is a unique optimal stationary policy u^* given by: $u^*(a_1, b_1) = u^*(a_2, b_2) = 1/2$, $u^*(a_1, b_2) = u^*(a_2, b_1) = 0$, whose value is $C = C(u^*) = 2$. It achieves $D(u^*) = 3$.

Consider now decentralized policies, i.e., policies in which randomizations are performed independently between the players. Let u be such a policy, and denote by $p := u^a(a_1)$ and $q := u^b(b_1)$ the probabilities that controller 1 chooses a_1 and that controller 2 chooses b_1. Then $u(a_1, b_1) = pq$, etc. The constraint $D(u) \leq V$ becomes:

$$D(u) = 4pq + Np(1 - q) + N(1 - p)q + 2(1 - p)(1 - q) \leq V.$$

We make the following observations:

- A necessary condition for this to hold is that $p(1 - q) \leq V/N$, so that either p is very small, or q is very close to 1 (for N large).
- A second necessary condition is that $(1 - p)q \leq V/N$, so that either p is very close to 1, or q very close to 0.

In order for both conditions to hold simultaneously, which is necessary for having $D(u) \leq V$, we need either p and q to be both close to 1, or both close to 0. The first possibility is however unfeasible, since it gives $D(u) \sim 4$. We thus conclude that any feasible stationary policy u has the property of choosing p and q close to 0. It then follows that any feasible policy has a value $C(u)$ around 4. This is twice the value we obtained by coordinating the randomization.

□

We conclude that stationary decentralized policies are not as good as those stationary policies that allow coordination, and restricting ourselves to them would result in non-optimal performance.

5.6 Open questions

The following questions are naturally posed for general CMDPs with a finite number of controllers:

(i) Having seen that optimal stationary decentralized policies do not exist in general, is there an optimal non-stationary decentralized policy, (in the sense that randomizations are always performed independently)?

(ii) If we restrict *a priori* **COP** to decentralized policies (in the sense that randomizations are always performed independently), does there exist an optimal policy?

(iii) If we restrict *a priori* **COP** to stationary decentralized policies, is there an optimal stationary policy among these?

(iv) How do we compute optimal or ε-optimal policies in (i) and (iii)?

The answer to (i) is positive in some cases. For the case of expected average cost with a unichain ergodic structure, randomization (in particular, one with coordination) can be replaced by deterministic decisions under fairly general conditions. Example of such policies are the ATS (Action Time Sharing) policies introduced in Section 3 in Altman and Shwartz (1991), and the PTS (Policy Time Sharing) introduced in Altman and Shwartz (1989, 1993). These are by definition decentralized, since they do not randomize. In cases other than the expected average cost, the answer to (i) can be negative. For example, if we consider the discounted costs with a discount factor of 0, then only the decisions at time 1 count, so that non-optimality among stationary policies means non-optimality among any class of policies.

Whenever one can replace randomized policies by equivalent non-randomized ones, then there exist optimal decentralized policies. Thus, if we restrict *a priori* to decentralized policies, we can still find one that is not only optimal for the decentralized problem but also for the original one. This, together with the answers to the previous point, is a partial answer to point (ii).

We now provide a partial answer to point (iii). Assume that we now restrict ourselves to decentralized stationary policies. Consider a **COP** with a finite set of actions and of states, two controllers (users), and either (i) the discounted cost is considered, or (ii) the expected average cost is used and the unichain assumption holds. Assume that for any stationary policy u of user 1, there is a stationary policy v for user 2, and vice versa, such that all the constraints are satisfied with strict inequalities:

$$D^k(\beta,(u,v)) < V_k, \quad k = 1,\dots,K.$$

Then if we restrict **COP** to stationary decentralized policies, there is an optimal one within this class. This result follows from a more general one in Altman and Shwartz (1995) on constrained games (these are CMDPs where each user has a different objective to minimize, and different constraints).

As for question (iv), computing optimal decentralized ATS or PTS policies from the given corresponding stationary policies is described by Altman and Shwartz (1993). The question of computing optimal policies when restricting to the class of decentralized stationary policies remains open.

PART II

Part Two: Infinite MDPs

CHAPTER 6

MDPs with infinite state and action spaces

We now generalize the model of the first part of the book in the following directions:

- We include more general spaces of states and of actions.
- We consider new classes of policies.
- We allow for initial distributions over the state rather than fixed states.
- We consider new cost criteria.

The fact that infinite state and action spaces are involved requires more care on topological and measurability aspects, which makes this part more technical than the first part of the book.

6.1 The model

In this part, we shall deal with MDPs $\{\mathbf{X}, \mathsf{A}, \mathcal{P}, c, d\}$ where

- $\mathbf{X}$ is a countable state space.
- A is a metric set of actions. $\mathsf{A}(x) \subset \mathsf{A}$ is the compact set of actions available at state x, equipped with its Borel sets $\mathbf{A}(x)$; set $\mathcal{K} = \{(x, a) : x \in \mathbf{X}, a \in \mathsf{A}(x)\}$, equipped with its Borel σ-algebra $\mathbb{K}$ generated by the rectangles $(x, \mathcal{A})$, $(\mathcal{A} \subset \mathsf{A}(x))$.
- $\mathcal{P}$ are the transition probabilities; thus, $\mathcal{P}_{xay}$ is the probability of moving from state x to y if action a is chosen.
- $c : \mathcal{K} \to \mathbb{R}$ is an immediate cost. This cost will be related to a cost function which we shall minimize.
- $d : \mathcal{K} \to \mathbb{R}^K$ is a K-dimensional vector of immediate costs, related to K constraints. With some abuse of notation, we denote $c(x, \gamma) = \int c(x, a)\gamma(da)$ for any probability measure γ over $\mathsf{A}(x)$, with a similar definition for $d(x, \gamma)$.

We make throughout the second part of the book the assumption that, for every state x,

$$c(x, \cdot) \text{ and } d^k(x, \cdot), k = 1, \ldots, K, \text{ are continuous on } \mathsf{A}(x). \tag{6.1}$$

The transition probabilities are continuous on $\mathsf{A}(x)$, i.e., if
$a(n) \to a$, where $a, a(n) \in \mathsf{A}(x)$, then (6.2)
$\lim_{n\to\infty} \mathcal{P}_{xa(n)y} = \mathcal{P}_{xay}, \forall y \in \mathbf{X}$.

Assumption (6.1) can also be rephrased as: c and d are continuous over $\mathcal{K}$.

Define a history at time t to be a sequence of previous states and actions, as well as the current state: $h_t = (x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t)$. Let $\mathbf{H}_t$ be the set of all possible histories of length t equipped with its Borel σ-algebra. A policy u is a sequence $u = (u_1, u_2, \ldots)$ (containing n elements) where $u_t : \mathbf{H}_t \to M_1(\mathsf{A})$ is a measurable function that assigns to any history of length $t \leq n$, a probability measure over the set of actions, where $M_1(G)$ stands for the set of probability measures over a set G endowed with the topology of weak convergence of measures. If the history h_t is observed at time t, then the controller chooses an action within $\mathcal{A}$ with probability $u_t(\mathcal{A} \mid h_t)$, where $\mathcal{A}$ is any Borel subset of $\mathsf{A}(x_t)$. The class of all policies defined as above is denoted by U, and is called the *behavioral* policies.

We shall again use the policies U_M, U_S and U_D defined in Chapter 2.

It will often be useful to extend the definition of a policy $u = (u_1, u_2, \ldots)$ so as to allow u_t to depend not only on h_t, but also on some initial randomizing mechanism. More precisely, for any class of policies $G \subset U$, we define $\overline{M}(G)$ to be the class of mixed policies generated by G, and we call these *mixed-G policies*. A mixed-G policy is identified with a distribution q over G; the controller first uses q to choose some policy $u \in G$, and then proceeds with that policy from time 1 onwards. A policy as above that uses a distribution q is denoted by $\hat{q}$. Define $\mathcal{U} := \overline{M}(U_D)$. In the above definition we implicitly assume some measurable structure, i.e., that together with G there is given some σ-algebra $\mathcal{G}$ of sets in G, that include singletons (sets that contain a single policy), so that a probability on G is well defined. We shall sometimes include $\mathcal{G}$ in the notation, i.e., denote by $\overline{M}(G, \mathcal{G})$, the class of mixed-$G$ strategies with respect to $\mathcal{G}$, and identify them by all probability measures on $(G, \mathcal{G})$. We delay the discussion on constructing such σ-algebras to Section 6.3.

Any given distribution β for the initial state (at time 1) and a policy u define a unique probability measure, P^u_β, over the space of trajectories of the states and actions. This defines the stochastic processes X_t and A_t of the states and actions. The construction of the probability space for $u \in U$ is standard, see e.g., Hinderer (1970). We denote by E^u_β the corresponding expectation operator. P^u_β is then a measurable function on the space of policies, see Feinberg (1986).

For mixed policies, the construction is done similarly. Moreover, for any mixed policy, the probability distribution for the state and action processes is the same as the one obtained by some equivalent policy in U. This was established for the more general setting of MDPs with several controllers (stochastic games) by Kuhn (1953), Aumann (1964) and Bernhard (1992).

When β is concentrated on some state x, we shall use the notation P_x^u (and E_x^u) instead of P_β^u (and E_β^u, respectively), which we had used in the first part of the book. β is then the Dirac measure $\beta(\mathcal{X}) = \delta_x(\mathcal{X}) := 1\{x \in \mathcal{X}\}$, for any $\mathcal{X} \subset \mathbf{X}$.

6.2 Cost criteria

In addition to the finite horizon, the discounted and the average costs defined in Chapter 2, we shall consider in this part the total cost criterion.

For a given set $\mathcal{M} \subset \mathbf{X}$, define the hitting time of the set $\mathcal{M}$ as

$$T_{\mathcal{M}} \stackrel{\text{def}}{=} \min\{n > 1 : X_n \in \mathcal{M}\}. \tag{6.3}$$

We shall often omit $\mathcal{M}$ from the notation. We shall understand $T_{\mathcal{M}} = \infty$ for $\mathcal{M} = \emptyset$. Define $M_u(\beta) \stackrel{\text{def}}{=} E_\beta^u T_{\mathcal{M}}$.

Define $p_\beta^u(t; \mathcal{X}) := P_\beta^u(X_t \in \mathcal{X}, T > t)$ and, with some abuse of notation, $p_\beta^u(t; \overline{\mathcal{K}}) := P_\beta^u((X_t, A_t) \in \overline{\mathcal{K}}, T > t)$, for $\mathcal{X} \subset \mathbf{X}, \overline{\mathcal{K}} \subset \mathcal{K}$. We have for all $\beta \in M_1(\mathbf{X})$ and policies u,

$$p_\beta^u(t; x) = p_\beta^u(t; x, \mathsf{A}(x)),$$

and for $t > 1$,

$$p_\beta^u(t; x) = \sum_{y \in \mathbf{X}} \int_{\mathsf{A}(y)} p_\beta^u(t-1; y, da) \mathcal{P}_{yax} 1\{x \notin \mathcal{M}\}. \tag{6.4}$$

The total expected cost until set $\mathcal{M}$ is reached, is defined as

$$C_{tc}^n(\beta, u) = \sum_{t=1}^{n} E_\beta^u c(X_t, A_t) 1\{T > t\}, \qquad C_{tc}(\beta, u) = \overline{\lim_{n \to \infty}} C_{tc}^n(\beta, u). \tag{6.5}$$

(The finite horizon cost (2.1) is obtained as a special case of (6.5) by setting $\mathcal{M} = \emptyset$.)

We use the other costs defined in Chapter 2 but with a general initial distribution β. We thus denote by $C_\alpha(\beta, u)$, the total discounted cost, etc. We define similarly the costs related to the constraints.

Let $C(\beta, u)$ stand for any of the costs previously defined. For a fixed vector $V = (V_1, \ldots, V_K)$ of real numbers, we define the constrained control problem **COP** as:

Find a policy that minimizes $C(\beta, u)$ subject to $D(\beta, u) \leq V$.

We shall often use below the following notation. For a given set $\mathcal{M} \subset \mathbf{X}$ and a matrix $P : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$, we define the Taboo matrix ${}_{\mathcal{M}}P$ obtained by replacing columns $j \in \mathcal{M}$ by columns with zero entries.

6.3 Mixed policies and topologic structure*

We would like to have some framework in which one can make precise objects such as 'mixed strategies' (as defined in Section 6.1) and 'convergence of policies', as will be discussed in Chapter 13.

In order to well define mixed strategies, i.e., strategies of the form $\overline{M}(\overline{U})$ for some $\overline{U} \subset U$, we need to construct some σ-algebra of subsets of $\overline{U}$, that includes in particular all singletons (i.e., sets that contain a single policy). In order to define convergence of policies within some class $\overline{U}$, we need to define some topology on $\overline{U}$. In the sequel, we introduce a metric on some sets of policies, and then define a topology and a σ-algebra which are generated by the Borel sets.

For each x, let $\mathcal{B}(\mathsf{A}(x))$ denote the set of Borel subsets of $\mathsf{A}(x)$. $M_1(\mathsf{A}(x))$ is the space of probability measures over $\mathcal{B}(\mathsf{A}(x))$ endowed with the topology of weak convergence, and it is a linear Hausdorff compact metric space. (Since $\mathsf{A}(x)$ is compact metric, it is also separable, and hence the set of probability measures over $\mathcal{B}(\mathsf{A}(x))$ is tight. By Prohorov's Theorem, this implies the compactness of $M_1(\mathsf{A}(x))$.)

Assume first that the sets $\mathsf{A}(x)$ are finite for all x. Then, for any time t, the set of histories $\mathbf{H}_t$ is countable, so that the set $\mathbf{H} = \cup_t \mathbf{H}_t$ is countable. Let $x : \mathbf{H} \to \mathbf{X}$ be the projection that assigns $x(h_t) = y$ if $h_t = (x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t)$, and $x_t = y$. U can be identified with all functions which have the countable set $\mathbf{H}$ as range, and a countable product of compact sets $\prod_{h\in\mathbf{H}} M_1(\mathsf{A}(x(h)))$ as image. Tychonov's Theorem therefore implies that $\prod_{h\in\mathbf{H}} M_1(\mathsf{A}(x(h)))$ is also a convex, compact set in the topology of weak convergence and it is metrizable by virtue of Theorem 4.14 in Royden (1988). Moreover, it is easily seen that the extreme points of U are the pure policies, i.e., those which do not use any randomization at any time (in response to any history).

We thus obtained a metric topology for U. Moreover, we can now define the Borel sets $\mathcal{B}_U$ of U, and the σ-algebra $\mathcal{G}_U$ generated by them. They include in particular all singletons. The set of mixed strategies $\overline{M}(U, \mathcal{G}_U)$ is now identified with the set of probability measures on the space $(U, \mathcal{G}_U)$. This class of policies is *non-behavioral* and is called the class of mixed policies.

The above topology and σ-field $\mathcal{G}_U$ do not extend to the case when $\mathsf{A}(x)$ are not finite, since the sets $\mathbf{H}_t$ are then not countable. However, we may still obtain similar results for U_M, U_S and U_D.

Both U_S and U_M can be represented as the set of functions that have some countable set $\mathbf{I}$ as range, and a countable product of compact sets (of measures) $\prod_{i\in\mathbf{I}} M_1(\mathsf{A}_i)$ as image. The same considerations as above show that U_S and U_M are also convex, compact in the topology of weak convergence, are metrizable, and they have as extreme points the sets U_D, and the set of pure Markov policies, respectively. We thus have a metric

topology for U_D, U_S and U_M. We now define the Borel sets $\mathcal{B}_M$ of U_M, and the σ-algebra $\mathcal{G}_M$ generated by them. The set of mixed Markov strategies $\overline{M}(U_M, \mathcal{G}_M)$ is identified with the compact set of probability measures (compact in the topology of weak convergence) on the space $(U_M, \mathcal{G}_M)$. We define similarly $\overline{M}(U_S)$ and $\overline{M}(U_D) = \mathcal{U}$. (It follows that U_D is indeed a measurable set, so that $\mathcal{U}$ is well defined. The measurability can be established using the argument in Borkar, 1994, p. 178.)

Finally, for the class of policies U, one can consider the discrete σ-algebra $\mathcal{G}_U^D$ (which is generated by singletons), and define $\overline{M}(U, \mathcal{G}_U^D)$ with respect to that σ-algebra.

6.4 The dominance of Markov policies

The class of Markov policies turns out to be rich, in the following sense. For any policy in U, or in $\overline{M}(U_M)$, there exists an equivalent policy in U_M that induces the same marginal probability measure. This result was obtained by Derman and Strauch (1966) and extended by Hordijk (1977) (see also Derman, 1970, p. 21, Dynkin and Yushkevich, 1979, p. 17). We prove below a related result for the measures $p_\beta^u(t)$.

If $u \in U_M$, then $p_\beta^u(t; \cdot)$ can be written in the following vector notation:

$$p_\beta^u(t) = \beta_{\mathcal{M}} P(u_1)_{\mathcal{M}} P(u_2) \cdots {}_{\mathcal{M}} P(u_{t-1}), \tag{6.6}$$

where $p_\beta^u(t)$ and β are considered to be row vectors, and ${}_{\mathcal{M}} P(u_i)$ are matrices whose (x, y) entry is given by ${}_{\mathcal{M}} P_{xy}(u_i) = \int \mathcal{P}_{xay} 1\{y \notin \mathcal{M}\} u_t(da \mid x)$.

Theorem 6.1 (*Sufficiency of Markov policies*)
(*i*) *Choose any initial distribution β, and any $\gamma \in M_1(U_M)$. Let $\hat{\gamma}$ be the corresponding policy in $\overline{M}(U_M)$. Then there exists some $v \in U_M$ such that for all t,*

$$p_\beta^{\hat{\gamma}}(t; \cdot, \cdot) = p_\beta^v(t; \cdot, \cdot) \tag{6.7}$$

(*ii*) *Choose any initial distribution β, and a distribution γ over U with a discrete support, i.e., $\gamma \in M_1(U, \mathcal{G}_U^D)$. Let $\hat{\gamma}$ be the corresponding policy in $\overline{M}(U, \mathcal{G}_U^D)$. Then there exists some $v \in U_M$ such that for all t,* (6.7) *holds.*

Proof. The proof of (i) is based on Dynkin and Yushkevich (1979), Section 3.5. We write $p_\beta^{\hat{\gamma}}$ in an integral form:

$$p_\beta^{\hat{\gamma}}(t; \cdot, \cdot) = \int_{U_M} \gamma(du) P_\beta^u(X_t = \cdot, A_t \in \cdot, T > t).$$

Define v to be the Markov policy given by

$$v_t(\mathcal{A} \mid x) := \frac{\int \gamma(du) P_\beta^u(X_t = x, A_t \in \mathcal{A}, T > t)}{\int \gamma(du) P_\beta^u(X_t = x, T > t)} \tag{6.8}$$

for all integers t, states x and $\mathcal{A} \subset \mathsf{A}(x)$, for which the denominator is

non-zero. When it is zero, define $v_t(\cdot \mid x)$ to be an arbitrary probability measure over $\mathsf{A}(x)$. The proof follows by induction. (6.7) clearly holds for $t = 1$, since for any policy $u \in U_M$,

$$P_\beta^u(X_1 = x, A_1 \in \mathcal{A}, T > 1) = \beta(x)u_1(\mathcal{A} \mid x),$$

and $\int \gamma(du)P_\beta^u(X_1 = x, T > 1) = \beta(x)$; this implies

$$\begin{aligned} P_\beta^v(X_1 = x, A_1 \in \mathcal{A}, T > t) &= \beta(x)v_1(\mathcal{A} \mid x) \\ &= \int \gamma(du)P_\beta^u(X_1 = x, A_1 \in \mathcal{A}, T > 1). \end{aligned}$$

Assume that (6.7) holds for some t, i.e.,

$$\begin{aligned} &\int \gamma(du)P_\beta^u(X_t = x, A_t \in \mathcal{A}, T > t) = P_\beta^v(X_t = x, A_t \in \mathcal{A}, T > t) \\ &= [\beta_{\mathcal{M}} P(v_1)_{\mathcal{M}} P(v_2) \cdots {}_{\mathcal{M}} P(v_{t-1})]_x v_t(\mathcal{A} \mid x). \end{aligned} \tag{6.9}$$

We show first that

$$\int \gamma(du)P_\beta^u(X_{t+1} = x, T > t+1) = [\beta_{\mathcal{M}} P(v_1)_{\mathcal{M}} P(v_2) \cdots {}_{\mathcal{M}} P(v_t)]_x. \tag{6.10}$$

Since

$$P_\beta^u(X_{t+1} = x, T > t+1 \mid X_t = y, A_t = a, T > t) = \mathcal{P}_{yax} 1\{x \notin \mathcal{M}\},$$

P_β^u almost sure (a.s.) for all $u \in U_M$, we obtain by conditioning on X_t, A_t and by (6.9), that the left-hand side of (6.10) equals

$$\sum_{y \in \mathbf{X}} [\beta_{\mathcal{M}} P(v_1)_{\mathcal{M}} P(v_2) \cdots {}_{\mathcal{M}} P(v_{t-1})]_y \int_{\mathsf{A}(y)} \mathcal{P}_{yax} 1\{x \notin \mathcal{M}\} v_t(da \mid x).$$

This implies (6.10). Combining now (6.10) with (6.8), we get

$$\begin{aligned} &\int \gamma(du)P_\beta^u(X_{t+1} = x, A_{t+1} \in \mathcal{A}, T > t+1) \\ &= v_t(\mathcal{A} \mid x) \int \gamma(du)P_\beta^u(X_{t+1} = x, T > t+1) \\ &= [\beta_{\mathcal{M}} P(v_1)_{\mathcal{M}} P(v_2) \cdots {}_{\mathcal{M}} P(v_t)]_x v_t(\mathcal{A} \mid x) \\ &= P_\beta^v(X_{t+1} = x, A_{t+1} \in \mathcal{A}, T > t+1). \end{aligned}$$

This concludes the proof of (i). (ii) is obtained by the same arguments (see also Derman and Strauch, 1966, and Hordijk, 1977). □

Remark 6.1 (*The converse*)
An interesting question is whether some converse exists, i.e., whether we can describe any Markov policy (which uses randomizations) as a mixture of policies within some class of simpler policies. The answer is positive, and that class can be taken as the class of pure Markov policies (which

do not use randomizations). This was established (for $\mathcal{M} = \emptyset$) by Feinberg (1982, Theorem 1) and Kadelka (1983). The same question can be posed for an arbitrary class of policies (not necessarily Markov); this more general problem has been studied by Feinberg (1986, 1991).

We saw that for any policy $u \in U$ or $u \in \overline{M}(U_M)$, there exists some $v(u) \in U_M$ such that for all t,

$$p_\beta^{v(u)}(t;\cdot,\cdot) = p_\beta^u(t;\cdot,\cdot). \tag{6.11}$$

We call $v(u)$ the Markov policy corresponding to u.

Extending the definition of the first part of the book, we have the following for an arbitrary initial distribution β.

Definition 6.1 *(Dominating policies)*
A class of policies $\overline{U}$ is said to be a dominating class of policies for **COP** *for one of the cost criteria introduced in Section 2.1 or 6.1, and for a given initial distribution β, if for any policy $u \in U$ there exists a policy $\overline{u} \in \overline{U}$ such that*

$$C(\beta,\overline{u}) \le C(\beta,u), \qquad \text{and} \qquad D(\beta,\overline{u}) \le D(\beta,u). \tag{6.12}$$

In the above definition, C, D, and **COP** stand for any one of the cost criteria previously defined.

We conclude the following:

Theorem 6.2 (*Dominance of Markov policies*)
The Markov policies are dominating for any cost criterion which is a function of the marginal distribution of states and actions or of the measures $p_\beta^u(t;\cdot,\cdot)$.

6.5 Aggregation of states*

We consider in this section MDPs that can be decomposed into classes of states, such that the transition probabilities between classes, as well as the immediate costs depend on the states only through the class to which the states belong (this will be made precise below). We show that Markov policies that depend only on the current class (rather than on the current state) are dominant.

This property will have several applications. In Section 6.6, we show that it implies that adding a further independent randomization mechanism at each time slot, in the definition of policies, does not allow us to obtain better performances. Another application of the results here will be given in Section 7.3 where an MDP with unbounded cost is shown to be equivalent to one with bounded cost (and in some cases, to a problem of minimizing the total expected time to reach some set, rather than the total expected cost).

The model

Consider a countable state space $\overline{\mathbf{X}}$ and assume that each state $\overline{x} \in \overline{\mathbf{X}}$ has two components: $\overline{x} = (x, i)$, where x belongs to a countable set $\mathbf{X}$, and each i belongs to a countable set $\mathbf{I}(x)$ (that may depend on x). We say that state $\overline{x}$ belongs to an equivalence class $g(\overline{x}) = x$. We shall denote by (X_t, I_t) the stochastic process describing the state. The partition into equivalent classes is of interest when the MDP is decomposable, as defined in the following.

Definition 6.2 *(Decomposable MDPs)*
The MDP $(\overline{X}, \overline{\mathsf{A}}, \overline{\mathcal{P}}, \overline{c}, \overline{d})$ is said to be decomposable if the following holds. For any two states in the same class, i.e., for any $\overline{x}_1, \overline{x}_2$ that satisfy $g(\overline{x}_1) = g(\overline{x}_2)$, we have

- $\overline{\mathsf{A}}(\overline{x}_1) = \overline{\mathsf{A}}(\overline{x}_2) =: \mathsf{A}(g(\overline{x}_1))$,
- $\overline{\mathcal{P}}_{\overline{x}_1,a,(y,\mathbf{I}(y))} = \overline{\mathcal{P}}_{\overline{x}_2,a,(y,\mathbf{I}(y))}$, $\forall y \in \mathbf{X}, a \in \overline{\mathsf{A}}(\overline{x}_1)$,
- $\overline{c}(\overline{x}_1, a) = \overline{c}(\overline{x}_2, a)$, $\overline{d}(\overline{x}_1, a) = \overline{d}(\overline{x}_2, a)$, $\forall a \in \overline{\mathsf{A}}(\overline{x}_1)$.
- *If the total cost is used, then the set $\overline{\mathcal{M}}$ is of the form: $\overline{\mathcal{M}} = \{\overline{x} = (x, i), x \in \mathcal{M}, i \in \mathbf{I}(x)\}$.*

We then say that the MDP can be aggregated to the MDP $(\mathbf{X}, \mathsf{A}, \mathcal{P}, c, d)$ where

$$\mathsf{A}(x) = \overline{\mathsf{A}}(\overline{x}), \quad \mathcal{P}_{xay} = \overline{\mathcal{P}}_{\overline{x},a,(y,\mathbf{I}(y))}, \quad c(x,a) = \overline{c}(\overline{x}, a), \quad d(x,a) = \overline{d}(\overline{x}, a);$$

note that all these quantities depend on $\overline{x}$ only through $g(\overline{x})$.

Definition 6.3 *(Simple policies)*
Consider the decomposable MDP $(\overline{X}, \overline{\mathsf{A}}, \overline{\mathcal{P}}, \overline{c}, \overline{d})$. A policy u is said to be simple if for all $t \in \mathbb{N}$, $u_t(\cdot \mid (x_1, i_1), a_1, \ldots, x_t, i_t))$ do not depend on $i_1, \ldots, i_t$.

The dominance of simple Markov policies

Theorem 6.3 (*Sufficiency of simple Markov policies*)
Consider the decomposable MDP $(\overline{X}, \overline{\mathsf{A}}, \overline{\mathcal{P}}, \overline{c}, \overline{d})$. Choose any initial distribution β, and any $u \in U_M$. Then there exists some simple Markov policy v such that for all t,
(i) the following holds:

$$P^v_\beta(X_t = x, A_t \in \mathcal{A}) = P^u_\beta(X_t = x, A_t \in \mathcal{A}), \tag{6.13}$$

$\forall \overline{x} = (x, i) \in \overline{\mathbf{X}}, \mathcal{A} \subset \overline{\mathsf{A}}(\overline{x})$, where $X_t = g(\overline{X}_t)$ is the first component of the state $\overline{X}_t$ at time t;
(ii) For any set $\mathcal{M} \subset \mathbf{X}$ and for $T = \min\{t : g(\overline{X}_t) \in \mathcal{M}\}$,

$$P^v_\beta(X_t = x, A_t \in \mathcal{A}, T > t) = P^u_\beta(X_t = x, A_t \in \mathcal{A}, T > t), \tag{6.14}$$

$\forall \overline{x} = (x, i) \in \overline{\mathbf{X}}, \mathcal{A} \subset \overline{\mathsf{A}}(\overline{x})$.

Proof. (i) Choose an arbitrary Markov policy u and define v as follows. For any $\overline{x} = (x, i) \in \overline{\mathbf{X}}$ and any $\mathcal{A} \subset \overline{\mathsf{A}}(\overline{x})$, we set

$$v_t(\mathcal{A} \mid \overline{x}) \stackrel{\text{def}}{=} P^u_\beta(X_t = x, A_t \in \mathcal{A} \mid X_t = x)$$

i.e.,

$$v_t(\mathcal{A} \mid \overline{x}) = \frac{P^u_\beta(X_t = x, A_t \in \mathcal{A})}{P^u_\beta(X_t = x)}$$

for x for which $P^u_\beta(X_t = x) > 0$, and otherwise v_t is an arbitrary probability distribution over $\overline{\mathsf{A}}(\overline{x})$. The proof is similar to the proof of Theorem 6.1, and proceeds by induction on t. For $t = 1$ we have for any $x \in \mathbf{X}$ and any Borel set $\mathcal{A} \subset \mathsf{A}(x)$,

$$\begin{aligned} P^u_\beta(X_t = x, A_t \in \mathcal{A}) &= \sum_{i \in \mathbf{I}(x)} \beta(x,i) \frac{P^u_\beta(X_1 = x, A_1 \in \mathcal{A})}{\sum_{i \in \mathbf{I}(x)} \beta(x,i)} \\ &= \sum_{i \in \mathbf{I}(x)} \beta(x,i) v_1(\mathcal{A} \mid (x,i)) \\ &= P^v_\beta(X_t = x, A_t \in \mathcal{A}). \end{aligned}$$

Next, we assume that (6.13) holds for some t, and show that it holds for $t+1$. We first show that

$$P^v_\beta(X_{t+1} = x) = P^u_\beta(X_{t+1} = x), \qquad \forall x \in \mathbf{X}. \tag{6.15}$$

Indeed,

$$\begin{aligned} P^v_\beta(X_{t+1} = x) &= \sum_{\overline{y} \in \overline{X}} \int_{\overline{\mathsf{A}}(\overline{y})} P^v_\beta(\overline{X}_t = \overline{y}, A_t = a) \overline{\mathcal{P}}_{\overline{y},da,(x,\mathbf{I}(x))} \\ &= \sum_{y \in X} \int_{\mathsf{A}(y)} P^v_\beta(X_t = y, A_t = a) \mathcal{P}_{y,da,x} \\ &= \sum_{y \in X} \int_{\mathsf{A}(y)} P^u_\beta(X_t = y, A_t = a) \mathcal{P}_{y,da,x} \\ &= \sum_{\overline{y} \in \overline{X}} \int_{\overline{\mathsf{A}}(\overline{y})} P^u_\beta(\overline{X}_t = \overline{y}, A_t = a) \overline{\mathcal{P}}_{\overline{y},da,(x,\mathbf{I}(x))} \\ &= P^u_\beta(X_{t+1} = x). \end{aligned}$$

Finally,

$$P^v_\beta(X_{t+1} = x, A_{t+1} \in \mathcal{A})$$

$$
\begin{aligned}
&= \sum_{\overline{x}\in\mathbf{I}(x)} P^v_\beta(A_{t+1}\in\mathcal{A}\mid \overline{X}_{t+1}=\overline{x})P^v_\beta(\overline{X}_{t+1}=\overline{x})\\
&= \sum_{\overline{x}\in\mathbf{I}(x)} v_{t+1}(\mathcal{A}\mid \overline{x})P^v_\beta(\overline{X}_{t+1}=\overline{x})\\
&= \sum_{\overline{x}\in\mathbf{I}(x)} P^u_\beta(A_{t+1}\in\mathcal{A}\mid X_{t+1}=x)P^v_\beta(\overline{X}_{t+1}=\overline{x})\\
&= P^u_\beta(A_{t+1}\in\mathcal{A}\mid X_{t+1}=x)\sum_{\overline{x}\in\mathbf{I}(x)} P^v_\beta(\overline{X}_{t+1}=\overline{x})\\
&= P^u_\beta(X_{t+1}=x, A_{t+1}\in\mathcal{A})
\end{aligned}
$$

This establishes (i). (ii) is obtained similarly. □

From Theorem 6.3 we conclude that:

Corollary 6.1 (*Dominance of simple policies*)
When an MDP is decomposable, then simple Markov policies are dominant for any one of the cost criterion introduced in Sections 2.1 *and* 6.1.

Remark 6.2 (*Equivalence between the MDPs*)
The decomposable MDP $(\overline{X},\overline{\mathsf{A}},\overline{\mathcal{P}},\overline{c},\overline{d})$ is equivalent to the aggregated one $(\mathbf{X},\mathsf{A},\mathcal{P},c,d)$. Indeed, there is an obvious one to one correspondence between simple Markov policies in the decomposable MDP, and general Markov policies in the aggregated MDP. It follows from Theorem 6.3 that they both generate the same distribution of (X_t, A_t) for each t for the two MDPs, and hence, the same distribution over the costs. In fact, the natural correspondence exists between arbitrary simple policies in the decomposable MDP, and arbitrary policies in the aggregated MDP. They generate the same distributions of the whole stochastic processes $\{X_t, A_t\}_t$.

6.6 Extra randomization in the policies*

As an application of the aggregation of MDPs, we analyze a new class of policies which we denote by U_R, that extend the behavioral ones (in Section 6.1) for the (original) MDP $\{\mathbf{X},\mathsf{A},\mathcal{P},c,d\}$ defined in Section 6.1.

We consider policies $u=(u_1,u_2,\ldots)$, where at each time unit t we allow the use of an *extra randomization mechanism*: we enhance the history h_t to include for each t not only all the past states and actions as well as the present state and action, but also the outcome of some independent random trials $\{\mathcal{I}_t\}$ taking values in a countable space $\mathbf{I}$. The distribution of $\mathcal{I}_t$ may depend on the current state X_t as well as the current action A_t; we denote the distribution of $\mathcal{I}_t$ at state x under action a by $q_{xa}(\cdot)$.

When using policies within U_R in the original MDP, this will be shown to be equivalent to an alternative MDP $\{\overline{\mathbf{X}},\overline{\mathsf{A}},\overline{\mathcal{P}},\overline{c},\overline{d}\}$ with a larger state space, but with policies restricted to the behavioral ones (as defined in

Section 6.1), i.e., to ones that depend only on the past state and action trajectories as well as the current state. In the new mechanism, the extra randomization mechanism will appear in the states, instead of appearing in the policies.

We now introduce the new MDP with extra randomization appearing in the state:

- The state space: $\overline{\mathbf{X}} = \mathbf{X} \times \mathbf{I}$;
- The action space: $\overline{\mathsf{A}}(x,i) = \mathsf{A}(x)$;
- The transition probabilities:

$$\overline{\mathcal{P}}_{(x,i),a,(y,J)} = \mathcal{P}_{xay} q_{xa}(J), \qquad x,j \in \mathbf{X}, i \in \mathbf{I}, J \subset \mathbf{I};$$

- The costs:

$$\overline{c}((x,i),a) = c(x,a), \qquad \overline{d}((x,i),a) = d(x,a).$$

- The policies for the new model are defined as in Section 6.1.

A one to one correspondence between policies in the two MDPs would be:

$$u = r(\overline{u}), \qquad \forall \text{ behavioral policies } \overline{u} \text{ in the new MDP} \tag{6.16}$$

where r is given by:

$$\begin{aligned} u_t(\mathcal{A} \mid x_1, i_1, a_1, \ldots, x_t, i_t) &= [r(\overline{u})]_t(\mathcal{A} \mid x_1, i_1, a_1, \ldots, x_t, i_t) \\ &= \overline{u}_t(\mathcal{A} \mid (x_1, i_1), a_1, \ldots, (x_t, i_t)). \end{aligned}$$

u, defined for the original MDP satisfies $u \in U_R$. It is easy to check that the two MDPs induce the same probability distribution over the trajectories $h_t = (x_1, i_1, a_1, \ldots, x_t, i_t)$ when the corresponding policies are used.

The new MDP is decomposable and the equivalent aggregated MDP is simply the original MDP. Simple policies (as defined in Definition 6.3) in the new MDP correspond to *behavioral policies* in the original MDP.

Since Corollary 6.1 and Remark 6.2 clearly apply to the new decomposable MDP, we may conclude by applying the correspondence with policies in the original MDP that the behavioral policies (and thus, in particular, the Markov policies) are dominating *in the original MDP*. Thus, adding extra randomization does not improve the performance.

Remark 6.3 (*Related results*)
A related model has been analyzed in Chapter 7 of Krass (1989). One can show that the behavioral policies are equivalent to non-behavioral ones (including mixed policies, and the policies with extra randomization) not only through the marginal probabilities of the state and actions, but in the actual probability distribution over the whole random processes of states and actions. The proof for the case of mixed policies can be found in Dynkin and Yushkevich (1979). (Similar results in the case of several controllers can be found in Kuhn, 1953, Aumann, 1964, and Bernhard, 1992.)

6.7 Equivalent quasi-Markov model and quasi-Markov policies*

The way we viewed our MDP $\{\mathbf{X}, \mathsf{A}, \mathcal{P}, c, d\}$ until now was as a discrete time model: transitions occur at every integer point. We note, however, that if $\mathcal{P}_{xax} > 0$ for some x and a, then when the state x is reached, it may remain at x for some random time greater than 1, if action a is used. It is therefore natural to investigate the state trajectories not only in the original discrete setting, but also at some embedded times.

One possible choice of these embedded times is simply the time instants at which state transition occur, i.e., time t at which $X_t \neq X_{t-1}$.

In cases where we use policies within U_R (defined in Section 6.6), we may consider, in addition, other embedded times that we now describe. Assume that at each time instant, the policies in U_R use a coin with two values: $\mathbf{I} = \{0, 1\}$. If the state at time t is x and the action used is a, then the outcome $\mathcal{I}_t$ equals 1 with probability

$$q(x, a) \stackrel{\text{def}}{=} q_{xa}(1),$$

and it equals 0 w.p. $1 - q(x, a)$. $q(x, a)$ will be some arbitrary fixed parameters that define U_R. We shall say that a *virtual* transition occurs whenever $\mathcal{I}_t = 1$.

Define $\tau(0) = 0$, and

$$\tau(n+1) := \inf\{n > \tau(n) : I_n = 1 \text{ or } X_n \neq X_{n-1}\},\ n \in \mathbb{N} \tag{6.17}$$

(with the convention that $\inf\{\emptyset\} = \infty$). $\tau(n)$ denotes the instant at which the nth transition (real or virtual) occurs.

Define also for $n \in \mathbb{N}$

$$\eta(n) := \max\{m \geq 0 : n \geq \tau(m)\}. \tag{6.18}$$

$\eta(n)$ denotes the number of transitions that occurred by time n.

Definition 6.4 *(Quasi-Markov policies)*
A policy $u \in U_R$ is said to be quasi-Markov if $u_t(\overline{h}_t)$ is only a function of the state at time t and of $\eta(t)$:

$$u_t(\overline{h}_t) = \hat{u}_{\eta(t)}(x_t)$$

for all $t \in \mathbb{N}$ and all histories $\overline{h}_t = (x_1, i_1, a_1, \ldots, x_t)$. $\hat{u} = \{\hat{u}_n(\cdot)\}_{n \in \mathbb{N}}$ is said to be a corresponding Markov policy.

Note that $X_t = X_{\tau_n}$ for all $t \in [\tau_n, \tau_{n+1})$. Thus the quasi-Markov policy uses at time t the state at the last time that a transition occurred (not later than t), as well as the number of transitions that have occurred. Under a quasi-Markov policy, our MDP can be viewed as sampled at transition times $\tau(n)$, and in between transitions the states and the probabilities used for selecting actions are unchanged.

Definition 6.5 *(Quasi-Markov cost)*
We say that a cost criterion $C(\beta,\cdot)$ has a quasi-Markov structure if for all $u \in U_R$ it depends on the distribution of the state and action trajectories only through the following quantities:

$$\mathsf{p}^u_\beta(n;x,\mathcal{A}) \stackrel{\text{def}}{=} E^u_\beta \sum_{t=\tau(n)}^{\tau(n+1)-1} 1\{X_t = x, A_t \subset \mathcal{A}, t > T\}$$

(*where T is the time that the process X_t hits some given set of states $\mathcal{M} \subset \mathbf{X}$*).

Note that $\mathsf{p}^u_\beta(n;\mathcal{K})$ is simply the expected duration of the nth transition that occurs before hitting $\mathcal{M}$.

The total cost criterion is quasi-Markov whenever it can be expressed as

$$C_{tc}(\beta,u) = \overline{\lim_{m\to\infty}} \sum_{n=1}^{m} \langle \mathsf{p}^u_\beta(n), c\rangle. \tag{6.19}$$

In particular, (6.19) holds for the case of non-negative immediate costs, as will be shown later.

Theorem 6.4 (*Sufficiency of quasi-Markov policies*)
Fix an arbitrary policy $u \in U_R$. Then there exists a quasi-Markov policy v achieving the same values of $\{\mathsf{p}^u_\beta(n)\}_n$.

Proof. Choose an arbitrary $u \in U_R$. Define

$$\hat{v}_n(\mathcal{A} \mid x) \stackrel{\text{def}}{=} \frac{\mathsf{p}^u_\beta(n;x,\mathcal{A})}{\mathsf{p}^u_\beta(n;x)}, \qquad x \in \mathbf{X}, \mathcal{A} \in \mathsf{A}(x)$$

and the quasi-Markov policy v by

$$v_t(\mathcal{A} \mid \overline{h}_t) \stackrel{\text{def}}{=} \hat{v}_{\eta(t)}(\mathcal{A} \mid \overline{x}_{\eta(t)}).$$

Step i:
Assume that the state distribution at time $\tau(n)$ is the same under u and v:

$$P^u(X_{\tau(n)} \in \mathcal{X}, T > \tau(n)) = P^v(X_{\tau(n)} \in \mathcal{X}, T > \tau(n)), \forall \mathcal{X} \in \mathbf{X}.$$

We show that this implies that

$$\mathsf{p}^u_\beta(n;x,\mathcal{A}) = \mathsf{p}^v_\beta(n;x,\mathcal{A}), \qquad \forall x \in \mathbf{X}, \mathcal{A} \subset \mathsf{A}(x).$$

For every integer $m \geq 1$ and every $x \in \mathbf{X}$,

$$\begin{aligned} &P^u_\beta\big(X_{\tau(n)+m} = x, \tau(n+1) > \tau(n)+m, T > \tau(n)\big) \\ &= \int_{\mathsf{A}(x)} P^u_\beta\big(X_{\tau(n)+m-1} = x, A_{\tau(n)+m-1} \in da, \\ &\qquad \tau(n+1) > \tau(n)+m-1, T > \tau(n)\big)\mathcal{P}_{xax}(1-q(x,a)). \end{aligned}$$

Hence

$$\begin{aligned}
\mathsf{p}^u_\beta(n;x) &= \sum_{m=0}^{\infty} P^u_\beta\big(X_{\tau(n)+m}=x, \tau(n+1)>\tau(n)+m, T>\tau(n)\big)\\
&= P^u_\beta(X_{\tau(n)}=x, T>\tau(n))\\
&\quad + \sum_{m=1}^{\infty} P^u_\beta\big(X_{\tau(n)+m}=x, \tau(n+1)>\tau(n)+m, T>\tau(n)\big)\\
&= P^u_\beta(X_{\tau(n)}=x, T>\tau(n)) +\\
&\quad \sum_{m=0}^{\infty}\int_{\mathsf{A}(x)} P^u_\beta\big(X_{\tau(n)+m-1}=x, A_{\tau(n)+m-1}\in da,\\
&\qquad\qquad \tau(n+1)>\tau(n)+m-1, T>\tau(n)\big)\mathcal{P}_{xax}(1-q(x,a))\\
&= P^u_\beta(X_{\tau(n)}=x, T>\tau(n)) + \int_{\mathsf{A}(x)} \mathsf{p}^u_\beta(n;x,da)\mathcal{P}_{xax}(1-q(x,a))\\
&= P^u_\beta(X_{\tau(n)}=x, T>\tau(n)) + \mathsf{p}^u_\beta(n;x)\int_{\mathsf{A}(x)} \mathcal{P}_{xax}(1-q(x,a))\frac{\mathsf{p}^u_\beta(n;x,da)}{\mathsf{p}^u_\beta(n;x)}\\
&= P^u_\beta(X_{\tau(n)}=x, T>\tau(n)) + \mathsf{p}^u_\beta(n;x)\int_{\mathsf{A}(x)} \mathcal{P}_{xax}(1-q(x,a))\hat{v}_n(da\mid x).
\end{aligned}$$

Thus,

$$\mathsf{p}^u_\beta(n;x) = \frac{P^u_\beta\big(X_{\tau(n)}=x, T>\tau(n)\big)}{1-\int_{\mathsf{A}(x)}\mathcal{P}_{xax}(1-q(x,a))\hat{v}_n(da\mid x)}.$$

Repeating the above for v, and recalling the assumption that $P^u_\beta(X_{\tau(n)}=x, T>\tau(n)) = P^v_\beta(X_{\tau(n)}=x, T>\tau(n))$, we conclude that indeed $\mathsf{p}^u_\beta(n;x) = \mathsf{p}^v_\beta(n;x)$. Hence

$$\mathsf{p}^u_\beta(n;x,\mathcal{A}) = \mathsf{p}^u_\beta(n;x)\hat{v}_n(\mathcal{A}\mid x) = \mathsf{p}^v_\beta(n;x)\hat{v}_n(\mathcal{A}\mid x) = \mathsf{p}^v_\beta(n;x,\mathcal{A}).$$

Step ii:

We show that $P^u_\beta(X_{\tau(n)}=x, T>\tau(n)) = P^v_\beta(X_{\tau(n)}=x, T>\tau(n))$.

Denote

$$\hat{\mathcal{P}}_{yax} = \begin{cases} \mathcal{P}_{yax}1\{x\notin\mathcal{M}\} & \text{if } y\neq x,\\ \mathcal{P}_{xax}q(x,a)1\{x\notin\mathcal{M}\} & \text{otherwise .}\end{cases}$$

Then

$$\begin{aligned}
&P^u_\beta(X_{\tau(n)}=x, T>\tau(n-1))\\
&\quad= \sum_{y\in\mathbf{X}}\int_{\mathsf{A}(y)}\sum_{m=0}^{\infty} P^u_\beta\big(X_{\tau(n-1)+m}=y, A_{\tau(n-1)+m}\in da,\\
&\qquad\qquad \tau(n)>\tau(n-1)+m, T>\tau(n-1)\big)\hat{\mathcal{P}}_{yax}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{y\in\mathbf{X}} \int_{\mathsf{A}(y)} \mathsf{p}^u_\beta(n-1;y,da)\hat{\mathcal{P}}_{yax} \\
&= \sum_{y\in\mathbf{X}} \int_{\mathsf{A}(y)} \mathsf{p}^v_\beta(n-1;y,da)\hat{\mathcal{P}}_{yax} \\
&= P^v_\beta(X_{\tau(n)} = x, T > \tau(n-1)).
\end{aligned}
$$

The proof of the theorem now follows by an inductive argument. □

We have thus shown that policy $u \in U_R$ can be replaced by a quasi-Markov policy v having the same p_β. This is also true if $u \in U$, since U can be identified with a subset of U_R, where policies do not make use of the extra randomization. We thus conclude:

Theorem 6.5 (*Dominance of quasi-Markov policies*)
Quasi-Markov policies are dominant for MDPs with quasi-Markov cost.

The following property of quasi-Markov policies is easily checked. Given the state at the nth transition epoch, the duration of the next transition and the state at the beginning of the next transition are independent. More precisely, let u be a quasi-Markov policy that uses a conditional measure $\hat{u}_n(\cdot \mid x)$ during the random interval $[\tau(n), \tau(n+1))$, if the state at time $\tau(n)$ is x. Then for any integer $t > 0$,

$$P^u_\beta(X_{\tau(n+1)} = y, \tau(n+1) - \tau(n) = t \mid X_{\tau(n)} = x) = \mathcal{P}_{x,y}R(x,t)$$

$P^u_\beta - a.s.$, where

$$R(x,t) \stackrel{\text{def}}{=} q(x)(1-q(x))^{t-1},$$

where $q(x) = \int_{\mathsf{A}(x)} q(x,a)\hat{u}_n(da \mid x)$ and where $n = \eta(t)$. This property follows from the fact that

$$
\begin{aligned}
&P^u(X_{\tau(n)+t} = y, X_{\tau(n)} = x, \tau(n+1) - \tau(n) = t) \\
&\quad = P^u\big(X_{\tau(n)} = x, \mathcal{I}_{\tau(n)+1} = \ldots = \mathcal{I}_{\tau(n)+t-1} = 0, \mathcal{I}_{\tau(n+1)} = 1), \\
&\qquad\qquad X_{\tau(n+1)} = y\big),
\end{aligned}
$$

and the fact that $\mathcal{I}_t$ has a Bernoulli distribution with parameter $q(x)$ at $t \in [\tau(n), \tau(n+1))$, if $X_{\tau(n)} = x$.

Remark 6.4 (*Y-embedded policies*)
The idea of introducing policies that depend on the number of visits in some states goes back to Feinberg (1986), who defined Y-embedded policies. For a fixed subset Y of the state space, we call a policy Y-embedded if all decisions are functions of the following three factors: (a) the current state x, (b) the number m of visits to Y (which is the total time spent in Y up to the current epoch), and (c) the time k passed after the last visit to Y. Given any subset $Y \subset \mathbf{X}$ and initial distribution, it is possible to construct for any policy u a corresponding randomized Y-embedded policy which selects

actions for any x, m and k with the same conditional probabilities as the ones that the original policy would choose. Feinberg (1968, Theorem 4.1) showed that for each $m, k = 0, 1, \ldots$, the distributions of state-actions are the same at each epoch: k units of time after the mth visit to Y, we have the same distribution under both policies. Note that for $Y = \mathbf{X}$, the set of Y-embedded policies is exactly the set of Markov policies. Feinberg's results imply that Y-embedded policies are dominant for the total cost criterion.

CHAPTER 7

The total cost: classification of MDPs

7.1 Transient and Absorbing MDPs

We begin by introducing transient and absorbing policies, and transient and absorbing MDPs. These definitions are due to Hordijk (1977, p. 60).

Definition 7.1 (*Transient and absorbing policies*)
Fix an initial distribution β. Consider a partition of $\mathbf{X}$ into two disjoint sets $\mathbf{X}'$ and $\mathcal{M}$. A policy u is said to be $\mathbf{X}'$-transient if

$$\sum_{t=1}^{\infty} P_{\beta}^{u}(X_t = x, T > t) < \infty \text{ for any } x \in \mathbf{X}. \tag{7.1}$$

It is called $\mathbf{X}'$-absorbing, or absorbing to $\mathcal{M}$, if

$$E_{\beta}^{u} T < \infty. \tag{7.2}$$

Condition (7.1) means that the expected time we spend (under policy u) in any state $x \in \mathbf{X}'$ is finite; condition (7.2) means that the expected life-time of the whole set $\mathbf{X}'$ is finite.

Definition 7.2 (*Transient and absorbing MDPs*)
Fix an initial distribution β. Consider a partition of $\mathbf{X}$ into two disjoint sets $\mathbf{X}'$ and $\mathcal{M}$. The MDP is called an $\mathbf{X}'$-transient MDP (absorbing to $\mathcal{M}$, respectively) if all policies are $\mathbf{X}'$-transient (absorbing to $\mathcal{M}$, respectively). An $\mathbf{X}$-transient MDP is called a transient MDP.

When using $\mathbf{X}'$-transient MDPs with the total cost criterion, we shall assume that the immediate costs are non-negative. We shall relax this condition for MDPs with uniform Lyapunov functions and contracting MDPs, defined in the following sections.

Sufficient conditions for determining that an MDP is $\mathbf{X}'$-transient or absorbing to $\mathcal{M}$ will be presented in Lemma 7.2, Remark 7.4, Remark 7.6 and Corollary 8.1.

Here are some properties of transient stationary policies.

Lemma 7.1 (*Stationary policies in transient and absorbing MDPs*)
Fix some initial distribution β on $\mathbf{X}$ and a stationary $\mathbf{X}'$-transient policy w. Then
(i)

$$f^*(x) := \sum_{t=1}^{\infty} p_{\beta}^{w}(t; x)$$

is the minimal solution to

$$f = \beta + f\,{}_{\mathcal{M}}P(w), \qquad f \geq 0 \tag{7.3}$$

(*where f and β are row vectors on $\mathbf{X}$, and ${}_{\mathcal{M}}P(w)$ is the transition probability matrix of the Markov chain corresponding to the stationary policy w, in which we replace columns corresponding to states in $\mathcal{M}$ by zeros*). *It is the unique solution to (7.3) among those solutions f that satisfy* $\lim_{n\to\infty} f[{}_{\mathcal{M}}P]^n(w) = 0$.
(*ii*) *Assume that the MDP is absorbing to $\mathcal{M}$. Fix some policy u and define $g^*(x) := \sum_{t=1}^{\infty} p_\beta^u(t;x)$. If g^* satisfies (7.3), then $g^*(x) = f^*(x)$ for all $x \in \mathbf{X}$.*

Proof. (i) It follows easily that f^* is indeed a solution of (7.3). Iterating (7.3), we get for all integers n:

$$\begin{aligned} f &= \beta + (\beta + f\,{}_{\mathcal{M}}P(w))\,{}_{\mathcal{M}}P(w) = \beta + \beta\,{}_{\mathcal{M}}P(w) + f\,{}_{\mathcal{M}}P^2(w) \\ &= \sum_{i=1}^{n-1} \beta\,{}_{\mathcal{M}}P^i(w) + f\,{}_{\mathcal{M}}P^n(w) = \sum_{t=1}^{n-1} p_\beta^w(t) + f\,{}_{\mathcal{M}}P^n(w). \end{aligned} \tag{7.4}$$

(i) follows since the above holds for all n and since $f \geq 0$.
(ii) follows from (i) since $g^*P^n(w)$ converges to zero. Indeed, define $\mathbf{1} : \mathbf{X} \to \mathbb{R}$ to be the function whose entries are all 1. Since w is absorbing to $\mathcal{M}$, $\langle g^*, \mathbf{1}\rangle < \infty$. For any integer n and $y \in \mathbf{X}$, the yth column of ${}_{\mathcal{M}}P^n(w)$ is bounded by $\mathbf{1}$, so by the the generalized dominance convergence theorem (Royden, 1988, Proposition 11.18),

$$\lim_{n\to\infty} g^*P^n(w) = g^*\left(\lim_{n\to\infty} P^n(w)\right) = 0. \tag{7.5}$$

To get the last equality, it suffices to show the following: let y be some state for which $g^*(y) > 0$. Then the yth row of the matrix ${}_{\mathcal{M}}P^\infty \stackrel{\text{def}}{=} \overline{\lim}_{n\to\infty}[{}_{\mathcal{M}}P]^n(w)$ is zero. Assume that for some z, ${}_{\mathcal{M}}P^\infty_{yz} \neq 0$. There exists some time t for which $p_\beta^u(t;y) > 0$. Consider a policy v that behaves like policy u until time t and then behaves like the stationary policy w. Then,

$$\sum_{s=1}^{\infty} p_\beta^v(s;z) \geq p_\beta^u(t,y) \sum_{n=0}^{\infty} [({}_{\mathcal{M}}P)^n(w)]_{yz} = \infty.$$

This contradicts the fact that the MDP is absorbing. This shows that ${}_{\mathcal{M}}P^\infty_{yz} = 0$ for all z, from which (7.5) follows. □

Definition 7.3 (*Unichain and communicating MDPs*)
An MDP is said to be unichain if under every $u \in U_D$, the state process is an ergodic Markov chain, i.e., all states communicate.
An MDP is said to be communicating (see Bather, 1973) if for any two states $x, y \in \mathbf{X}$, there exists a policy $u \in U_D$ (that may depend on x and y) such that y is reached from x with positive probability.

7.2 MDPs with uniform Lyapunov functions

We define in this section the framework of uniform Lyapunov functions for MDPs, due originally to Hordijk (1977) and further investigated and used by many authors (see Van Der Wal, 1981a, Cavazos-Cadena and Hernández-Lerma, 1992, Arapostathis *et al.*, 1993, and references therein).

We first present the definition of Cavazos-Cadena and Hernández-Lerma (1992) related to the total expected life-time. The results in that reference are useful for the total cost problem, provided that the costs are bounded.

Definition 7.4 (*Uniform Lyapunov function for total expected life-time*) Consider a partition of $\mathbf{X}$ to the disjoint sets $\mathbf{X}'$ and $\mathcal{M}$. A function $\mu : \mathbf{X} \to [1, \infty)$ is said to be a uniform Lyapunov function for the total expected life-time if (i)-(iii) below hold:
(i) For all $(x, a) \in \mathcal{K}$,

$$1 + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \mu(y) \leq \mu(x).$$

(ii) For each $x \in \mathbf{X}$, the mapping $a \to \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \mu(y)$ is continuous in $\mathsf{A}(x)$ (i.e., if $a_n \to a$, then $\sum_{y \in \mathbf{X}'} \mathcal{P}_{x,a_n,y} \mu(y)$ converges to $\sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \mu(y)$).
(iii) For each $x \in \mathbf{X}$ and $u \in U_D$,

$$\lim_{t \to \infty} E_x^u [\mu(X_t) 1\{T_{\mathcal{M}} > t\}] = 0.$$

Lemma 7.2 (*Sufficient conditions for absorbing MDPs*)
Consider a partition of $\mathbf{X}$ *to the disjoint sets* $\mathbf{X}'$ *and* $\mathcal{M}$.
The following statements are equivalent:
(*i*) *There exists a uniform Lyapunov function for the total expected life-time.*
(*ii*) $x \to \sup_{u \in U} E_x^u [T_{\mathcal{M}}]$ *is a Lyapunov function for the total expected life-time.*
(*iii*) *For all* x, *the MDP is absorbing to* $\mathcal{M}$ *and* $E_x^u [T_{\mathcal{M}}]$ *is continuous over* U_M.

The proof of the Lemma can be found in Cavazos-Cadena and Hernández-Lerma (1992). We shall present and prove a more general equivalence result in Theorem 7.3.

Consider a non-negative function $\nu : \mathcal{K} \to \mathbb{R}$. We make throughout the assumption that for every state x, $\nu(x, \cdot)$ are continuous on $\mathsf{A}(x)$. They will serve as some bounds on the immediate costs.

Definition 7.5 (*General Uniform Lyapunov function for the total cost*) Consider a partition of $\mathbf{X}$ to the disjoint sets $\mathbf{X}'$ and $\mathcal{M}$. A function $\mu : \mathbf{X} \to [1, \infty)$ is said to be a uniform Lyapunov function if M1(i)–M1(iii) below hold:

M1(i) For all $(x,a) \in \mathcal{K}$,

$$\nu(x,a) + 1 + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay}\mu(y) \le \mu(x). \tag{7.6}$$

M1(ii) For each $x \in \mathbf{X}$, the mapping $a \to \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}\mu(y)$ is continuous in $\mathsf{A}(x)$ (i.e., if $a_n \to a$, then $\sum_{y\in\mathbf{X}'} \mathcal{P}_{x,a_n,y}\mu(y)$ converges to $\sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}\mu(y)$).
M1(iii) For each $x \in \mathbf{X}$ and $u \in U_D$,

$$\lim_{n\to\infty} \sum_{y\in\mathbf{X}} [{}_{\mathcal{M}}P(u)]^n_{xy}\mu(y) = 0, \tag{7.7}$$

where ${}_{\mathcal{M}}P(u)$ is the Taboo probability matrix obtained by replacing columns $j \in \mathcal{M}$ by columns with zero entries.

If such a function μ exists, then the MDP is said to have a uniform Lyapunov function.

When using MDPs with uniform Lyapunov functions, we shall assume that the immediate costs satisfies (i) or (ii) in the following definition.

Definition 7.6 (*Bounds on the immediate costs*)
(i) The immediate costs are said to be ν-bounded if

$$\|c\|_\nu \stackrel{\text{def}}{=} \sup_{x,a} \frac{|c(x,a)|}{\nu(x,a)} < \infty, \tag{7.8}$$

with similar relations for $d^k, k = 1, \ldots, K$.
(ii) The immediate costs are said to be ν-bounded from below if their negative parts are ν-bounded. In other words, let $c^-(x,a) \stackrel{\text{def}}{=} \min(c(x,a), 0)$. Then

$$\sup_{x,a} \frac{|c^-(x,a)|}{\nu(x,a)} < \infty, \tag{7.9}$$

with similar relations for $d^k, k = 1, \ldots, K$.

We show in the next section that the Definitions 7.4 and 7.5 for uniform Lyapunov functions are equivalent under some simple transformation of the transition probabilities and costs. This allows one to make use of many results obtained for the first definition. The transformation will enable us to leave total expected costs unchanged, and yet, restrict to bounded immediate costs. The parameters in the new MDP will be denoted by adding a bar to the original one.

The following two sections establish general important properties of MDPs with uniform Lyapunov functions; the reader may skip these in a first reading of the monograph.

7.3 Equivalence of MDP with unbounded and bounded costs*

We show that an MDP satisfying Definition 7.5 can be transformed into an equivalent new one with bounded cost, satisfying Definition 7.4.

Roughly speaking, in the new MDP, whenever for some state and action (x, a), the immediate cost $c(x, a)$ is larger in absolute value than one, then we replace the cost by $c(x, a)/(\nu(x, a) + 1)$, and we 'compensate' for that by staying in a state x, $\nu(x, a)+1$ time longer (in some probabilistic sense) than in the original MDP. The latter is done by decreasing the transition probabilities out of that state by a factor of $1 + \nu(x, a)$.

Definition 7.7 *(Equivalent MDP)*
Define the following **New MDP** *with bounded costs:*

- *State space:* $\overline{\mathbf{X}} = \mathbf{X}$.
- *Actions:* $\overline{\mathsf{A}}(x) = \mathsf{A}(x)$.
- *Transition probabilities:*

$$\overline{\mathcal{P}}_{xay} = \begin{cases} \dfrac{\mathcal{P}_{xay}}{1 + \nu(x, a)} & \text{if } y \neq x, \\ \dfrac{\nu(x, a) + \mathcal{P}_{xax}}{1 + \nu(x, a)} & \text{otherwise .} \end{cases}$$

- *Immediate cost:*

$$\overline{c}(x, a) = \frac{c(x, a)}{1 + \nu(x, a)}.$$

- *Initial distribution:* $\overline{\beta} = \beta$.

We shall see that the new MDP is related to the original one in that the state and action at time n in the original MDP, under any policy $u \in U_M$, can be coupled with the state and action in the new MDP at time $\tau(n)$ under some equivalent quasi-Markov policy $\overline{u}$ (see Section 6.7). Moreover, the expected cost at time n in the initial MDP corresponds to the total expected cost during $[\tau(n), \tau(n + 1))$. ($\tau(n)$ is defined in (6.17) for quasi-Markov policies.)

More precisely, we consider the class of policies U_R (defined in Section 6.6) for the new MDP. Recall that these policies use some extra independent randomizations at each step. We assume that this is done by tossing a coin whose outcome is within $\mathbf{I} \stackrel{\text{def}}{=} \{0, 1\}$. We set the probabilities q (in the definition of the extra randomization in these policies) of obtaining 1, to be

$$q(x, a) = \frac{1}{1 + \nu(x, a)}.$$

For any $u \in U_M$ in the original MDP, we now define the quasi-Markov policy $\overline{u} = s(u) \in U_R$ as

$$\overline{u}_t(\mathcal{A}|h_t) = u_{\eta(t)}(\mathcal{A}|x), \quad t \in \mathbb{N}, \tag{7.10}$$

where η is defined in (6.18), and $h_t = (x_1, i_1, a_1, \ldots, x_t)$ is the history observed in the new MDP.

Theorem 7.1 (*Equivalence between unbounded and bounded MDPs*) *Consider any Markov policy u in the original MDP, and the corresponding quasi-Markov policy $s(u)$ in the new MDP (see (7.10)). Then*
(*i*) *The total expected cost under both MDPs are the same:*

$$C_{tc}(\beta, u) = \overline{C}_{tc}(\overline{\beta}, s(u)).$$

(*ii*) *The total expected time under a quasi-Markov policy $\overline{u}$ until the set $\overline{\mathcal{M}}$ is hit in the new MDP, starting from an initial distribution $\overline{\beta}$, is given by $\hat{M}_{\overline{u}}(\overline{\beta}) = \hat{M}(\beta, u)$, where*

$$\hat{M}(\beta, u) \stackrel{\text{def}}{=} E^u_\beta \sum_{n=1}^{\infty} [1 + \nu(X_n, A_n)] 1\{T > n\}. \tag{7.11}$$

(*iii*) *The optimal value for* **COP** *with total cost criterion is the same in the two MDPs, and a Markov policy is optimal in the original MDP* (*with unbounded cost*) *if and only if $s(u)$ is optimal in the new MDP* (*with bounded cost*)

Proof. Choose some Markov policy u for the original MDP. It is easily seen that the distribution of the process $\{\overline{X}_{\tau(n)}\}_n$ under the policy $s(u)$ is related to the distribution of $\{X_n\}_n$ under the policy u, by

$$\begin{aligned} & P^{s(u)}(\overline{X}_{\tau(n+1)} = y | \overline{X}_{\tau(n)} = x, \overline{T} > \tau(n)) \\ = \;& P^u(X_{n+1} = y | X_n = x, T > n) \end{aligned} \tag{7.12}$$

(quantities with a bar correspond to the new MDP). Denote by $\overline{\mathcal{F}}_t$ the σ-field generated by the history until time t of the new MDP. Denote by $\overline{\mathcal{F}}_{\tau(n)}$ the σ-algebra generated by events $\{\tau(n) \le t\} \cap A$, $A \in \overline{\mathcal{F}}_t$. We note that

$$\begin{aligned} & E^{s(u)}_{\overline{\beta}} \left(1\{\overline{T} > \tau(n)\} (\tau(n+1) - \tau(n)) | \overline{\mathcal{F}}_{\tau(n)} \right) \\ = \;& 1\{\overline{T} > \tau(n)\} [1 + \nu(\overline{X}_{\tau(n)}, \overline{A}_{\tau(n)})] \end{aligned} \tag{7.13}$$

(we understand above $1\{\overline{T} > \tau(n)\}(\tau(n+1) - \tau(n)) = 0$ when $\tau(n) = \infty$). Hence, the total cost under $s(u)$ is given by

$$\begin{aligned} \overline{C}_{tc}(\overline{\beta}, s(u)) &= \sum_{t=1}^{\infty} E^{s(u)}_{\overline{\beta}} \overline{c}(\overline{X}_t, \overline{A}_t) 1(\overline{T} > t) \\ &= \sum_{n=1}^{\infty} E^{s(u)}_{\overline{\beta}} 1\{\overline{T} > \tau(n)\} \sum_{r=\tau(n)}^{\tau(n+1)-1} \overline{c}(\overline{X}_r, \overline{A}_r) \\ &= \sum_{n=1}^{\infty} E^{s(u)}_{\overline{\beta}} \Big[E^{s(u)}_{\overline{\beta}} \big(1\{\overline{T} > \tau(n)\} \times \\ &\qquad (\tau(n+1) - \tau(n)) \overline{c}(\overline{X}_{\tau(n)}, \overline{A}_{\tau(n)}) | \overline{\mathcal{F}}_{\tau(n)} \big) \Big] \end{aligned}$$

$$= \sum_{n=1}^{\infty} E_{\overline{\beta}}^{s(u)} \left[c(\overline{X}_{\tau(n)}, \overline{A}_{\tau(n)}) 1\{\overline{T} > \tau(n)\} \middle| \overline{\mathcal{F}}_{\tau(n)} \right]$$

$$= \sum_{n=1}^{\infty} E_{\beta}^{u} c(X_n, A_n) 1\{T > n\} = C_{tc}(\beta, u),$$

where the equality before the last follows from (7.12). This establishes (i). (ii) follows from (7.13) by similar arguments. (iii) follows since both Markov as well as quasi-Markov policies are dominant for the total cost criterion (Theorems 6.2, 6.5). □

In the special case where the immediate costs are bounded below by some positive constant, we conclude from Theorem 7.1 that the problem of minimization of the total expected cost until a set $\mathcal{M}$ is reached can be transformed into a problem of minimization of the expected total time until the set $\mathcal{M}$ is reached. This is done as follows. Let $\underline{c} > 0$ be a lower bound on the immediate cost. Choose $\nu(x,a) = 2|c(x,a)|/\underline{c} - 1$. We note that c is indeed ν-bounded with $\|c\|_\nu \leq \underline{c}$. The new immediate cost is a constant $\overline{c}(x,a) = \underline{c}/2$. Therefore the total expected time to reach $\mathcal{M}$ is proportional to the total expected cost until it is reached.

Remark 7.1 (*Unbounded immediate costs*)
It follows from the proof of Theorem 7.1 (i) that it holds in fact for immediate costs that need not be ν-bounded; in fact, all that is required for it to hold is that the expectations and summation in the definition of the total cost $C_{tc}(\beta, u)$ are well defined. This is in particular the case when the immediate costs are non-negative.

Before proceeding to the equivalence between the Lyapunov functions, we present a useful Lemma (see Hordijk, 1977).

Lemma 7.3 (*Policy independent total costs*)
Assume that the MDP has a uniform Lyapunov function for the total cost. Define

$$\nu'(x,a) \stackrel{\text{def}}{=} \mu(x) - \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \mu(y) - 1. \tag{7.14}$$

Then

$$\hat{M}'(x,u) \stackrel{\text{def}}{=} E_x^u \sum_{n=1}^{\infty} [1 + \nu'(X_n, A_n)] 1\{T > n\} = \mu(x), \quad \forall u.$$

Proof. It easily follows that M1 holds with ν' replacing ν. Iterating (7.14) we get

$$E_x^u \sum_{n=1}^{m} [(1 + \nu'(X_n, A_n)) 1\{T > n\}] = \mu(x) - E_x^u \mu(X_{m+1}) 1\{T > m+1\}. \tag{7.15}$$

$E_x^u \mu(X_{m+1}) 1\{T > m+1\}$ converges to 0 for any $u \in U_D$ by M1(iii). In fact, it is proved in Hordijk (1977, pp. 43-44) that it converges to 0 for any policy (and that the convergence is uniform in the policies). Therefore $\hat{M}'(x,u) = \hat{M}'(x) \stackrel{\text{def}}{=} \sup_u \hat{M}'(x,u) = \mu(x)$. □

Remark 7.2 (*Monotonicity properties*)
It follows from the representation (7.15) that for any policy and state x, $E_x^u \mu(X_{m+1}) 1\{T > m+1\}$ is monotone non-increasing in m, if condition M1(i) of the Lyapunov function holds, even if the other conditions do not hold. (We conclude from the Lemma that the limit is zero if μ is a uniform Lyapunov function.)

Theorem 7.2 (*Relation between the Lyapunov conditions*)
(a) Assume that the original MDP has a uniform Lyapunov function μ (Definition 7.5) for the total cost. Then the new MDP has the same uniform Lyapunov function (Definition 7.4) μ for the total expected time until the set $\mathcal{M}$ is reached.
(b) Assume that the new MDP has a uniform Lyapunov function μ for the total expected time until the set $\mathcal{M}$ is reached. Then μ is a uniform Lyapunov function for the original MDP.

Proof. (a) Assume that the original MDP has a uniform Lyapunov function. Choose some x, and denote $q(x,a) := [1+\nu(x,a)]^{-1}$, $\bar{q} := 1-q$. The following holds:

$$\sum_{\overline{y} \notin \mathcal{M}} \overline{\mathcal{P}}_{\overline{x}a\overline{y}} \mu(\overline{y}) = q(x,a) \sum_{y \notin \mathcal{M}} \mathcal{P}_{xay} \mu(y) + \bar{q}(x,a)\mu(x). \tag{7.16}$$

Since μ is a uniform Lyapunov function for the original MDP, we have

$$\begin{aligned} 1 + \sum_{\overline{y} \notin \mathcal{M}} \overline{\mathcal{P}}_{\overline{x}a\overline{y}} \mu(\overline{y}) &= 1 + q(x,a)\Big(1 + \nu(x,a) + \sum_{y \notin \mathcal{M}} \mathcal{P}_{xay}\mu(y)\Big) \\ &\qquad - (1+\nu(x,a))q(x,a) + \bar{q}(x,a)\mu(x) \\ &= q(x,a)\Big(1 + \nu(x,a) + \sum_{y \notin \mathcal{M}} \mathcal{P}_{xay}\mu(y)\Big) + \bar{q}(x,a)\mu(x) \\ &\leq q(x,a)\mu(x) + \bar{q}(x,a)\mu(x) = \mu(x). \end{aligned} \tag{7.17}$$

Hence, condition (i) of Definition 7.5 is satisfied by μ.

Next, we check (ii). Definition 7.5 for the original MDP can be formulated as $\sum_{y \notin \mathcal{M}} \mathcal{P}_{xay}\mu(y)$ being continuous in a for all x. The continuity is established by noticing that q and $\bar{q}$ are continuous in a, and by using (7.16).

We now check condition (iii). Define

$$\overline{\nu}'(\overline{x},\overline{a}) \stackrel{\text{def}}{=} \frac{1+\nu'(x,a)}{1+\nu(x,a)} - 1,$$

where ν' is defined in Lemma 7.3, and where $\overline{x} = x$ and $\overline{a} = a$. We obtain

$$\begin{aligned} 1 + \overline{\nu}'(\overline{x},\overline{a}) + \sum_{\overline{y}\notin\mathcal{M}} \mathcal{P}_{\overline{xay}}\mu(\overline{y}) \\ = \quad 1 + \overline{\nu}'(x,a) + q(x,a)\Big(1 + \nu'(x,a) + \sum_{y\notin\mathcal{M}} \mathcal{P}_{xay}\Big) \\ - (1+\nu'(x,a))q(x,a) + \overline{q}(x,a)\mu(x) = \mu(x). \end{aligned}$$

Let $u \in U_D$ be a deterministic policy in the original MDP, and $s(u)$ the corresponding quasi-Markov policy in the new one. Iterating the above equation, we obtain:

$$\sum_{t=1}^{n} E_{\overline{x}}^{s(u)}(1+\overline{\nu}'(\overline{X}_t,\overline{A}_t))1(\overline{T}>t) + E_{\beta}^{s(u)}\mu(\overline{X}_{n+1})1\{\overline{T}>n+1\} = \mu(x).$$

From Theorem 7.1 (and Remark 7.1), it follows that

$$\begin{aligned} \sum_{t=1}^{\infty} E_{\overline{x}}^{s(u)}(1+\overline{\nu}'(\overline{X}_t,\overline{A}_t))1(\overline{T}>t) \\ = \quad \sum_{t=1}^{\infty} E_{\overline{x}}^{u}(1+\nu(X_t,A_t))1(T>t) = \mu(x). \end{aligned}$$

We conclude by combining the two last equations that

$$\lim_{n\to\infty} E_{\overline{x}}^{s(u)}\mu(\overline{X}_{n+1})1\{\overline{T}>n+1\} = 0.$$

Any stationary deterministic policy in the new MDP is in particular a quasi-Markov policy, and can be written as $s(u)$, where $u \in U_D$ is a stationary deterministic policy in the original MDP. We conclude that M1(iii) holds for the new MDP.

(b) Assume that the new MDP has a uniform Lyapunov function as stated. Recall the definition of q from the first part of the proof. We have from (7.16) for any x

$$\begin{aligned} \sum_{y\notin\mathcal{M}} \mathcal{P}_{xay}\mu(y) &= \frac{1}{q(x,a)}\left(\sum_{\overline{y}\notin\overline{\mathcal{M}}} \overline{\mathcal{P}}_{\overline{x}a\overline{y}}\mu(\overline{y}) + \bar{q}(x,a)\mu(x)\right) \\ &\le \frac{1}{q(x,a)}\left(\mu(\overline{x}) - 1 + \bar{q}(x,a)\mu(x)\right) \\ &= \mu(x) - \frac{1}{q(x,a)} \le \mu(x) - 1 - \nu(x,a). \end{aligned}$$

This establishes condition (i).

The proof of (ii) follows from (7.16). We finally check condition (iii).

$$\begin{aligned}&\left([_{\mathcal{M}}P(u)]^n\mu\right)_x\\ &= E_x^u(\mu(X_n)1\{T_{\mathcal{M}} > n\}) = E_{\overline{x}}^{\overline{u}}(\mu(\overline{X}_{\tau(n)})1\{\overline{T}_{\overline{\mathcal{M}}} > \tau(n)\})\\ &\leq E_{\overline{x}}^{\overline{u}}(\mu(\overline{X}_n)1\{\overline{T}_{\overline{\mathcal{M}}} > n\}) = \left([_{\overline{\mathcal{M}}}\overline{P}(\overline{u})]^n\mu\right)_{\overline{x}}.\end{aligned}$$

The last inequality follows since $\mu(\overline{X}_n)1\{\overline{T}_{\overline{\mathcal{M}}} > n\}$ is a super-Martingale and $\tau(n)$ is a stopping time, see e.g., p. 99 in Williams (1992). This implies condition (iii). □

7.4 Properties of MDPs with uniform Lyapunov functions*

Ten different equivalent conditions are presented in Cavazos-Cadena and Hernández-Lerma (1992) for an MDP to have a uniform Lyapunov function for the total expected time until a state x is reached. The equivalence that we established in the previous section between MDPs with general uniform Lyapunov function, and MDPs with a uniform Lyapunov function for the total expected time, may be helpful in extending the conditions to our original MDP. We shall often present direct proofs for completeness.

Remark 7.3 The equivalence results that we present below are more general than that of Cavazos-Cadena and Hernández-Lerma (1992) in the following points:
(i) The Taboo set $\mathcal{M}$ will not be restricted to a singleton.
(ii) We do not make the unichain assumption.
(iii) We obtain results for unbounded costs.

We begin by stating properties that are related to the total cost criterion.

- **(M1)** The MDP has a uniform Lyapunov function μ, i.e., it satisfies conditions M1(i), M1(ii) and M1(iii) in Definition 7.5.
- **(M1′)** For all $x \in \mathbf{X}$,
$$\lim_{n\to\infty} \sup_u E_x^u \mu(X_n)1\{T > n\} = 0.$$
- **(M2)** $\hat{M}(x) \stackrel{\text{def}}{=} \sup_{u\in U} \hat{M}(x,u) < \infty$ is a uniform Lyapunov function, where
$$\hat{M}(x,u) \stackrel{\text{def}}{=} E_x^u \sum_{n=1}^{\infty} [1 + \nu(X_n, A_n)]1\{T > n\} \tag{7.18}$$
(M2(i), M2(ii) and M2(iii) are defined to be the corresponding properties in Definition 7.5.)
- **(M3)** For each $x \in \mathbf{X}$, the following hold:
(i) for all $u \in U_D$, $\hat{M}(x,u) < \infty$ and

(ii) $u \to \hat{M}(x,u)$ is continuous over U_D.

- **(M4)** For each $x \in \mathbf{X}$, the following hold:
(i) $\lim_{n\to\infty} \sup_{u\in U} E_x^u \hat{M}(X_n) 1\{T > n\} = 0$,
(ii) $\hat{M}(x) < \infty$,
(iii) $E_x^u \hat{M}(X_n) 1\{T > n\}$ is continuous over U_M for all $x \in \mathbf{X}$ and $n \in \mathbb{N}$.

- **(M5)** For each $x \in \mathbf{X}$, the function $u \to \hat{M}(x,u)$ is continuous over U_M and is finite.

- **(M5′)** For each $x \in \mathbf{X}$, the following hold:
(i) For all $u \in U_M$, $C_{tc}(x,u) < \infty$, $M_u(x) < \infty$ (M_u is defined below (6.3), and
(ii) $u \to C_{tc}(x,u)$ and $u \to M_u(x)$ are continuous over U_M.

- **(M6)** For all $x \in \mathbf{X}$,

$$\lim_{n\to\infty} \sup_{u\in U} \frac{E_x^u \mu(X_n)}{n} = 0.$$

- **(M7)** There is a policy $w \in U_D$ such that for every initial state x, $\hat{M}(x) = \hat{M}(x,w) < \infty$.

- **(M8)** (i) For every initial state x, $\sup_{u\in U} \hat{M}(x,u) < \infty$,
(ii) For all $\varepsilon > 0$, there exists $u(\varepsilon) \in U_D$ such that $\hat{M}(x,u(\varepsilon)) > \hat{M}(x)(1-\varepsilon)$ for all $x \in \mathbf{X}$.

- **(M9)** $\hat{M}(x)$ is a solution of the optimality equation

$$\sup_a \left\{ 1 + \nu(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} v(y) \right\} = v(x). \tag{7.19}$$

Remark 7.4 (*Some derived properties*)
(i) For the special case where $\mathcal{M} = \{0\}$, where 0 is some arbitrary state, it follows from the statement $\hat{M}(x) < \infty$ that state 0 is recurrent under any policy; in particular, the expected time between two consecutive visits of state 0 is uniformly bounded.
(ii) Any one of the properties M2, M4 or M7 implies that the MDP is absorbing for any initial state x.

Theorem 7.3 (*Equivalence between conditions*)
(*i*) *M1* $\Leftrightarrow$*M2* $\Leftrightarrow$*M4* $\Leftrightarrow$*M5* $\Rightarrow$*M3* $\Rightarrow$*M7.*

(*ii*) *If the MDP is unichain and* $\mathcal{M}$ *is a singleton, then M3* $\Rightarrow$*M2.*
(*iii*) *M1* $\Rightarrow$*M1′, M1* $\Rightarrow$*M6 , M1* $\Rightarrow$*M5′.*

Proof. The proof of M1 $\Leftrightarrow$M2 $\Rightarrow$M5 is given in Lemma 7.5(iv). M5 $\Rightarrow$M7 and M4 $\Rightarrow$M2 are proved in Lemma 7.4.
M5 $\Rightarrow$M2 and M1 $\Rightarrow$M4 are established in Lemma 7.6 and Lemma 7.5(iii), respectively.
(ii) was established by Cavazos-Cadena and Hernández-Lerma (1992) for the case of bounded cost. The proof carries over directly to our case by using Theorem 7.3.
(iii) M1 $\Rightarrow$M1′ and M1 $\Rightarrow$M6 are established in Lemma 7.5(i).
M1 $\Rightarrow$M5′ is proved in Lemma 7.5 (v). □

Lemma 7.4 (*i*) *M2(i), M8(ii) and M9 hold.*
(*ii*) *M5 or M3 imply M7.*
(*iii*) *M4* $\Rightarrow$*M2.*

Proof. M9 is well known, see e.g., Hordijk (1977, Theorem 6.1). This implies M2(i). M8(ii) follows from Theorem 13.7 of Hordijk (1977) (see also Feinberg and Sonin, 1984, Van Der Wal, 1981b, and references therein). Since the class of stationary policies is compact, we obtain (ii).

(iii) M2(iii) follows from M4(i) and M2(ii) follows from M4(iii). M4(ii) together with M9 imply M2(i). □

Lemma 7.5 (*i*) *M1 implies M1′, M6 and M6′ where*
(M6): *For all* $x \in \mathbf{X}$,

$$\lim_{n\to\infty} \sup_{u\in U} \frac{E_x^u \hat{M}(X_n)}{n} = 0.$$

(*ii*) *M1 implies that* $\hat{M} \le \mu$, *and* $\hat{M}(\beta) \le \langle \beta, \mu \rangle$.
(*iii*) *M1 implies M4.*
(*iv*) *Conditions M1 and M2 are equivalent and imply M5.*
(*v*) *M1 implies M5′.*

Proof. Assume M1. It is proved in Hordijk (1977, pp. 43-44) that M1′ holds, and that M6′ holds (although the proof is written for $\mathcal{M} = \{z\}$ for some state z, it is unchanged for an arbitrary $\mathcal{M}$). We show that M6 holds:

Recall the definition of ν' in Lemma 7.3. Since, by definition, M1 holds also for ν' with the Lyapunov function μ, this implies that property M6′ holds for ν', which implies M6.

Iterating (7.6), one obtains for any x and $u \in U_M$

$$\sum_{n=1}^{k-1} E_x^u(1 + \nu(X_s, A_s))1\{T > n\} \le \mu(x) - E_x^u \mu(X_n)1\{T > n\} \le \mu(x). \tag{7.20}$$

Taking the limit as $n \to \infty$, we obtain the first statement of (ii). The second follows from the bounded convergence theorem.

(iii) M4(i) was established by Hordijk (1977, pp. 43-44), and M4(ii) follows from statement (ii) of our Lemma. M4(iii) is obtained by showing that $E_x^u \hat{\mu}(X_n) 1\{T > n\}$ is continuous over U_M for all $x \in \mathbf{X}$ and $n \in \mathbb{N}$, where $\hat{\mu}$ is any function that is bounded by the Lyapunov function μ. To show that, we first note that $E_\beta^u \hat{\mu}(X_2) 1\{T > 2\}$ is continuous in $u \in U_M$; this follows from condition M1(ii). We proceed by induction to show that

$$E_x^u \hat{\mu}(X_n) 1\{T > n\} = \left[\left({}_{\mathcal{M}}P(u_1) {}_{\mathcal{M}}P(u_2) \cdots {}_{\mathcal{M}}P(u_{n-1}) \right) \hat{\mu} \right]_x \tag{7.21}$$

is continuous in u. Suppose (7.21) holds for some n. We have

$$E_x^u \hat{\mu}(X_{n+1}) 1\{T > n+1\} = \sum_{y \in \mathbf{X}'} {}_{\mathcal{M}}P_{xy}(u_1) Z_y(u)$$

where

$$Z_y(u) \stackrel{\text{def}}{=} \left[\left({}_{\mathcal{M}}P(u_2) {}_{\mathcal{M}}P(u_3) \cdots {}_{\mathcal{M}}P(u_n) \right) \hat{\mu} \right]_x .$$

By Iterating (7.6) in property M1(i), it follows that $Z_y(u) \leq \mu(y)$ for any u and y. Moreover, the inductive assumption implies that Z_y is continuous in $u \in U_M$. Since $[{}_{\mathcal{M}}P(u_1)\mu]_x$ is continuous in u_1, it follows from the generalized dominated convergence that

$$[{}_{\mathcal{M}}P(u_1) Z(u)]_x = E_x^u \hat{\mu}(X_{n+1}) 1\{T > n+1\}$$

is continuous in $u \in U_M$, which establishes the induction.

(iv) That M2 implies M1 is trivial.

Assume that M1 holds. M2(i) follows from property M9 (see Lemma 7.4 (i)). M2(ii) follows from part (ii) of our Lemma and the dominated convergence theorem. M2(iii) follows from property M4(iii) (and statement (iii) of our Lemma).

Next, we show that M1 implies M5.

Let $c' : \mathcal{K} \to \mathbb{R}$ be any continuous function satisfying $|c'| \leq \nu + 1$. It follows from property M1(i) that for all $t \geq 1$ and all x,

$$E_\beta^u c'(X_t, A_t)) 1\{T > t\} + E_\beta^u \mu(X_{t+1}) 1\{T > t+1\} \leq E_\beta^u \mu(X_t) 1\{T > t\} \tag{7.22}$$

where $\beta = \delta_x$. By summing over $t \in n, \ldots, n+m$, we get

$$E_\beta^u \sum_{t=n}^{n+m} [c'(X_t, A_t) 1\{T > t\}] \leq E_\beta^u \mu(X_n) 1\{T > n\}.$$

Taking the limit as $m \to \infty$, we get

$$E_\beta^u \sum_{t=n}^{\infty} [c'(X_t, A_t) 1\{T > t\}] \leq \sup_{u \in U} E_\beta^u \mu(X_n) 1\{T > n\}, \tag{7.23}$$

which tends to zero as $n \to \infty$ according to property M4(i).

Thus, $E_\beta^u \sum_{t=1}^n [c'(X_t, A_t) 1\{T > t\}]$ converges to $C'_{tc}(\beta, u)$ uniformly in $u \in U_m$ ($C'_{tc}(\beta, u)$ is the total cost corresponding to the immediate cost c').

For any n, we shall show that $E^u_\beta \sum_{t=1}^n [c'(X_t, A_t)1\{T > t\}]$ is continuous in u. The proof is then established by combining this continuity with the uniform convergence (that we established in the previous paragraph).

Let $c'(x, u_t) \stackrel{\text{def}}{=} \int c'(x, a)u_t(da|x)$. We have

$$E^u_\beta [c'(X_t, A_t)1\{T > t\}] = \sum_{x\in\mathbf{X}} P^u_\beta(X_t = x, T > t)c'(x, u_t).$$

We note that $c'(x, u_t)$ is continuous in u_t for every x, since it $c'(\cdot, \cdot)$ is continuous. Moreover, $P^u_\beta(X_t = x, T > t)$ are continuous over $u \in U_M$; this follows from (7.21). Since $c'(x, u_t) \le \nu(x, t) + 1 \le \mu(x)$ for any u and x, and since $\sum_{x\in\mathbf{X}} P^u_\beta(X_t = x, T > t)\mu(x)$ is continuous, it follows from the dominated convergence theorem that $E^u_\beta [c'(X_t, A_t)1\{T > t\}]$ is continuous in u. M5 follows by setting $c' = 1 + \nu$.

(v) The statements for M_u and $C_{tc}(x, u)$ are obtained by replacing $c'(x, a)$ by 1 and by c, respectively, in the proof of M1 $\Rightarrow$M5 in part (iv). □

Remark 7.5 An alternative way to prove part (iv) of the Lemma is by adapting the proof of Theorem 6.1 in Cavazos-Cadena and Hernández-Lerma (1992): one may use the proof there to show that (iii) holds for the modified MDP given in Definition 7.7. (This is done by replacing the indicators $1(X_n \neq z)$ in Cavazos-Cadena and Hernández-Lerma (1992) by $1(X_n \notin \mathcal{M})$.) The result then follows by applying Theorems 7.1 and 7.2.

Lemma 7.6 *M5 implies M2.*

Proof. Since M5 $\Rightarrow$ M7, there is a stationary policy w achieving $\hat{M}(x) = \hat{M}(x, w)$. Let $v_a \in U_M$ be given by $v_a = (a, w, w, w, \dots,)$. Then

$$\hat{M}(x, v_a) = 1 + \nu(x, a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}\hat{M}(y).$$

Since $\nu(x, a)$ and $\hat{M}(x, v_a)$ are continuous in a (by M5), it follows that $\sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}\hat{M}(y)$ is continuous in a, which establishes M2(ii).

Consider an arbitrary $u \in U_D$ and define $v(n) = [u \times n, w]$ to be the policy that follows u for the n first periods, and then switches to the policy w. Then

$$\hat{M}(x, v(k-1)) = \sum_{m=1}^{k-1} E^u_x(1 + \nu(X_m, A_m))1\{T > m\} + E^u_x\hat{M}(X_k)1\{T > k\} \tag{7.24}$$

M5 implies

$$\hat{M}(x, u) = \lim_{n\to\infty} \hat{M}(x, v(n)) = \hat{M}(x, u) + \lim_{n\to\infty} E^u_x\hat{M}(X_n)1\{T > n\}$$

which establishes M2(iii). □

Since M1 implies M2, we conclude that an MDP with a uniform Lyapunov function is absorbing.

7.5 Properties for fixed initial distribution*

The equivalent or necessary conditions in the previous section all involved *all initial states*. It is natural to investigate the following questions:

- (i) Can one conclude that a property holds for all initial state by checking at only one initial state?
- (ii) Do properties hold for arbitrary initial distributions β, rather than for fixed initial states?

Cavazos-Cadena and Hernández-Lerma (1992) showed that some conditions that hold for all initial states are equivalent to conditions that hold for a particular *single state* z, provided that $\mathcal{M} = \{z\}$. Below, we present similar results for any initial distribution, which provides an answer to the above questions.

Definition 7.8 *For a fixed initial distribution β, define a state x to be β-accessible if there exists some policy $v \in U_M$ (that may depend on x) and an integer n such that*

$$P_\beta^v(X_n = x, T > n) > 0. \tag{7.25}$$

Let $\hat{\mathbf{X}}(\beta) \stackrel{\text{def}}{=} \{x \in \mathbf{X} : x \text{ is } \beta\text{-accessible}\}$.

Note that $\hat{\mathbf{X}}(\beta)$ contains all states in the support of β. Thus,

Lemma 7.7 $\mathbf{X} = \hat{\mathbf{X}}(\beta)$ *if any of the following conditions hold:*
(i) $\beta(y) > 0$ *for all* $y \in \mathbf{X}$,
(ii) $\mathcal{M} = \{z\}$ *for some* z, *the MDP is communicating and* $\beta = \delta_z$,
(iii) $\mathcal{M} = \{z\}$ *for some* z, *and for any* $y \neq z$ *there exists a policy* $u \in U_M$ *such that* $P_x^u(X_n = y) > 0$ *for some integer* n.

We introduce the following conditions:

- **(N1)** M1 holds with $\hat{\mathbf{X}}$ replacing $\mathbf{X}$ and $\hat{\mathbf{X}} \cap \mathbf{X}'$ replacing $\mathbf{X}'$. Moreover, the function μ satisfies $\beta\mu < \infty$.

- **(N2)** $\hat{M}(x)$ satisfies N1.

- **(N3)** The following hold:
  (i) for all $u \in U_D$, $\hat{M}(\beta, u) < \infty$ and
  (ii) $u \to \hat{M}(\beta, u)$ is continuous over U_D.

- **(N4)** The following hold:
  (i) $\lim_{n\to\infty} \sup_{u\in U} E_\beta^u \hat{M}(X_n) 1\{T > n\} = 0$,
  (ii) $\hat{M}(\beta) < \infty$,
  (iii) $E_\beta^u \hat{M}(X_n) 1\{T > n\}$ is continuous over U_M for all $n \in \mathbb{N}$.

- **(N5)** The function $u \to \hat{M}(\beta, u)$ is continuous over U_M and is finite.

- **(N5′)** (i) For all $u \in U_M$, $C_{tc}(\beta, u) < \infty$, $M_u(\beta) < \infty$, and (ii) $u \to C_{tc}(\beta, u)$ and $u \to M_u(\beta)$ are continuous over U_M.

- **(N6)** For each $x \in \hat{\mathbf{X}}$, the function $u \to \hat{M}(x, u)$ is continuous over U_M and is finite.

Remark 7.6 (*Some derived properties*)
Fix an initial distribution β. Then any one of the properties N3, N4, N5 or N6 implies that the MDP is absorbing for any initial state x.

Theorem 7.4 (*Equivalence between conditions*)
M1 $\Rightarrow$N4 $\Leftrightarrow$N1 $\Leftrightarrow$N2 $\Leftrightarrow$N5 $\Leftrightarrow$N6 $\Rightarrow$N3. M1 $\Rightarrow$N5′.

Proof. The proof follows from Lemma 7.8 and Lemma 7.9 below. □

Corollary 7.1 *If $\hat{\mathbf{X}}(\beta) = \mathbf{X}$, then N4, N5, N6, M1, M2, M4, M5 are equivalent.*

Lemma 7.8 *M1 $\Rightarrow$N4.*

Proof. M1 $\Rightarrow$N4: Define the sequence $\xi_n \in \mathbb{R}^{\mathbf{X}}$:
Let $\xi_n(x) \stackrel{\text{def}}{=} \sup_u E_x^u \mu(X_n) 1\{T > n\}$. It follows from M1′ (which is implied by M1, see Lemma 7.5(i)) that $\xi_n \to 0$ and that $\xi < \mu$. Applying the dominated convergence theorem, we obtain

$$\lim_{n\to\infty} E_\beta^u \mu(X_n) 1\{T > n\} \le \lim_{n\to\infty} \beta\xi_n = \beta \lim_{n\to\infty} \xi_n = 0.$$

□

Lemma 7.9 (*i*) *N5 implies N6.* (*ii*) *N6 $\Leftrightarrow$N1 $\Leftrightarrow$N2 $\Rightarrow$N4 , N1 $\Rightarrow$N5 $\Rightarrow$N3, N1 $\Rightarrow$N5′.*

Proof. N5 $\Rightarrow$N6: Let $x \in \hat{\mathbf{X}}(\beta)$. Let u and n be as in (7.25). Choose an arbitrary sequence of Markov policies $u(m)$ converging to some limit Markov policy which we denote by w. Define the Markov policy $v(m)$, $m = 1, 2, \ldots$ as follows:

$$v_t(m) = \begin{cases} u_t & \text{if } t < n, \\ w_{t-n+1}(m) & \text{if } t \ge n. \end{cases} \tag{7.26}$$

Define similarly the policy v, where $w(m)$ is replaced by w. One can show using the Markov property that

$$\hat{M}(\beta, v(m)) = E_\beta^u \sum_{k=1}^{n-1} \big(1 + \nu(X_k, A_k)\big) 1\{T > k\} \tag{7.27}$$

$$+ \sum_{y \in \hat{\mathbf{X}}(\beta)} P_\beta^u(X_n = y, T > n) E_y^{w(m)} \sum_{k=1}^{\infty} \big(1 + \nu(X_k, A_k)\big) 1\{T > k\}.$$

The same relation holds for v (with w replacing $w(m)$). By Fatou's Lemma, we have

$$\underline{\lim}_{m\to\infty} \hat{M}(\beta, v(m)) \geq E^u_\beta \sum_{k=1}^{n-1} \big(1+\nu(X_k, A_k)\big)1\{T>k\} + \tag{7.28}$$

$$\sum_{y\in\hat{\mathbf{X}}(\beta)} P^u_\beta(X_n = y, T>n)\ \underline{\lim}_{m\to\infty} E^{w(m)}_y \sum_{k=1}^{\infty} \big(1+\nu(X_k, A_k)\big)1\{T>k\}.$$

For any $y \in \mathbf{X}$ and $t \in \mathbb{N}$, one can show that

$$\underline{\lim}_{m\to\infty} E^{w(m)}_y \big(1+\nu(X_t, A_t)\big)1\{T>t\} \geq E^w_y \big(1+\nu(X_t, A_t)\big)1\{T>t\}$$

(for exact details, see the Lemma 8.1 (i) and its proof). If strict inequality holds for some t for $y \in \hat{\mathbf{X}}(\beta)$ (i.e., assumption N6 does not hold), then this, together with (7.27) and (7.28), would imply that

$$\underline{\lim}_{m\to\infty} \hat{M}(\beta, v(m)) > \hat{M}(\beta, v).$$

But since $v(m)$ converges to v (this follows from the convergence of $w(m)$ to w and the definition of $v(m)$ and v), this contradicts assumption N5 (N5 implies equality instead of the strict inequality in the above equation). Hence, if N6 does not hold, then N5 does not hold. This implies (i).

(ii) We begin by showing N1 $\Leftrightarrow$N2 $\Leftrightarrow$N6. One may consider a new MDP which is obtained by restricting the original one to $\hat{\mathbf{X}}(\beta)$ (thus, in particular, $\mathcal{P}_{xay}$ are the same as the original MDP for all $x, y \in \hat{\mathbf{X}}(\beta)$). For each Markov policy u in the new MDP we may associate a set $V(u)$ of Markov policies in the original MDP such that $u_t(\cdot|y) = v_t(\cdot|y)$ for each $v \in V(u)$, $t \in \mathbb{N}$ and $y \in \hat{\mathbf{X}}(\beta)$; one can then construct a probability space for which the trajectories of the states and actions are the same for the original MDP under u, and the new MDP under v, so they have the same costs. Applying Theorem 7.3, we then obtain the equivalence of N6 and the properties M1 and M2 for the new MDP which are properties N1 and N2 in the original MDP.

Next, we show that N1 $\Rightarrow$N4 $\Rightarrow$N2. That N1 implies N4 is established exactly as in the proof of Lemma 7.8. The implication N4 $\Rightarrow$N2 follows by arguments similar to Lemma 7.4(iii). Indeed, N4(ii) implies that $\hat{M}(x)$ is finite for all $x \in \hat{\mathbf{X}}$. This, together with M9, implies N2(i).
Next we establish N2(iii). For any $u(m) \in U_D$ converging to some $u \in U_D$, we have

$$\underline{\lim}_{m\to\infty} E^{u(m)}_x \hat{M}(X_2)1\{T>2\} \geq E^u_x \hat{M}(X_2)1\{T>2\}.$$

This follows directly by using Fatou's Lemma, since $\mathcal{P}_{x,a,y}$ are continuous over $\mathsf{A}(x)$. Assume that N2(iii) does not hold. Then there are some deterministic policies $u(m) \in U_D$ converging to some limit policy $u \in U_D$ and

some $x \in \hat{\mathbf{X}}$, such that

$$\underline{\lim}_{m\to\infty} E_x^{u(m)} \hat{M}(X_2) 1\{T > 2\} > E_x^u \hat{M}(X_2) 1\{T > 2\}. \tag{7.29}$$

Let v and n satisfy (7.25). Define the Markov policies

$$w(m) = (\underbrace{v, v, \ldots, v}_{n-1}, u(m), u(m), u(m), \ldots), \qquad w = (\underbrace{v, v, \ldots, v}_{n-1}, u, u, u, \ldots). \tag{7.30}$$

Then the strong Markov property implies

$$E_\beta^{w(m)} = \sum_{y\in\mathbf{X}} P_\beta^v(X_n = y, T > n) E_y^{u(m)} \hat{M}(X_2) 1\{T > 2\}.$$

Applying Fatou's Lemma and (7.29), we obtain

$$\begin{aligned}
&\underline{\lim}_{m\to\infty} E_\beta^{w(m)} \hat{M}(X_{n+1}) 1\{T > n+1\} \\
&\quad \geq \quad \underline{\lim}_{m\to\infty} \sum_{y\in\mathbf{X}} P_\beta^v(X_n = y, T > n) E_y^{u(m)} \hat{M}(X_2) 1\{T > 2\} \\
&\quad > \quad \sum_{y\in\mathbf{X}} P_\beta^v(X_n = y, T > n) E_y^u \hat{M}(X_2) 1\{T > 2\} \\
&\quad = \quad E_\beta^w \hat{M}(X_{n+1}) 1\{T > n+1\}.
\end{aligned}$$

This contradicts N4(iii), which implies the relation N4 $\Rightarrow$N2(ii).
We establish next N2(iii). Assume that it does not hold, i.e., there exists some $u \in U_D$ such that

$$\overline{\lim}_{m\to\infty} E_x^u \hat{M}(X_m) 1\{T > m\} > 0.$$

Let v and n and x be as in (7.25), and let u define the policy w as in (7.30). Then by the strong Markov property, we have

$$\begin{aligned}
&E_\beta^w \hat{M}(X_{m+n}) 1\{T > m+n\} \\
&\quad \geq \quad P_\beta^v(X_n = x, T > n) E_x^u \hat{M}(X_m) 1\{T > m\}.
\end{aligned}$$

This implies that

$$\overline{\lim}_{m\to\infty} E_\beta^w \hat{M}(X_m) 1\{T > m\} > 0,$$

which contradicts N4(i). This establishes the relation N4 $\Rightarrow$N2(iii).

N1 $\Rightarrow$N5 and N1 $\Rightarrow$N5$'$ follow by the same arguments as in the proof of Lemma 7.5 (iv)-(v).

N5 $\Rightarrow$N3 is trivial. □

We finally remark that the following relation holds:

Lemma 7.10 *M3 $\Rightarrow$N3.*

7.6 Examples of uniform Lyapunov functions

A natural question that arises is whether a given MDP has a uniform Lyapunov function, and if it does, how can we compute it.

As we see from property M2 (Section 7.4), a candidate for a Lyapunov function is obtained by solving an MDP in which we *maximize* the total expected cost until we hit a set $\mathcal{M}$, with respect to the immediate cost $1+\nu$. This might suggest that in order to minimize the total expected cost, we have first to consider the problem of maximization of another total cost problem.

This also illustrates another weakness of the uniform Lyapunov-function approach: in order to apply this approach, the worst possible policy (the one that maximizes some total expected cost, instead of the one minimizing it) has to lead to a finite expected cost and to have some good properties.

Fortunately, it turns out in practice that when uniform Lyapunov functions exist, it is often easy to identify and to compute one (which is typically different than the candidate from property M2). This is in particular the case in many problems in the control of queues, as we shall show below. When uniform Lyapunov functions do not exist but the immediate costs are bounded below, there is still a rich theory that can be applied (in particular, the one for transient MDPs, which we develop in the two following chapters).

When costs are not bounded below and uniform Lyapunov functions do not exist, then the characterization and computation of "good" policies are particularly complex. The dynamic programming techniques that apply in the other cases for the non-constrained control problem do not hold any more, and optimal policies need not exist. We shall illustrate these problems in Section 9.8.

In the remainder of this section we consider several applications in which uniform Lyapunov functions exist and can be easily computed.

Example 7.1 (*Flow and service control*)
Consider the example of Chapter 5, this time with an infinite buffer $L = \infty$. The model is unchanged, except that we do not need to make the assumption $0 \in \mathsf{B}(x)$. Assume that

- The cost c is polynomially bounded;
- $b_{max} < a_{min}$.

Let $\mathcal{M} \stackrel{\text{def}}{=} \{0\}$ (it thus contains only the state of an empty queue).

Define now

$$\rho \stackrel{\text{def}}{=} \frac{b_{max}\bar{a}_{min}}{\bar{b}_{max}a_{min}}$$

and note that $\rho < 1$.

Define further

$$\mu(x) \stackrel{\text{def}}{=} Cr^x$$

where r is some arbitrary number satisfying $r \in (1, \rho^{-1})$ and C is a positive real number that will be determined later. Then for $x > 1$,

$$\begin{aligned}
&\sum_{y\neq 0} \mathcal{P}_{xaby}\mu(y) - \mu(x) \\
&\quad = \bar{b}a(\mu(x-1) - \mu(x)) + b\bar{a}(\mu(x+1) - \mu(x)) \\
&\quad \leq \bar{b}_{max}a_{min}(\mu(x-1) - \mu(x)) + b_{max}\bar{a}_{min}(\mu(x+1) - \mu(x)) \\
&\quad = Cr^x(r-1)\left(b_{max}\bar{a}_{min} - \bar{b}_{max}a_{min}r^{-1}\right) \\
&\quad = Cr^x(r-1)r^{-1}b_{max}\bar{a}_{min}\left(r - \rho^{-1}\right) < 0. \qquad (7.31)
\end{aligned}$$

In fact, if we denote

$$q \stackrel{\text{def}}{=} -(r-1)r^{-1}b_{max}\bar{a}_{min}\left(r - \rho^{-1}\right),$$

then we have by (7.31) that $q > 0$ and

$$\sum_{y\neq 0} \mathcal{P}_{xaby}\mu(y) \leq (1-q)\mu(x). \qquad (7.32)$$

It is easily seen that (7.31) and hence also the above equation hold also for $x = 0$ and $x = 1$.

Let $\nu(x,a,b) \stackrel{\text{def}}{=} |c(x)| + |d_1(a)| + |d_2(b)|$. Since the immediate cost is polynomially bounded, we obtain by choosing C sufficiently large that $1 + \nu(x,a) < q\mu(x)$ for all states x and actions a. Hence we get

$$1 + \nu(x,a,b) + \sum_{y\neq 0} \mathcal{P}_{xaby}\mu(y) \leq \mu(x).$$

Thus, μ satisfies condition M1(i) in Definition 7.5. Condition M1(ii) clearly holds (since the action space is finite). Condition M1(iii) is obtained by iterating (7.32). □

Remark 7.7 (*Relaxing the stability conditions*)
Note that the condition $b_{max} < a_{min}$ in the above example is a strong stability condition. In practice, this condition can be completely dropped when dealing with uniform Lyapunov functions for the discounted cost (this will be shown in Section 10.5). There are ways to relax the above condition (and in fact the requirement that condition M1(i) holds for all actions) even for the total expected cost and the expected average cost. For more details, see Altman, Hordijk and Spieksma (1997).

Example 7.2 (*Optimal priority assignment*)
Consider N infinite discrete time queues and a Bernoulli arrival processes to each queue. More precisely, in the beginning of each time unit there is a probability $\hat{\lambda}_i$ of an arrival of a packet to queue i. The packets in each queue are served according to the first-in-first-out (FIFO) order. If there is at least one packet in queue i and this queue is being served at time

slot t, then at the end of the time slot the service is completed and the packet is transmitted successfully and leaves the system with probability $\hat{\mu}_i$; otherwise, with probability $1-\hat{\mu}_i$, it stays in the queue. We assume that the arrivals at different queues and at different slots are independent, that the completion of services are independent, and the latter are independent of the arrival processes. This defines an MDP as follows.

- The state space is $\mathbf{X} = \mathbb{N}^{\mathrm{N}}$, where $x \in \mathbf{X}$ stands for the number of packets in each queue after the beginning of a time slot (and after the possible arrivals occurred in that slot).
- The action space is $\mathsf{A} = \{1, \ldots, N\}$; if action a is chosen, then the queue to be served is queue number a. In particular, the available actions at state x are $\mathsf{A}(x) = \{a : x_a > 0\}$. In other words, service can be assigned to non-empty queues only.

This problem corresponds to the assignment of the access to a communication channel among different types of traffic; each traffic type (voice, data, video, etc.) has its own queue. A typical problem is to minimize a weighted sum of the expected delays of non-interactive traffic (such as data transfer) subject to constraints on the expected delays of interactive traffic (voice, video). The fact that service may fail may correspond to a noisy channel.

The immediate costs that appear both in the objective function to be minimized as well as in the constraints are assumed to be linear in the queue sizes, with non-negative coefficients. In other words, $c(x, a)$ has the form

$$c(x, a) = \sum_{i=1}^{N} c_i x_i$$

(they do not depend on a), where $c_i \geq 0$. The costs $d^k(x, a)$ have a similar linear form. The reason for choosing linear costs is that by the well-known Little Theorem, the expected queue lengths are proportional to the expected delays (see e.g., Kleinrock, 1976, p. 17).

We assume that

$$\sum_{i=1}^{N} \frac{\hat{\lambda}_i}{\hat{\mu}_i} < 1.$$

Under this assumption, it has been shown in Makowski and Shwartz (1987) that the queue lengths distribution is tight, and that the queue lengths are uniformly integrable (in time). As will be shown in Section 11.9, this implies the existence of a Lyapunov function for this problem.

Consider the problem of minimizing the total expected cost until the set $\mathcal{M} = \{0\}$ is reached, where 0 corresponds to the state in which all queues are empty. As was shown on p. 145 in Spieksma (1990), a possible choice

of a Lyapunov function μ is given by

$$\mu(x) = \prod_{i=1}^{N} (1 + \varepsilon_i)^{x_i},$$

where $\varepsilon_i, i = 1, \ldots, N$ are some strictly positive real numbers.

Moreover, the function ν in Definition 7.5 can be chosen to be proportional to μ. This means, in particular, that also immediate costs that are polynomially bounded have a uniform Lyapunov function.

In fact, the above uniform Lyapunov function is derived in Chapter 9 of Spieksma (1990) for an even more general control problem, in which packets that are successfully transmitted may be rerouted back to one of the queues with positive probabilities.

It turns out that the constrained problem can be fully solved using a linear program with *finitely many decision variables*, see Altman and Shwartz (1989). For other solution approaches, see Nain and Ross (1986). □

We illustrated in the above examples the usefulness of the approach based on the uniform Lyapunov function. Another example of a routing problem into a two-center open Jackson network can be found in Chapter 9 of Spieksma (1990).

7.7 Contracting MDPs

Let $\mu : \mathbf{X} \to [1, \infty)$ be given. For any functions $q : \mathbf{X} \to \mathbb{R}$, $Q : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$, define their μ-norms

$$\|q\|_\mu = \sup_{x \in \mathbf{X}} \frac{q(x)}{\mu(x)}, \qquad \|Q\|_\mu = \sup_{x \in \mathbf{X}} \frac{\sum_{y \in \mathbf{X}} Q_{xy} \mu(y)}{\mu(x)}. \tag{7.33}$$

It is easily verified that $\|\cdot\|_\mu$ is indeed a norm. In particular, it satisfies $\|Qq\|_\mu \leq \|Q\|_\mu \|q\|_\mu$, and for $Q^1, Q^2 : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$, we have $\left\|Q^1 Q^2\right\|_\mu \leq \left\|Q^1\right\|_\mu \left\|Q^2\right\|_\mu$. We say that q and Q are μ-bounded if $\|q\|_\mu < \infty$ and $\|Q\|_\mu < \infty$, respectively. We define F^μ to be the set of functions from $\mathbf{X}$ to $\mathbb{R}$ having finite μ-norm, and M^μ to be the set of non-negative measures over $\mathbf{X}$ given by $\mathsf{M}^\mu := \{q : \sum_{x \in \mathbf{X}} q(x)\mu(x) < \infty\}$ (we shall use the notation $\langle q, \mu \rangle$ for $\sum_{x \in \mathbf{X}} q(x)\mu(x)$.)

We shall say that a function $f : \mathcal{K} \to \mathbb{R}$ is in $\overline{\mathsf{F}^\mu}$ if the function whose x entry is $\sup_{a \in \mathsf{A}(x)} |f(x, a)|$, is in F^μ. Similarly, a non-negative measure q defined on $\mathcal{K}$ is said to be in $\overline{\mathsf{M}^\mu}$ if the measure $\overline{q}$ is in M^μ, where $\overline{q}(x) := q(x, \mathsf{A}(x))$.

Definition 7.9 *(Contracting MDPs)*
Let $\mathbf{X}'$ and $\mathcal{M}$ be two disjoint sets of states with $\mathbf{X} = \mathbf{X}' \cup \mathcal{M}$. An MDP

is said to be contracting (on $\mathbf{X}'$*) if there exist a scalar* $\xi \in [0,1)$ *(called the contracting factor) and a vector* $\mu : \mathbf{X} \to [1,\infty)$*, such that for all* $x \in \mathbf{X}, a \in \mathsf{A}(x)$*,*

$$\sum_{y \notin \mathcal{M}} \mathcal{P}_{xay}\mu(y) \leq \xi\mu(x). \tag{7.34}$$

When using contracting MDPs, we shall make the following assumptions on the initial distribution, the transition probabilities and the costs:

- $\langle \beta, \mu \rangle < \infty$.
- *The transition probabilities are* μ*-continuous, i.e., if* $a(n) \to a$*, then*

$$\lim_{n\to\infty} \sum_{y \in \mathbf{X}} |\mathcal{P}_{xa(n)y} - \mathcal{P}_{xay}|\, \mu(y) = 0. \tag{7.35}$$

- $c, d^k \in \overline{\mathsf{F}^\mu}, 1 \leq k \leq K$: *they are* μ*-bounded by a constant* $\overline{b} < \infty$. (7.36)

An alternative way to write (7.34) is

$$\sup_{w \in U_D} \|_{\mathcal{M}} P(w)\|_\mu \leq \xi. \tag{7.37}$$

The μ-continuity, defined in (7.35), is related to standard continuity as follows:

Lemma 7.11 (μ*-continuity,* Lemma 5.1 in Spieksma, 1990, p. 96)
The following assertions are equivalent for a matrix $Q(u)$*, with* $u \in U_S$ *and* $\|Q(u)\|_\mu < \infty$*:*
(i) $Q(u)$ *is* μ*-continuous on* U_S*,*
(ii) $Q(u)$ *and* $Q(u)\mu$ *are pointwise continuous on* U_S*,*
(iii) For any pointwise converging sequence q_n *of* μ*-bounded functions with* $\sup_{n\in\mathbb{N}} \|q_n\|_\mu < \infty$*, and for any converging sequence of stationary policies* u_n *with a limit* u^**,*

$$\lim_{n\to\infty} [Q(u_n)q_n]_x = [Q(u^*)q^*]_x, \qquad \forall x \in \mathbf{X}.$$

The μ-continuity of the transition probabilities imply also the following.

Lemma 7.12 (μ*-continuity and uniform integrability*)
The μ*-continuity of the transition probabilities, defined in (7.35), together with the assumption that*

$$\sum_{y \notin \mathcal{M}} \mathcal{P}_{xay}\mu(y) \leq \xi'(x)\mu(x) \tag{7.38}$$

for some $\xi(\cdot)' > 0$*, implies that*

(i) For any fixed x, $\{\mathcal{P}_{xay}\}_a$ are integrable with respect to μ uniformly in a, i.e., for any sequence of compact sets $\mathbf{X}_n$ increasing to $\mathbf{X}$,

$$\lim_{n\to\infty} \sup_a \sum_{y\notin\mathbf{X}_n} \mathcal{P}_{xay}\mu(y) = 0.$$

(ii) Let $u^n \to u$ (weakly) where u and u^n are probability measures over $\mathsf{A}(x)$. Then

$$\lim_{n\to\infty} \sum_{y\in\mathbf{X}} |P_{xy}(u^n) - P_{xy}(u)|\, \mu(y) = 0. \tag{7.39}$$

Proof. (i) follows directly from Lemma 17.4 (ii) in the Appendix.
(ii) The weak convergence of u^n implies that $P_{xy}(u^n)$ converges to $P_{xy}(u)$ for all x and y. (i) implies that $P_{x\cdot}(u^n)$ are integrable with respect to μ, uniformly in n, so that

$$\lim_{n\to\infty} \sum_{y\in\mathbf{X}'} P_{xy}(u^n)\mu(y) = \sum_{y\in\mathbf{X}'} P_{xy}(u)\mu(y),$$

see Lemma 17.4. This implies (7.39) (by Scheffé's Lemma, see e.g., Williams, 1992, p. 55). □

Lemma 7.13 (*Rate of convergence*)
Consider the contracting MDP. Then

$$p^u_x(t;\mathbf{X}') \le \sum_{y\in\mathbf{X}'} p^u_x(t;y)\mu(y) \le \mu(x)\xi^{t-1} \tag{7.40}$$

(*thus the μ-norm of $p^u_{(\cdot)}(t)$ converges to 0 at a geometric rate, uniformly over all $u \in U$, i.e., for any $x \in \mathbf{X}$*) *Moreover,*

$$p^u_\beta(t;\mathbf{X}') \le \sum_{y\in\mathbf{X}'} p^u_\beta(t;y)\mu(y) \le \langle\beta,\mu\rangle\xi^{t-1}, \tag{7.41}$$

and

$$\sum_{t=1}^{\infty} p^u_\beta(t;\mathbf{X}') \le \sum_{t=1}^{\infty}\sum_{y\in\mathbf{X}'} p^u_\beta(t;y)\mu(y) \le \frac{\langle\beta,\mu\rangle}{1-\xi}, \tag{7.42}$$

which implies that the MDP is $\mathbf{X}'$-absorbing.

Proof. Choose any $u \in U$ and let $v = v(u)$ be the corresponding Markov policy given in (6.11). Viewing $p^u_{(\cdot)}(t;\cdot)$ as a function $\mathbf{X}\times\mathbf{X} \to \mathbb{R}$, we have

$$\|p^u_\cdot(t;\cdot)\|_\mu \le \|_{\mathcal{M}}P(v_1)\|_\mu \, \|_{\mathcal{M}}P(v_2)\|_\mu \cdots \|_{\mathcal{M}}P(v_{t-1})\|_\mu \le \xi^{t-1},$$

which implies (7.40) and (7.41). (7.42) easily follows. □

Hence, contracting MDPs are a subclass of absorbing MDPs, which are a subclass of transient MDPs. The converse need not hold; if $\mathbf{X} = \mathbb{N}$ and $P_{n,n+1}(w) = 1$ for some $w \in U_S$, then w is transient but non-absorbing.

For the case of finite state and action spaces, however, the converse holds, and any transient policy is contracting, see Kallenberg (1983).

Lemma 7.1 can be strengthened:

Lemma 7.14 *(Uniqueness of a bounded solution)*
Consider a contracting MDP. Fix a stationary transient policy w on $\mathbf{X}'$. Fix some initial distribution β such that $\langle \beta, \mu \rangle < \infty$. Then

$$f(x) = \sum_{t=1}^{\infty} p_\beta^w(t; x)$$

is the unique μ-bounded solution of

$$f = \beta + {}_{\mathcal{M}} f P(w). \tag{7.43}$$

Proof. Let f' be some μ-bounded solution of (7.43). Iterating (7.43), we get:

$$\begin{aligned} f'(y) &= \beta(y) + p_\beta^w(2; y) + \sum_{x \in \mathbf{X}'} f'(x)[{}_{\mathcal{M}} P^2(w)]_{xy} \qquad (7.44) \\ &= \beta(y) + p_\beta^w(2; y) + p_\beta^w(3; y) + \sum_{x \in \mathbf{X}'} f'(x)[{}_{\mathcal{M}} P^3(w)]_{xy} \\ &= \ldots = \beta(x) + p_\beta^w(2; x) + \ldots + p_\beta^w(t; x) + \sum_{x \in \mathbf{X}'} f'(x)[{}_{\mathcal{M}} P^t(w)]_{xy}. \end{aligned}$$

We have (as in Lemma 7.13)

$$\left| \sum_{x \in \mathbf{X}'} f'(x)[{}_{\mathcal{M}} P^t(w)]_{xy} \right| \leq \mu(y) \left\| f' \right\|_\mu \left\| {}_{\mathcal{M}} P^t(w) \right\|_\mu \leq \mu(y) \left\| f' \right\|_\mu \xi^t \to 0.$$

The proof is established by taking the limit as $t \to \infty$ in (7.44). □

Remark 7.8 (*Relation to other definitions*)
The definition introduced in this section for contracting MDPs is taken from Dekker and Hordijk (1988) and Spieksma (1990). It is weaker than (and thus implies) previous definitions, such as the one by Wessels (1977), who considers a similar definition but with an empty set $\mathcal{M}$. Allowing a non-empty set, $\mathcal{M}$, turns out to be especially important in the average cost case.

Theorem 7.5 (*Contracting MDP implies a uniform Lyapunov function*)
Assume that an MDP is contracting. Then it has a uniform Lyapunov function with the same function μ, up to a multiplicative constant. Moreover, ν can be chosen proportional to μ.

Proof. Assume that the MDP is contracting with a geometric Lyapunov function μ. For any constant C, and for all $(x, a) \in \mathcal{K}$,

$$|c(x, a)| + 1 + \sum_{y \notin \mathcal{M}} \mathcal{P}_{xay} C \mu(y) \leq \overline{b} \mu(x) + 1 + \sum_{y \notin \mathcal{M}} \mathcal{P}_{xay} C \mu(y)$$

$$\leq \quad \overline{b}\mu(x) + 1 + \xi C\mu(x) = C\mu(x) + 1 + (\overline{b} + (\xi - 1)C)\mu(x).$$

For all C sufficiently large, $1 + (\overline{b} + (\xi - 1)C)\mu(x)$ is negative, so that $C\mu$ satisfies M(i) in Definition 7.5.
M(ii) follows from the μ-continuity of the transition probabilities (see (7.35)).
M(iii) follows since $\|[{}_{\mathcal{M}}P(u)]^n\|_\mu \leq \xi^n$ (see (7.37)). □

We note that in all the examples presented in Section 7.6, the MDPs were in fact contracting.

CHAPTER 8

The total cost: occupation measures and the primal LP

We study in this chapter the total cost criterion for $\mathbf{X}'$-transient MDPs, i.e., the total expected cost until a set of states $\mathcal{M} = \mathbf{X} \setminus \mathbf{X}'$ is reached.

8.1 Occupation measure

For any given initial distribution β and policy u, define the occupation measure $f_{tc}(\beta, u)$ on $(\mathcal{K}, \mathbb{K})$ related to the total cost criterion by

$$f_{tc}^n(\beta, u; \overline{\mathcal{K}}) = \sum_{t=1}^{n} p_\beta^u(t; \overline{\mathcal{K}}), \qquad f_{tc}(\beta, u; \overline{\mathcal{K}}) = \sum_{t=1}^{\infty} p_\beta^u(t; \overline{\mathcal{K}}), \qquad \overline{\mathcal{K}} \subset \mathcal{K}.$$

With some abuse of notation, we use f_{tc} also for the restriction over $\mathbf{X}'$, i.e., $f_{tc}(\beta, u; \mathcal{X}) := f_{tc}(\beta, u; \cup_{x \in \mathcal{X}}(x, \mathsf{A}(x)))$. Define

$$\mathbf{L}_{\overline{U}}(\beta) = \bigcup_{u \in \overline{U}} \{f_{tc}(\beta, u)\} \text{ for any class of policies } \overline{U}, \tag{8.1}$$

$$\mathbf{Q}_{tc}(\beta) = \tag{8.2}$$
$$\left\{ \begin{array}{c} \rho \in M(\mathcal{K}) : \sum_{y \in \mathbf{X}} \int_{\mathsf{A}(y)} \rho(y, da)(\delta_x(y) - \mathcal{P}_{yax} 1\{x \in \mathbf{X}'\}) = \beta(x) \\ \\ \text{and } \rho(x, \mathsf{A}(x)) < \infty \text{ for all } x \in \mathbf{X}, \end{array} \right\}$$

where $M(\mathcal{K})$ is the set of non-negative measures over $\mathcal{K}$ and δ_x is the Dirac probability measure concentrated on x. We set $\mathbf{L}(\beta) = \mathbf{L}_U(\beta) \cup \mathbf{L}_{\overline{M}(U_M)}(\beta)$.

Define

$$\mathbf{Q}_{tc}^{\mathrm{b}}(\beta) \stackrel{\mathrm{def}}{=} \text{ the subset of finite measures among } \mathbf{Q}_{tc}(\beta),$$

$$\mathbf{Q}_{tc}^{\nu}(\beta) \stackrel{\mathrm{def}}{=} \{q \in \mathbf{Q}_{tc}^{\mathrm{b}}(\beta) : \langle q, \nu \rangle < \infty\}, \qquad \mathbf{Q}_{tc}^{\mu}(\beta) \stackrel{\mathrm{def}}{=} \mathbf{Q}_{tc}(\beta) \cap \overline{\mathsf{M}^\mu}. \tag{8.3}$$

The two definitions in (8.3) are for MDPs with a uniform Lyapunov function and for contracting MDPs (the parameters ν and μ are introduced in Definitions 7.5 and 7.9, respectively).

For any sets B, B_1, B_2 of measures (equipped with a given topology) on some measurable space, define

- $\overline{co}B :=$ the closed convex hull of a set B (see definition in Dunford and Schwartz, 1988, p. 414);
- $\min B :=$ the set of minimal elements in B, i.e., $\rho \in \min B$ if there does not exist some $\rho' \in B$, $\rho' \leq \rho$ (i.e., $\rho'(\mathcal{A}') \leq \rho(\mathcal{A}')$ for any $\mathcal{A}'$), such that $\rho'(\mathcal{A}) < \rho(\mathcal{A})$ for some $\mathcal{A}$;
- $B_1 \prec B_2$ if $\forall \rho_2 \in B_2$ there exists $\rho_1 \in B_1$, such that $\rho_1 \leq \rho_2$.

Definition 8.1 (*Completeness for the total cost criterion*)
A class of policies $\overline{U}$ is called complete for the total cost criterion (for a given initial distribution β) if $\mathbf{L}_{\overline{U}}(\beta) = \mathbf{L}(\beta)$. It is called weakly complete if $\mathbf{L}_{\overline{U}}(\beta) \prec \mathbf{L}(\beta)$.

Theorem 8.1 (*Completeness of stationary policies*)
(*i*) *Consider an* $\mathbf{X}'$*-transient MDP. Then the set of stationary policies is weakly complete.*
(*ii*) *If the MDP is absorbing to* $\mathcal{M}$*, then the set of stationary policies is complete.*

Proof. Choose a policy $u \in U$ and let w be a stationary policy satisfying

$$w_y(\mathcal{A}) = \frac{f_{tc}(\beta, u; y, \mathcal{A})}{f_{tc}(\beta, u; y)}, \ y \in \mathbf{X}, \mathcal{A} \subset \mathsf{A}(y) \tag{8.4}$$

whenever the denominator is non-zero. (When it is zero, $w_y(\cdot)$ is chosen arbitrarily.) We show that $f_{tc}(\beta, w) \leq f_{tc}(\beta, u)$. For any $x \in \mathbf{X}$,

$$\begin{aligned}
f_{tc}(\beta, u; x) &= \beta(x) + \sum_{t=2}^{\infty} p^u_\beta(t, x) \\
&= \beta(x) + \sum_{t=2}^{\infty} \sum_{y \in \mathbf{X}} \int_{\mathsf{A}(y)} p^u_\beta(t-1; y, da) \mathcal{P}_{yax} 1\{x \in \mathbf{X}'\} \\
&= \beta(x) + \sum_{y \in \mathbf{X}} \int_{\mathsf{A}(y)} f_{tc}(\beta, u; y, da) \mathcal{P}_{yax} 1\{x \in \mathbf{X}'\} \qquad (8.5) \\
&= \beta(x) + \sum_{y \in \mathbf{X}} f_{tc}(\beta, u; y) \int_{\mathsf{A}(y)} \mathcal{P}_{yax} w_y(da) 1\{x \in \mathbf{X}'\} \\
&= \beta(x) + \sum_{y \in \mathbf{X}} f_{tc}(\beta, u; y)_{\mathcal{M}} P_{yx}(w). \qquad (8.6)
\end{aligned}$$

Hence, by Lemma 7.1 (i), $f_{tc}(\beta, w; x) \leq f_{tc}(\beta, u; x)$ for all $x \in \mathbf{X}$. This implies by the definition of w that $f_{tc}(\beta, w) \leq f_{tc}(\beta, u)$, so that the set of stationary policies is weakly complete.

(ii) Follows from Lemma 7.1 (ii) and (8.6). □

Next we present some examples that show that the stationary policies are not complete in the general transient case.

Example 8.1 *(Incompleteness of stationary policies)*
Consider the following MDP:

- State space: $\mathbf{X} = \{1, 2\}$.
- Actions: $\mathsf{A}(1) = \{a\}$; $\mathsf{A}(2) = \{a, b\}$.
- Transition probabilities: $\mathcal{P}_{1,a,1} = \mathcal{P}_{2,a,1} = 1$, $\mathcal{P}_{2,b,2} = 1$.

Let $\mathcal{M} = \{1\}$; we consider the occupation measure corresponding to the total time until hitting $\mathcal{M}$.

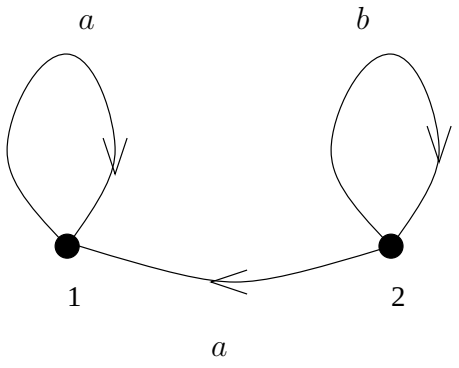

Figure 8.1 *Incompleteness of stationary policies (Example 8.1)*

Let $g(p)$ be the stationary policy that chooses action b (at state 2) with probability p. We shall take 2 as initial state and consider the occupation measure until state 1 is reached. Then

$$f_{tc}(2, g(p); 2) = \begin{cases} \infty & \text{if } 0 \le p < 1, \\ 0 & \text{if } p = 1. \end{cases}$$

However, the Markov policy $u(p)$ that uses action a at time 1 with probability p, and then always uses actions b at state 2, achieves $f_{tc}(2, u(p); 2) = 1 - p, 0 \le p \le 1$. To conclude, the values of $f_{tc}(2, u; 2)$ achievable by the stationary policies are two isolated points: $\{0, \infty\}$, whereas those achievable by the Markov policy $u(p)$ are the whole interval $[0, \infty]$.

□

In the above example, the MDP is not transient: under action b we remain forever at state b. Next, we present an example due to Feinberg and Sonin (1996, Section 4), where the MDP is transient and the stationary policies are not complete.

Example 8.2 *(Incompleteness of stationary policies, a transient MDP)*
Consider the following MDP:

- State space: $\mathbf{X} = \{-1, 0, 1, 2, \ldots\}$; let $\mathbf{X}' = \{0, 1, 2, \ldots\}$.

- Actions: $\mathsf{A} = \{f, b\} = \{forward, backword\}$; $\mathsf{A}(0) = \{f\}, \mathsf{A}(x) = \mathsf{A}$ for $x > 0$.
- Transitions: $\mathcal{P}_{x,f,x+1} = \gamma_{x+1}, \mathcal{P}_{x,f,x-1} = \gamma_x$, $x = 1, 2, \ldots$; the transition probabilities to all other states in $\mathbf{X}'$ are zero.

Let $(d_x)_{x=1}^{\infty}$ be a sequence of positive integers. d defines a policy u in the following way. If a state $x = 1, 2, \ldots$ is visited for the mth time, then policy u selects b if $m \leq d_x$, and selects f if $m > d_x$. We consider 0 as the initial state.

Feinberg and Sonin chose $d_x = 2^{x-1}$, $x = 1, 2, \ldots$. The γ_x, $x = 1, 2, \ldots$ were chosen to satisfy the following constraints:
(i) $P_0^u(T = \infty) > 0.9$,
(ii) $\gamma_x \leq \gamma_{x+1}$,
(iii) $\gamma_x < 1$.

For these values, they show that

$$f_{tc}(0, u; x) \geq 0.9(32^{x-1} + 1) > 5$$

for $x \geq 3$. If a stationary policy w achieving the same occupation measures as u existed, it would necessarily satisfy (8.4). However, Feinberg and Sonin show that for such w, $f_{tc}(0, w; x) \leq 5$. We conclude that the stationary policies are not complete.

That this MDP is indeed transient follows from the fact that for each policy u and for each state x, the probability to return to state x (before being absorbed in state -1) is bounded above by γ_x. Hence,

$$f_{tc}(y, u; x) \leq \frac{1}{1 - \gamma_x} < \infty$$

for all $y \in \mathbf{X}$ and for all $x = 1, 2, \ldots$. Finally, for $x = 0$, we have $f_{tc}(y, u; 0) \leq f_{tc}(y, u; 1)$.

8.2 Continuity of occupation measures

Definition 8.2 (*μ-continuity of policies*)
Consider some $\overline{U} \subset U_M$ and $Q : \overline{U} \times \mathbf{X} \to \mathbb{R}$. Q is said to be μ-continuous on $\overline{U}$ if for any converging sequence $u(n) \in \overline{U}$ with limit $u \in \overline{U}$

$$\lim_{n \to \infty} \sum_{y \in \mathbf{X}} |Q(u(n), y) - Q(u, y)| \mu(y) = 0.$$

(The convergence of policies is understood with respect to the topology over U_M defined in Section 6.3.)

Definition 8.3 (*Lower semi-continuity*)
Consider a class of policies $\overline{U}$ and some Borel space Y. Consider a measured-value function $f : \overline{U} \to M(Y)$, where $f(u; \mathcal{Y})$ stands for the measure it assigns to a Borel set $\mathcal{Y} \subset Y$ for $u \in \overline{U}$. It is said to be weakly lower

semi-continuous (l.s.c.) on $\overline{U}$ if for any $u \in \overline{U}$ and any sequence $u^n \in \overline{U}$ that weakly converges to u, we have

$$\underline{\lim}_{n\to\infty} \langle f(u^n), c\rangle \geq \langle \underline{\lim}_{n\to\infty} f(u^n), c\rangle,$$

for any non-negative function $c : Y \to \mathbb{R}$ which is lower semi-continuous.

Lemma 8.1 (*Continuity properties of* f_{tc})
(*i*) *For transient MDPs, the measures* $f_{tc}(\beta,\cdot)$ *defined on* $\mathcal{K}$ *are weakly l.s.c. over* U_M. *The state occupation measures* $f_{tc}(\beta,u;\cdot)$ *are weakly continuous over* U_M. *The above statements also hold for the policies* $\overline{M}(U_M)$.
(*ii*) *Consider an MDP with a uniform Lyapunov function. For any continuous function* $c' : \mathcal{K} \to \mathbb{R}$ *satisfying* $|c'| \leq \nu + 1$, $\langle f_{tc}(\beta,\cdot), c'\rangle$ *are continuous over* U_M (*thus the measures* $f_{tc}(\beta,\cdot)$, *defined over* $\mathcal{K}$, *are weakly continuous over* U_M). *The above statement holds also for the policies* $\overline{M}(U_M)$.
(*iii*) *Consider a contracting MDP. Then the state occupation measures* $f_{tc}(\beta,\cdot)$ (*defined on* $\mathbf{X}$) *are* μ*-continuous over* U_M.

Proof. (i) Assume that $u^n \to u$, where $u^n, u \in U_M$ (i.e., for any $x \in \mathbf{X}$ and t, $u_t^n(x)$ converges weakly to $u_t(x)$). Then ${}_{\mathcal{M}}P_{xy}(u_1^n) \to {}_{\mathcal{M}}P_{xy}(u_1)$ for all $x, y \in \mathbf{X}$. By the bounded convergence theorem, this implies that $\sum_x \beta(x)P_{xy}(u_1^n) \to \sum_x \beta(x)P_{xy}(u_1)$ for all $y \in \mathbf{X}'$. Moreover, the m step probabilities also converge, i.e., for all integers m:

$$\begin{aligned}\lim_{n\to\infty} p_\beta^{u^n}(m;y) &= \lim_{n\to\infty} \sum_{x\in\mathbf{X}} \beta(x)[{}_{\mathcal{M}}P(u_1^n){}_{\mathcal{M}}P(u_2^n)\cdots{}_{\mathcal{M}}P(u_{m-1}^n)]_{xy} \\ &= \sum_{x\in\mathbf{X}} \beta(x)[{}_{\mathcal{M}}P(u_1){}_{\mathcal{M}}P(u_2)\cdots{}_{\mathcal{M}}P(u_{m-1})]_{xy} = p_\beta^u(m;y), \quad (8.7)\end{aligned}$$

for all $y \in \mathbf{X}$. (8.7) is established by induction. It holds for $m = 1$. Assume it holds for arbitrary m. Consider the probability measures over $\mathbf{X}$: $\nu(n) := p_\beta^{u^n}(m;\cdot)$ and $\nu := p_\beta^u(m;\cdot)$, and let $q_y(n)$ and q_y be the y column of ${}_{\mathcal{M}}P(u_m^n)$ and ${}_{\mathcal{M}}P(u_m)$, respectively. Then,

$$p_\beta^{u^n}(m+1;y) = \int_{\mathbf{X}'} q_y(n)d\nu(n), \qquad p_\beta^u(m+1;y) = \int_{\mathbf{X}'} q_y d\nu.$$

The entries of q_y are bounded by 1, so by applying the generalized dominance convergence theorem (Royden, 1988, Proposition 11.18), we get

$$p_\beta^{u^n}(m+1;y) \to p_\beta^u(m+1;y),$$

from which (8.7) follows.

Denote

$$Z^n(m,y) := \int_{\mathsf{A}(y)} p_\beta^{u^n}(m;y,da)c'(y,a),$$

$$Z(m,y) := \int_{\mathsf{A}(y)} p_\beta^u(m;y,da)c'(y,a),$$

where c' is an arbitrary non-negative l.s.c. function. Since $p_\beta^{u^n}(m)$ converges weakly to $p_\beta^u(m)$ (from (8.7)) for all m and x, it follows that

$$\underline{\lim}_{n\to\infty} Z^n(m,y) \geq Z(m,y),$$

see Doob (1994, p. 133) (in the related theorem in Doob, the cost c' is assumed to be bounded; however, it can easily be seen that only boundedness from below is used in the proof of that theorem). Applying Fatou's Lemma (Royden, 1988, Proposition 11.17) with respect to the (infinite) measure over $\mathbb{N}\times\mathbf{X}$ generated by $\mu_n(m,y)=\mu(m,y)=1\{y\in\mathbf{X}'\}$ for all m and y, we obtain

$$\underline{\lim}_{n\to\infty}\langle f_{tc}(\beta,u^n),c'\rangle = \underline{\lim}_{n\to\infty}\sum_{m,y} Z^n(m,y) \geq \sum_{m,y} Z(m,y) = \langle f_{tc}(\beta,u),c'\rangle,$$

which concludes the proof of the first statement in (i). (The equalities above follow from the fact that the integrand is non-negative, see Tonelli's Theorem in Royden, 1988, Theorem 12.20.) This implies all the statements in (i) concerning the lower semi-continuity in U_M.

The lower semi-continuity on $\overline{M}(U_M)$ follows from that on U_M, see Doob (1994, p. 133).

(ii) Due to Theorem 7.4, we can conclude that $C_{tc}(\beta,u)$ is continuous over U_M (property N5′ there). One can easily show that for any policy u, $C_{tc}(\beta,u)=\langle f_{tc}(\beta,u),c\rangle$ (for a detailed proof of this, see Theorem 8.3). Hence the statements are established for the Markov policies.

Let γ^n be a sequence in $M_1(U_M)$ converging weakly to some γ. Let $\hat{\gamma}^n$ and $\hat{\gamma}$ be the corresponding policies in $\overline{M}(U_M)$. Then the continuity on $\overline{M}(U_M)$ follows since $f_{tc}(\beta,\cdot)$ are bounded and continuous functions on U_M, so that the weak convergence of γ^n implies (Billingsley, 1968, Theorem 2.1):

$$\begin{aligned}&\lim_{n\to\infty} f_{tc}(\beta,\hat{\gamma}^n)\\ &\quad= \lim_{n\to\infty}\langle\gamma^n, f_{tc}(\beta,\cdot)\rangle = \langle \lim_{n\to\infty}\gamma^n, f_{tc}(\beta,\cdot)\rangle = f_{tc}(\beta,\hat{\gamma}).\end{aligned}$$

(iii) Contracting MDPs have a uniform Lyapunov function with the same μ, and for which ν can be chosen proportional to μ (Theorem 7.5). This implies by (ii) that

$$\lim_{n\to\infty}\sum_{y\in\mathbf{X}} f_{tc}(\beta,u^n;y)\mu(y) = \sum_{y\in\mathbf{X}} f_{tc}(\beta,u;y)\mu(y).$$

The result now follows from Scheffé's Lemma, see e.g., Williams (1992, p. 55). □

As an example where the occupation measures are l.s.c. in the policies, and not continuous, consider the following example.

Example 8.3 Consider $\mathbf{X}$ to be the natural numbers, let $\mathcal{M}=\{0\}$ and

assume that $\mathsf{A}(x) = \{a, b\}$ for all x. For any state $x > 1$, action a results in a transition to state $x + 1$, and action b results in a transition to state 1. From state 1 we leave the set $\mathbf{X}'$ in one step. Assume that we start at some state $y > 1$. For any n, a policy u^n that chooses action b for the first time at some instant larger than n achieves $f_{tc}(\beta, u^n; 1) = 1$. But the policy u that is obtained as the weak limit of u^n, i.e., the policy that always chooses action a, achieves $f_{tc}(\beta, u; 1) = 0$.

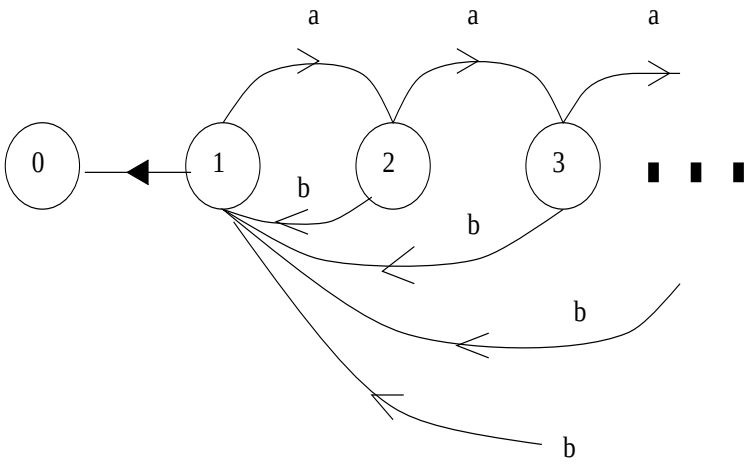

Figure 8.2 *Discontinuity of f_{tc} (Example 8.3)*

Note that the condition

$$\lim_{n\to\infty} \sup_{u\in U_M} \sum_{t=n}^{\infty} p_\beta^u(t; \mathbf{X}) = 0 \tag{8.8}$$

is not satisfied. Condition (8.8) would hold if the MDP had a uniform Lyapunov function since

$$\sum_{t=n}^{\infty} p_\beta^u(t; \mathbf{X}) \le \sum_{t=n}^{\infty} E_\beta^u(1 + \nu(X_t, A_t))1\{T > t\} \le E_\beta^u \hat{M}(X_n)1\{T > n\}$$

(see definitions in Section 7.4), which tends to zero as $n \to \infty$, uniformly in u. This follows from property N4 in Theorem 7.4. In our example we have, however, $\sum_{t=n}^{\infty} p_\beta^{u^n}(t; 1) = 1$. Thus

$$\lim_{n\to\infty} \sup_{u\in U_M} \sum_{t=n}^{\infty} p_\beta^u(t; \mathbf{X}) \ge 1.$$

□

In the above example, the MDP is transient. However, continuity may fail also for absorbing MDPs. We illustrate this in the following counter example, due to Fisher and Ross (1968), which was further studied in Spieksma (1990).

Example 8.4 (Fisher and Ross' example)
Consider the following (unichain) MDP:

- State space: $\mathbf{X} = \{0, 1, 1', 2, 2', 3, 3', \ldots\}$.
- Actions: a single (trivial) action at states 0 and i', actions a and b at state i.
- Transition probabilities:

$$\begin{aligned}
\mathcal{P}_{0,a,i} &= \mathcal{P}_{0,a,i'} = \frac{3}{2}\left(\frac{1}{4}\right)^i, \qquad i > 0, \\
\mathcal{P}_{i,a,0} &= 1 - \mathcal{P}_{i,a,i'} = \left(\frac{1}{2}\right)^i, \\
\mathcal{P}_{i,b,0} &= 1 - \mathcal{P}_{i,b,i+1} = \frac{1}{2}, \\
\mathcal{P}_{i',a,0} &= 1 - \mathcal{P}_{i',a,i'} = \left(\frac{1}{2}\right)^i.
\end{aligned}$$

Let $u(n)$ be the stationary policy that chooses action b for $i \leq N$, and action a for $i > N$. Let u be the policy that chooses action b for all i.

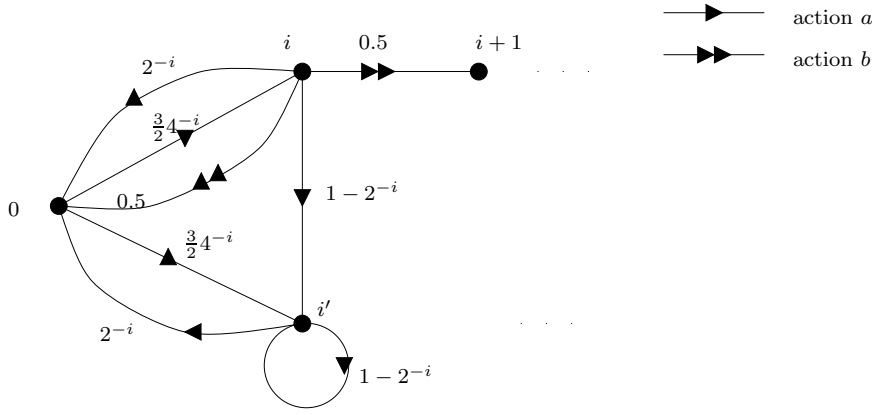

Figure 8.3 *Discontinuity of f_{tc}, absorbing MDP (Example 8.4)*

Fisher and Ross computed the total expected time between two consecutive visits of the state 0:

$$E_0^{u(n)} T = f_{tc}(0, u(n); \mathbf{X}') = 5 - \sum_{j=1}^{n-1} \frac{3}{2}\left(\frac{1}{4}\right)^j \left(\frac{1}{2}\right)^{n-j-1} - 3\sum_{j=n}^{\infty}\left(\frac{1}{4}\right)^j.$$

On the other hand, $E_0^u T = 7/2$ (this follows (11.17) in Spieksma (1990)). Hence f_{tc} is discontinuous over the stationary policies in this example:

$$5 = \lim_{n\to\infty} E_0^{u(n)} T > E_0^u T = \frac{3}{2}.$$

□

Lemma 8.2 (*Splitting in a state*)
Choose $w \in U_S$ and a state y. Define $w^a \in U_S$ to be the policy that always chooses action a when in state y, and otherwise behaves exactly like w. Then, there exists a probability measure γ over $\mathsf{A}(y)$ such that

$$f_{tc}(\beta, w) = \int_{\mathsf{A}(y)} \gamma(da) f_{tc}(\beta, w^a).$$

Proof. Define the stopping times $\mathcal{T}(y) \stackrel{\text{def}}{=} \inf_{r>1}\{X_r = y\}$, $y \in \mathbf{X}$, with the convention that $\inf\{\emptyset\} = \infty$. Define the probability of ever reaching state y from state x before hitting $\mathcal{M}$:

$$\overline{p}(u; x, y) := P_x^u(\mathcal{T}(y) < T_{\mathcal{M}}).$$

Define γ in the following way: for any $\mathcal{A} \subset \mathsf{A}(y)$,

$$\gamma(\mathcal{A}) \stackrel{\text{def}}{=} \int_{\mathcal{A}} \frac{w_y(da)(1 - \overline{p}(w^a; y, y))}{1 - \overline{p}(w; y, y)}.$$

It follows from standard properties of Markov chains (see Kemeney *et al.*, 1976, Corollary 4-20) that

$$f_{tc}(x, w; y) = 1\{x = y\} + \overline{p}(w; x, y) f_{tc}(y, w; y).$$

By setting $x = y$, we get

$$f_{tc}(y, w; y) = \frac{1}{1 - \overline{p}(w; y, y)} = \frac{\int w_y(da)}{1 - \overline{p}(w; y, y)}$$

$$= \int \frac{w_y(da)(1 - \overline{p}(w^a; y, y))}{1 - \overline{p}(w; y, y)} f_{tc}(y, w^a; y) = \int_{\mathsf{A}(y)} \gamma(da) f_{tc}(y, w^a; y),$$

which establishes the proof for the case $x = y$. For other x,

$$\begin{aligned} f_{tc}(x, w; y) &= \overline{p}(w; x, y) f_{tc}(y, w; y) \\ &= \int_{\mathsf{A}(y)} \gamma(da) \overline{p}(w; x, y) f_{tc}(y, w^a; y) \\ &= \int_{\mathsf{A}(y)} \gamma(da) f_{tc}(x, w^a; y). \end{aligned}$$

□

8.3 More properties of MDPs*

Lemma 8.3 (*Boundedness of the life-time*)
(*i*) *For* $\mathbf{X}'$*-transient MDPs,* $\sup_{u\in U} f_{tc}(\beta, u; y) < \infty$ *for all* $y \in \mathbf{X}$.
(*ii*) *For an MDP absorbing to* $\mathcal{M} = \mathbf{X} \setminus \mathbf{X}'$, $\sup_{u\in U} f_{tc}(\beta, u; \mathbf{X}) < \infty$.
(*iii*) *For a contracting MDP,* $\sup_{u\in U} \sum_{y\in\mathbf{X}} f_{tc}(\beta, u; y)\mu(y) < \infty$, *and* $f_{tc}(\beta, u; \cdot)$ *are integrable w.r.t.* μ *uniformly over all policies.*

Proof. (i) and (ii) follow from a result on positive dynamic programming, see p. 108 in Dynkin and Yushkevich (1979). It states the following: given a non-negative cost c, for any $\varepsilon > 0$, there exists some policy w such that

$$\sup_{u\in U} C_{tc}(x, u) < C_{tc}(x, w) + \varepsilon, \quad \forall x \in \mathbf{X}. \tag{8.9}$$

Hence

$$\sup_{u\in U} C_{tc}(\beta, u) < C_{tc}(\beta, w) + \varepsilon. \tag{8.10}$$

By choosing $c(x, a) = 1\{x = y\}$, we obtain (i). By considering $c(x, a) = 1$, we obtain (ii). The first part of (iii) was established in (7.42), and can be obtained alternatively by (8.10) with $c(x, a) = \mu(x)$ (see Spieksma, 1990, Lemma 5.3(ii) and its proof). For the uniform integrability, we shall restrict ourselves, without loss of generality, to Markov policies. The uniform integrability can be obtained from the continuity of $f_{tc}(\beta, u)$, Lemma 8.1 (iii) and Lemma 17.4(ii) from the appendix. □

The statements of the above theorem also hold for mixed policies due to Theorem 6.1. It is shown in Theorem 13.7 in Hordijk (1977, p. 122) that a relation similar to (8.9) holds with $w \in U_S$: for any $\varepsilon > 0$ there exists $w \in U_S$ such that $\sup_{u\in U} C_{tc}(x, u)(1 - \varepsilon) \le C_{tc}(x, w)$, $\forall x \in \mathbf{X}$, which implies that

$$\sup_{u\in U} C_{tc}(\beta, u)(1 - \varepsilon) < C_{tc}(\beta, w). \tag{8.11}$$

Theorem 6.1 and (8.11) imply the following:

Corollary 8.1 (*Sufficient conditions for transient and absorbing MDPs*)
A sufficient condition for an MDP to be $\mathbf{X}'$*-transient* (*resp.,* $\mathbf{X}'$*-absorbing*) *is that every stationary policy is* $\mathbf{X}'$*-transient* (*resp.,* $\mathbf{X}'$*-absorbing*).

8.4 Characterization of the sets of occupation measure

Theorem 8.2 (*Characterization of the sets of occupation measure*)
(*i*) *For transient MDPs,* $\mathbf{L}(\beta)$ *is convex, and*

$$\min \mathbf{Q}_{tc}(\beta) = \mathbf{L}_{U_S}(\beta) \prec \mathbf{L}_{U_M}(\beta) = \mathbf{L}(\beta) \subset \mathbf{Q}_{tc}(\beta).$$

(*ii*) *For MDPs with uniform Lyapunov functions,* $\mathbf{L}_{U_S}(\beta)$ *is convex and compact, and satisfies*

$$\mathbf{L}_{\mathcal{U}}(\beta) = \mathbf{L}(\beta) = \mathbf{L}_{U_S}(\beta) = \overline{co}\mathbf{L}_{U_D}(\beta) = \mathbf{Q}^{\mathrm{b}}_{tc}(\beta) = \mathbf{Q}^{\nu}_{tc}(\beta) = \min \mathbf{Q}_{tc}(\beta).$$

(iii) For contracting MDPs, $\mathbf{L}_{U_S}(\beta)$ is convex and compact, and satisfies

$$\mathbf{L}_{\mathcal{U}}(\beta) = \mathbf{L}(\beta) = \mathbf{L}_{U_S}(\beta) = \overline{co}\mathbf{L}_{U_D}(\beta) = \mathbf{Q}^{\mu}_{tc}(\beta) = \min \mathbf{Q}_{tc}(\beta).$$

Proof. (i) Theorem 6.1 implies that $\mathbf{L}(\beta) = \mathbf{L}_{U_M}(\beta)$ is convex. The weak completeness of $\mathbf{L}_{U_S}(\beta)$ was established in Theorem 8.1. That $\mathbf{L}(\beta) \subset \mathbf{Q}_{tc}(\beta)$ follows from (8.5). Finally, we show that $\mathbf{L}_{U_S}(\beta) = \min \mathbf{Q}_{tc}(\beta)$. For any $\rho \in \mathbf{Q}_{tc}(\beta)$, define $w(\rho)$ to be any stationary policy such that $w_y(\mathcal{A}) = \rho(y, \mathcal{A})[\rho(y, \mathsf{A}(y))]^{-1}$ whenever the denominator is non-zero. We have

$$\begin{aligned}
\rho(x, \mathsf{A}(x)) &= \beta(x) + \int_{\mathcal{K}} \rho(d\kappa)\mathcal{P}_{\kappa x} 1\{x \notin \mathcal{M}\} \\
&= \beta(x) + \sum_y \rho(y, \mathsf{A}(y)) \int_{\mathsf{A}(y)} \mathcal{P}_{yax} 1\{x \notin \mathcal{M}\} w_y(da) \\
&= \beta(x) + \sum_y \rho(y, \mathsf{A}(y))_{\mathcal{M}} P_{yx}(w). \qquad (8.12)
\end{aligned}$$

By Lemma 7.1 (i), we conclude that $f_{tc}(\beta, w(\rho); x) \leq \rho(x, \mathsf{A}(x))$ for all $x \in \mathbf{X}$. By the definition of $w(\rho)$, it follows that $f_{tc}(\beta, w(\rho)) \leq \rho$.

(ii) That $\mathbf{L}(\beta) = \mathbf{L}_{U_S}(\beta)$ follows from Theorem 8.1 (ii); hence $\mathbf{L}_{U_S}(\beta)$ is convex. The compactness of $\mathbf{L}_{U_M}(\beta)$ follows since by Section 6.3 and Lemma 8.1 it is the image of the compact set U_M under the continuous function $f_{tc}(\beta, \cdot)$. We show that $\mathbf{L}_{U_S}(\beta)$ is equal to the closed convex hull of $\mathbf{L}_{U_D}(\beta)$ (and thus to $\mathbf{L}_{\mathcal{U}}(\beta)$). Since it is compact, by the Krein-Milman theorem (see Krein and Milman, 1940, or Dunford and Schwartz, 1988, p. 440), it is the closed convex hull of its extreme points. Choose some extreme point ρ of $\mathbf{L}_{U_S}(\beta)$. Define $w(\rho)$ to be again a stationary policy such that $w_y(\mathcal{A}) = \rho(y, \mathcal{A})[\rho(y, \mathsf{A}(y))]^{-1}$ whenever the denominator is non-zero. When it is zero, we let w_y be concentrated on $a(y)$, where $a(y)$ is some arbitrary action in $\mathsf{A}(y)$. It follows from the proof of Theorem 8.1 that $f_{tc}(\beta, w) = \rho$. Assume that $w \notin U_D$. Then there exists some $y \in \mathbf{X}$ such that $\rho(y, \mathsf{A}(y)) > 0$. But then by Lemma 8.2, $f_{tc}(\beta, w)$ is not an extreme point of $\mathbf{L}_{U_S}(\beta)$, as it can be expressed as a convex combination of the distinct points $f_{tc}(\beta, w^a)$ (where w^a are given in Lemma 8.2).

For all $f \in \mathbf{L}(\beta)$, $\langle f, \nu \rangle < \infty$ and f is a finite measure. Indeed, for an MDP with a uniform Lyapunov function,

$$\langle f, \nu \rangle = E^u_\beta \sum_{n=1}^{\infty} \big(1 + \nu(X_n, A_n)\big) 1\{T > n\} < \infty;$$

the last inequality follows from Theorem 7.4 (see property N2 there). The first equality follows easily from the monotone convergence theorem (for a precise proof, see Theorem 8.3 part (i)). Since by the part (i) of our theorem, $\mathbf{L}(\beta) \subset \mathbf{Q}_{tc}(\beta)$, we conclude that $\mathbf{L}(\beta) \subset \mathbf{Q}^{\nu}_{tc}(\beta) \subset \mathbf{Q}^{\mathrm{b}}_{tc}(\beta)$. It remains to show that $\mathbf{Q}^{\mathrm{b}}_{tc}(\beta) \subset \mathbf{L}_{U_S}(\beta)$.

Let $\rho \in \mathbf{Q}^{\mathrm{b}}_{tc}(\beta)$, and let w be the stationary policy as in the proof of (i). Since $[{}_{\mathcal{M}}P_{yx}(w)]^n$ converges to 0 pointwise, it follows from the bounded convergence theorem (e.g., Royden, 1988, Proposition 11.18) that

$$\lim_{n\to\infty} \rho_{\mathcal{M}} P^n(w) = 0.$$

By combining (8.12) with Lemma 7.1 (i), we conclude that $\rho = f_{tc}(\beta, w) \in \mathbf{Q}^{\mathrm{b}}_{tc}(\beta)$. The rest follows from part (i).
(iii) Since contracting MDPs are a special case of MDPs with uniform Lyapunov functions (Theorem 7.5), all the statements in (ii) hold. According to this theorem, we may choose ν to be proportional to μ. This implies that $\mathbf{Q}^{\mu}_{tc}(\beta) \subset \mathbf{L}_{U_S}(\beta)$, which concludes the proof. □

Note that the closure of the convex hull in Theorem 8.2 (ii) and (iii) is taken in the weak convergence topology.

8.5 Relation between cost and occupation measure

We have the following properties of the total costs:

Theorem 8.3 (*Linear representation and boundedness of the cost*)
For any $u \in U$ and $u \in \overline{M}(U_M)$,

$$C_{tc}(\beta, u) = \langle f_{tc}(\beta, u), c\rangle := \int_{\mathcal{K}} c(\kappa) f_{tc}(\beta, u; d\kappa) \tag{8.13}$$

holds if
(i) the immediate costs are non-negative, or
(ii) the MDP has a uniform Lyapunov function, and the immediate costs are ν-bounded from below.
If the MDP is contracting, then the finite and infinite horizon total costs are uniformly μ-bounded over all policies:

$$\|C^n_{tc}(\cdot, u)\|_\mu \le \frac{\overline{b}}{1-\xi} \qquad\qquad \|C_{tc}(\cdot, u)\|_\mu \le \frac{\overline{b}}{1-\xi}. \tag{8.14}$$

($C_{tc}(\cdot, u)$ is the vector of total cost corresponding to all initial states, and $\overline{b}$ is the μ-bound on the immediate costs.)

Proof. (i)

$$\begin{aligned} C_{tc}(\beta, u) &= \sum_{t=1}^{\infty} E^u_\beta c(X_t, A_t) = \sum_{t=1}^{\infty} \int_{\mathcal{K}} p^u_\beta(t; d\kappa) c(\kappa) \\ &= \int_{\mathcal{K}} \sum_{t=1}^{\infty} p^u_\beta(t; d\kappa) c(\kappa) = \langle f_{tc}(\beta, u), c\rangle, \end{aligned}$$

where the change between integration and summation follows since the integrand is non-negative (see Royden, 1988, Corollary 11.14).

(ii) For the finite horizon cost, we have by Fubini's Theorem (Royden, 1988, Theorem 12.19),

$$\begin{aligned} C^t_{tc}(\beta,u) &= \langle \sum_{s=1}^{t} p^u_\beta(s), c\rangle \\ &= \langle \sum_{s=1}^{t} p^u_\beta(s), \nu + c\rangle - \langle \sum_{s=1}^{t} p^u_\beta(s), \nu\rangle. \end{aligned}$$

By part (i) of the theorem, this difference converges to

$$\langle f_{tc}(\beta,u), \nu + c\rangle - \langle f_{tc}(\beta,u), \nu\rangle = \langle f_{tc}(\beta,u), c\rangle;$$

the above difference is indeed well defined since $\langle f_{tc}(\beta,u), \nu\rangle \le \langle \beta, \mu\rangle < \infty$ by property M2 of the uniform Lyapunov functions.

(8.14) follows from

$$\|C_{tc}(u)\|_\mu = \|\langle f_{tc}(\cdot,u), c\rangle\|_\mu \le \overline{b}\,\|f_{tc}(\cdot,u)\|_\mu \le \frac{\overline{b}}{1-\xi}, \tag{8.15}$$

and

$$\left\|C^t_{tc}(\cdot,u)\right\|_\mu = \left\|\langle f^t_{tc}(\cdot,u), c\rangle\right\|_\mu \le \overline{b}\,\|f_{tc}(\cdot,u)\|_\mu \le \frac{\overline{b}}{1-\xi}. \tag{8.16}$$

□

Lemma 8.4 (*The transient case: lower semi-continuity of the costs*) *Consider the transient framework, (Definition 7.1 with non-negative costs). Then $C(\beta,\cdot)$ (and $D^k(\beta,\cdot), k = 1,\dots,K$) are lower semi-continuous on U_M and on $\overline{M}(U_M)$.*

Proof. Follows directly from Lemma 8.1 (i). □

Lemma 8.5 (*Uniform convergence and continuity of the costs*) *Assume that the MDP has a uniform Lyapunov function and the immediate costs are ν-bounded. Then*
(i) $C^t_{tc}(\beta,u)$ converges to $C_{tc}(\beta,u)$ uniformly over U and $\overline{M}(U_M)$ as $t \to \infty$. In particular, if the MDP is contracting then for any u,

$$|C_{tc}(\beta,u) - C^t_{tc}(\beta,u)| \le \frac{\overline{b}\langle \beta,\mu\rangle \xi^t}{1-\xi}.$$

(ii) $C_{tc}(\beta,u)$ is continuous on U_M and on $\overline{M}(U_M)$.

Assume that the the immediate costs are only ν-bounded below. Then
(iii) $C(\beta,\cdot)$ (and $D^k(\beta,\cdot), k = 1,\dots,K$) are lower semi-continuous on U_M and on $\overline{M}(U_M)$.

Proof. (i) For any policy u,

$$|C_{tc}(\beta,u) - C^t_{tc}(\beta,u)|$$

$$\leq \sum_{s=t+1}^{\infty} E_\beta^u \nu(X_t, A_t) 1\{T > s\}$$
$$\leq E_\beta^u \hat{M}(X_{t+1}) 1\{T > t+1\}. \tag{8.17}$$

(ν and $\hat{M}$ are given in Definition 7.5 and in Section 7.4, respectively.) The right-hand side of (8.17) converges to zero uniformly in $u \in U_M$ due to property N4 (i) and Theorem 7.4. The generalization to any policy follows from Theorem 6.1.

For contracting MDPs, we have for any policy u,

$$|C_{tc}(\beta, u) - C_{tc}^t(\beta, u)|$$
$$\leq \sum_{s=t+1}^{\infty} E_\beta^u |c(X_t, A_t)| 1\{T > s\} = \sum_{s=t+1}^{\infty} \int_{\mathcal{K}} p_\beta^u(s; d\kappa) |c(\kappa)|$$
$$\leq \overline{b} \sum_{s=t+1}^{\infty} \sum_{y \in \mathbf{X}'} p_\beta^u(s; y) \mu(y) \leq \frac{\overline{b} \langle \beta, \mu \rangle \xi^t}{1 - \xi}.$$

The last inequality follows from Lemma 7.13.
(ii) The continuity follows by combining Theorem 8.3 with Lemma 8.1 (ii). (For the Markov policies, this is established in Theorem 7.4, by applying N5′.)
(iii) This is obtained by using the linear representation of the cost (Theorem 8.3) and by decomposing the immediate cost into the negative and positive parts. Then Lemmas 8.4 and 8.5(ii) are used for the positive and negative parts, respectively. □

8.6 Dominating classes of policies

Theorem 8.4 (*Dominating policies*)
(*i*) *Consider the transient framework,* (*Definition* 7.1)*, together with non-negative costs. Then both* U_S *and* $\mathcal{U}$ *are dominating classes of policies.*
(*ii*) *Consider an MDP with a uniform Lyapunov function and immediate costs that are* ν*-bounded from below. Then any complete class of policies* (*Definition* 8.1) *is a dominating class of policies.*
(*iii*) *Under the assumptions of either* (*i*) *or* (*ii*)*, if* **COP** *is feasible, then there exist optimal policies in* U_S *and in* $\mathcal{U}$*.*

Proof. (i) follows from the linear representation of the cost (Theorem 8.3) as well as the weak completeness of the set of stationary policies (Theorem 8.1). We delay the proof for $\mathcal{U}$ to the next chapter (Corollary 9.1).
(ii) follows from similar arguments.
(iii) Recall that the sets U_S of stationary policies and $\mathcal{U}$ are compact. Under the assumptions of (i) or (ii), the costs are lower semi-continuous on U_S and on $\mathcal{U}$ (Lemmas 8.5, 8.4). This implies that the feasible set of stationary

policies $\Pi_V := \{u : u \in U_S, D_{tc}(\beta, u) \leq V\}$ is compact, since it is obtained as the intersection of the compact set U_S and the inverse map of the closed sets $(-\infty, V_k]$. Finally, by the lower semi-continuity of $C_{tc}(\beta, u)$ on Π_V, we conclude that $C_{tc}(\beta, u)$ achieves its minimum on $\Pi_V(\beta, u)$, i.e., there exists an optimal stationary policy for COP. Similarly, it follows that there exists an optimal policy within $\mathcal{U}$. □

8.7 Equivalent Linear Program

We show below that **COP** is equivalent to an LP with an infinite set of decision variables and a countable set of constraints. Such equivalence was obtained for the total cost criterion for finite states and actions by Kallenberg (1983). The LP formulation constitutes an important method for computing stationary optimal policies.

Consider the following LP that will correspond to the transient case:
$\mathbf{LP_1}(\beta)$: Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle \rho, c \rangle$ subject to:

$$\mathcal{D}^k(\rho) := \langle \rho, d^k \rangle \leq V_k, k = 1, \ldots, K, \qquad \rho \in \mathbf{Q}_{tc}(\beta) \tag{8.18}$$

where $\mathbf{Q}_{tc}(\beta)$ was defined in (8.2).

$\mathbf{LP_1^b}(\beta)$ for the case of uniform Lyapunov function is defined similarly, with the set $\mathbf{Q}_{tc}^{\mathrm{b}}(\beta)$ replacing the set $\mathbf{Q}_{tc}$.

We show that there is a one to one correspondence between feasible (and optimal) solutions to the LP, and the feasible (and optimal) solutions to **COP**.

Theorem 8.5 (*Equivalence between COP and LP, the transient case*)
Consider a transient MDP and non-negative immediate costs. Then
(i) $\mathcal{C}^* = C_{tc}(\beta)$.
(ii) For any $u \in U$, $\rho(u) := f_{tc}(\beta, u) \in \mathbf{Q}_{tc}(\beta)$, $C_{tc}(\beta, u) = \mathcal{C}(\rho(u))$ *and* $D_{tc}(\beta, u) = \mathcal{D}(\rho(u))$*; conversely, for any* $\rho \in \mathbf{Q}_{tc}(\beta)$*, the stationary policy* $w(\rho)$ (*defined above* (8.12)) *satisfies* $C_{tc}(\beta, w(\rho)) \leq \mathcal{C}(\rho)$ *and* $D_{tc}(\beta, w(\rho)) \leq \mathcal{D}(\rho)$.
(iii) $\mathbf{LP_1}(\beta)$ *is feasible if and only if* **COP** *is. Assume that* **COP** *is feasible. Then there exists an optimal solution* ρ^* *for* $\mathbf{LP_1}(\beta)$*, and the stationary policy* $w(\rho^*)$ *is optimal for* **COP**.

Proof. We start from (ii). The first claim follows from (8.5). The claims on the costs follow from Theorem 8.3 and Theorem 8.1.
(i) and (iii) now follow from (ii) and Theorem 8.4. □

For the case of a uniform Lyapunov function, we similarly get

Theorem 8.6 (*Equivalence between COP and LP for MDPs with a uniform Lyapunov function*)
Assume that the MDP has a uniform Lyapunov function and that the immediate costs are ν*-bounded from below. Then*
(i) $\mathcal{C}^* = C_{tc}(\beta)$.

(ii) For any $u \in U$, $\rho(u) := f_{tc}(\beta, u) \in \mathbf{Q}^{\mathrm{b}}_{tc}(\beta)$, $C_{tc}(\beta, u) = \mathcal{C}(\rho(u))$ *and* $D_{tc}(\beta, u) = \mathcal{D}(\rho(u))$*; conversely, for any* $\rho \in \mathbf{Q}^{\mathrm{b}}_{tc}(\beta)$ *the stationary policy* $w(\rho)$ *(defined above* (8.12) *satisfies* $C_{tc}(\beta, w(\rho)) = \mathcal{C}(\rho)$ *and* $D_{tc}(\beta, w(\rho)) = \mathcal{D}(\rho)$.
(iii) $\mathbf{LP}^{\mathrm{b}}_{\mathbf{1}}(\beta)$ *is feasible if and only if* **COP** *is. Assume that* **COP** *is feasible. Then there exists an optimal solution* ρ^* *for* $\mathbf{LP}^{\mathrm{b}}_{\mathbf{1}}(\beta)$ *and the stationary policy* $w(\rho^*)$ *is optimal for* **COP**.

8.8 The dual program

Next, we present the formal dual program DP for the LP above. The decision variables are $\phi : \mathbf{X} \to \mathbb{R}$ and the K-dimensional non-negative vectors $\lambda \in \mathbb{R}^K_+$.

$$\begin{aligned} \mathbf{DP_1}(\beta): \qquad & \text{Find } \Theta^* := \sup_{\phi,\lambda} \langle \beta, \phi \rangle - \langle \lambda, V \rangle \text{ subject to} \\ \phi(x) \quad &\leq \quad c(x,a) + \langle \lambda, d(x,a) \rangle + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y), \ x \in \mathbf{X}, a \in \mathsf{A}(x). \end{aligned}$$

We shall show in the next chapter that when choosing the decision variables ϕ to be in the appropriate linear space, then there is no duality gap, and

$$\Theta^* = \mathcal{C}^* = C_{tc}(\beta), \tag{8.19}$$

for both transient MDPs as well as for MDPs with uniform Lyapunov functions. In particular, for the case of a uniform Lyapunov function with ν-bounded costs, we shall restrict ourselves to $\phi \in \mathsf{F}^\mu$; and for the transient framework with non-negative immediate costs, a possible choice is for ϕ to vanish outside some compact set.

CHAPTER 9

The total cost: Dynamic and Linear Programming

In the previous chapter we obtained an LP that was seen to be equivalent to **COP**; it yields the same value, and can be used to compute an optimal stationary policy for **COP**. That LP was the starting point for the analysis of constrained MDP by Derman and Veinott (1972). $\mathbf{DP_1}(\beta)$, the dual of that LP, was obtained directly by Kallenberg (1983) from dynamic programming arguments (for the finite state and action spaces without constraints).

We follow in this chapter a similar approach to obtain $\mathbf{DP_1}(\beta)$, using dynamic programming arguments and Lagrangian techniques. Then, by using standard saddle-point theorems, we show that there is no duality gap between $\mathbf{DP_1}(\beta)$ and $\mathbf{LP_1}(\beta)$. The derivation of $\mathbf{DP_1}(\beta)$ is independent of the geometric description of achievable occupation measures developed in the previous chapter.

In obtaining the Linear Program for the general transient case, we establish a calculation approach for the value function of **COP** based on finite state approximation. Unlike previous approaches for state approximations for **COP** (most of which were derived for the contracting framework, see Chapter 16 and Altman, 1993, 1994), we do not need here any Slater-type condition (see (9.32) below).

Some analysis of constrained MDPs was presented in the past by considering directly the Lagrangian formulation for a single constraint, see Beutler and Ross (1985, 1986), Sennott (1991, 1993). The use of Lagrangian techniques for several constraints is quite recent (see e.g., Arapostathis *et al.*, 1993, Piunovskiy, 1993, 1994, 1995, 1996, 1997a, 1997b, and Altman and Spieksma, 1995), and has not been much exploited. Not only does the Lagrangian approach enable one to derive different linear programming formulations (as we illustrate in this chapter), but also to obtain many results on asymptotic behavior of constrained MDPs (this is done in Chapters 13-16).

We finally present a different LP approach for computing the optimal values and optimal mixed strategies. Although in practice this alternative approach has a numerical complexity which is too high (in the case of finite states and actions), it has special features that will make it very useful in the study of sensitivity analysis, see e.g., Tidball and Altman (1996).

9.1 Non-constrained control: Dynamic and Linear Programming

We describe in this section the dynamic programming formulation for solving unconstrained MDPs. This approach has been developed starting from Shapely (1953), in the context of Markov games, which generalize MDPs to a setting with several controllers. For more detailed presentation, algorithmic procedures and references, see e.g., Puterman (1994).

Lemma 9.1 (*Uniform optimality and optimality for a given* β)
Fix an initial distribution β*. Assume either* (*i*) *that the cost is non-negative, or (ii) that the MDP has a uniform Lyapunov function, the immediate costs are* ν*-bounded from below, and* $\langle \beta, \nu \rangle < \infty$.
If u *is uniformly optimal* (*see Definition* 2.1)*, then it minimizes* $C_{tc}(\beta, u)$.

Proof. Let u be uniformly optimal. For any policy v,

$$\begin{aligned} C_{tc}(\beta, u) &= \sum_{x \in \mathbf{X}} \beta(x) C_{tc}(x, u) \\ &\leq \sum_{x \in \mathbf{X}} \beta(x) C_{tc}(x, v) = C_{tc}(\beta, v), \end{aligned}$$

which establishes the proof. The equalities follow by changing the order of expectations, since, by Theorem 8.3, we have:

$$C_{tc}(\beta, v) = \sum_{x \in \mathbf{X}} \beta(x) \langle f_{tc}(x, v), c \rangle = \langle \sum_{x \in \mathbf{X}} \beta(x) f_{tc}(x, v), c \rangle.$$

Under (i), this change is justified since the integrands are non-negative (see Royden, 1988, Corollary 11.14). For case (ii), it follows since

$$\begin{aligned} C_{tc}(\beta, v) &= \sum_{x \in \mathbf{X}} \beta(x) \left(\langle f_{tc}(x, v), \nu + c \rangle - \langle f_{tc}(x, v), \nu \rangle \right) \\ &= \langle f_{tc}(\beta, v), \nu + c \rangle - \langle f_{tc}(\beta, v), \nu \rangle \\ &= \langle f_{tc}(\beta, v), c \rangle. \end{aligned}$$

The integrands are again non-negative and the integrals are well defined (in particular, $\langle f_{tc}(\beta, v), \nu \rangle \leq \langle \beta, \mu \rangle$, see Lemma 7.5 (ii)). □

Introduce the dynamic programming inequality:

$$\phi(x) \geq \min_{a \in \mathsf{A}(x)} \left[c(x, a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \phi(y) \right], \qquad x \in \mathbf{X}. \tag{9.1}$$

Theorem 9.1 (*Dynamic programming: the transient case*)
Consider the transient framework (*Definition* 7.1 *and non-negative immediate costs*)*. Then*
(*i*) *The optimal value* $C_{tc}(x), x \in \mathbf{X}$*, is the smallest* (*componentwise*) *non-negative solution of* (9.1).

(ii) For any state x, let $\mathcal{A}(x)$ be the set of actions that achieve the minimum of $[c(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y)]$. A stationary policy g is uniformly optimal (see Definition 2.1) if it chooses actions within $\mathcal{A}(x), x \in \mathbf{X}$ w.p.1 (i.e., for which $g(\mathcal{A}(x)) = 1$ for all $x \in \mathbf{X}$). If, moreover, for all $x \in \mathbf{X}$

$$\lim_{t\to\infty} E_x^u \phi(X_{t+1}) 1\{T > t+1\} = 0, \qquad \forall u \in U_S \tag{9.2}$$

holds with $\phi = C_{tc}$, then also the converse holds.
Fix some initial distribution β. Let v be a stationary policy that does not choose among $\mathcal{A}(x)$ at some x for which $f_{tc}(\beta, v; x) > 0$. Assume that (9.2) holds for all x in the support of β. Then $C_{tc}(\beta, v) > C_{tc}(\beta)$.
(iii) The optimal value $C_{tc}(x), x \in \mathbf{X}$, achieves (9.1) with strict equality.

Proof. (i) We consider any non-negative solution ϕ of (9.1) and let w be a stationary policy that chooses at state x an action that achieves the minimum of $[c(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}\phi(y)]$. We iterate (9.1) and obtain:

$$\begin{aligned}
\phi(x) &\geq c(x,w) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xwy}\phi(y) = c(x,w) + E_x^w \phi(X_2) 1\{T>2\} \\
&\geq c(x,w) + E_x^w \left[c(X_2, A_2) 1\{T>2\} + E_{X_2}^w \phi(X_3) 1\{T>3\}\right] \\
&= c(x,w) + E_x^w c(X_2, A_2) 1\{T>2\} + E_x^w \phi(X_3) 1\{T>3\} \\
&\geq \ldots \geq \sum_{t=1}^{n} E_x^w c(X_t, A_t) 1\{T>t\} + E_x^w \phi(X_{n+1}) 1\{T > n+1\} \\
&\geq C_{tc}^n(x,w)
\end{aligned} \tag{9.3}$$

where the last inequality follows from the fact that ϕ is non-negative. Since (9.3) holds for all integers n, we conclude that $\phi(x) \geq C_{tc}(x)$.

On the other hand,

$$\begin{aligned}
C_{tc}(x) &= \inf_{u\in U_M} C_{tc}(x,u) = \inf_{u\in U_M} \left[c(x,u_1) + E_x^u \sum_{t=2}^{\infty} c(X_t, A_t) 1\{T>t\} \right] \\
&= \inf_{u\in U_M} \left[c(x,u_1) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xu_1y} C_{tc}(y) \right] \\
&= \min_{a\in \mathsf{A}(x)} \left[c(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y) \right].
\end{aligned} \tag{9.4}$$

This establishes (i) and (iii).

If g is a policy as in (ii), then it follows by applying the first part of the proof of (i), with $\phi = C_{tc}$ and $w = g$ that $C_{tc}(x) \geq C_{tc}(x,g)$, and hence g is uniformly optimal. To obtain the converse, assume that $u \in U_S$ does not satisfy the condition in the theorem (which g satisfies), i.e., for some

$x \in \mathbf{X}$ and $\delta > 0$,

$$c(x,u) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xuy} C_{tc}(y) = \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y) \right] + \delta.$$

Since by (iii), (9.1) is obtained with equality for $\phi = C_{tc}$, we have

$$\begin{aligned} C_{tc}(x) + \delta &= c(x,u) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xuy} C_{tc}(y) \\ &= c(x,u) + E_x^u [C_{tc}(X_2) 1\{T > 2\}] \\ &\leq c(x,u) + E_x^u [c(X_2, A_2) 1\{T > 2\} + E_{X_2}^u C_{tc}(X_3) 1\{T > 3\}] \\ &\leq \ldots \leq \sum_{t=1}^{n} E_x^u c(X_t, A_t) 1\{T > t\} + E_x^u C_{tc}(X_{n+1}) 1\{T > n+1\}. \end{aligned}$$

Taking the limit as $n \to \infty$ (and using part (i)), we obtain $C_{tc}(x,u) \geq C_{tc}(x) + \delta$. Hence u is not uniformly optimal.

The statements concerning β is obtained by similar arguments. This establishes (ii). □

In the following example we show that, indeed, (9.1) may have several solutions larger than the value function.

Example 9.1 *(On the necessity of the restrictions on ϕ)*
Consider a discrete time queueing model. At each time period t, there may be an arrival of a customer with probability λ, or a departure from the queue with probability μ'. We assume that $\mu' > \lambda$. The arrivals and departures in different time periods are independent. The state space is the set of integers, and a state x has the meaning that there are x customers in the queue. There is no control here (thus, we may assume that $\mathsf{A}(x) = \{a\}$ contains a dummy control action a at all states). We wish to compute $M(x)$:= the expected time it takes the queue to empty (i.e., to reach state 0) starting from state x. In other words, we wish to compute the total expected cost with respect to the immediate cost $c(x) = c(x,a) = 1$ until the set $\mathcal{M} = \{0\}$ is reached. The solution satisfies

$$\begin{aligned} M(0) &= 1 + \lambda M(1), \\ M(x) &= 1 + \mu' M(x-1) + (1 - \mu' - \lambda) M(x) + \lambda M(x+1) \end{aligned} \quad (9.5)$$

for $x > 0$. In particular, it satisfies (9.1). (9.5) is a difference equation whose solution $\phi(x)$ is given by

$$\phi(x) = const \cdot \left[\left(\frac{\mu'}{\lambda} \right)^x - 1 \right] + \frac{x}{\mu' - \lambda}, \qquad x > 0, \quad (9.6)$$

and $\phi(0) = 1 + \lambda\phi(1)$. The expected time to reach 0 is

$$M(x) = \begin{cases} \dfrac{x}{\mu' - \lambda} & \text{for } x > 0, \\ \dfrac{\mu'}{\mu' - \lambda} & \text{for } x = 0. \end{cases}$$

But any other constant in (9.6) yields another solution of (9.1), so $M(x)$ is not the unique solution, nor the smallest one. It is, however, the smallest non-negative solution of (9.6) (and of (9.1)).

□

Theorem 9.2 (*Dynamic programming: uniform Lyapunov function*)
Consider an MDP with a uniform Lyapunov function.
(*i*) *Assume that the immediate costs are ν-bounded. Then the optimal value $C_{tc}(x), x \in \mathbf{X}$ is the unique solution of* (9.1) *in the class of μ-bounded functions. If the immediate costs are ν-bounded only from below, then $C_{tc}(x), x \in \mathbf{X}$ is the minimal solution of* (9.1) *in the class of functions which are μ-bounded from below. In both cases C_{tc} achieves* (9.1) *with strict equality.*
(*ii*) *Assume that the immediate costs are ν-bounded from below. For any state x, let $\mathcal{A}(x)$ be the set of actions that achieve the minimum of $[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y)]$. Any stationary policy g that chooses actions within $\mathcal{A}(x), x \in \mathbf{X}$ with probability one (i.e., for which $g_x(\mathcal{A}(x)) = 1$ for all $x \in \mathbf{X}$) is uniformly optimal. If the immediate costs are ν-bounded, then the converse also holds.*
Fix some initial distribution β. Let v be a stationary policy that does not choose among $\mathcal{A}(x)$ at some x for which $f_{tc}(\beta, v; x) > 0$. Assume that either
– the immediate costs are ν-bounded from below, and (9.2) *holds for all x in the support of β; or*
– the immediate costs are ν-bounded.
Then $C_{tc}(\beta, v) > C_{tc}(\beta)$.

Proof. Assume that the immediate costs are ν-bounded and let $\phi \in \mathsf{F}^{\mu}$ be a solution to (9.1). Since $\phi \in \mathsf{F}^{\mu}$, it follows from property M1(iii) of the uniform Lyapunov function that $\lim_{t\to\infty} E^u_\beta \phi(X_t) 1\{T > t\} = 0$. We take the limit in (9.3) as $n \to \infty$ and obtain $\phi(x) \geq C_{tc}(x)$. (i) then follows from (9.4). For the case that the immediate costs are only ν-bounded from below, the proof of (i) is obtained by combining the arguments here with those of Theorem 9.1, by considering separately the positive and negative parts of ϕ.
The proof of (ii) is the same as in Theorem 9.1. □

Remark 9.1 Consider Example 9.1. If we choose the μ norm to be $\mu(x) = \alpha^x$, for some $1 < \alpha < \mu'/\lambda$, then it can be seen that we obtain a contracting

Markov chain with respect to this norm and to the set $\mathcal{M} = \{0\}$. We see that M is indeed the unique μ-bounded solution of (9.6) (and of (9.1)).

Remark 9.2 By the same arguments as in the proof of Theorem 9.2, one can show that for any stationary policy $w \in U_S$, the cost $C_{tc}(x, w)$ is the unique solution in F^μ of

$$\phi(x) \geq c(x, w) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xwy}\phi(y), \qquad x \in \mathbf{X} \tag{9.7}$$

and the above inequality is then achieved as equality.

9.2 Super-harmonic functions and Linear Programming

Definition 9.1 *(Super-harmonic functions)*
Fix some $\mathbf{X}' \subset \mathbf{X}$. A function $\phi : \mathbf{X} \to \mathbb{R}$ is called super-harmonic (for the total cost criterion) if it satisfies for all $x \in \mathbf{X}$ and $a \in \mathsf{A}(x)$:

$$\phi(x) \leq c(x, a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay}\phi(y). \tag{9.8}$$

Super-harmonic functions have an important role in MDPs, illustrated by the following theorem (which is closely related to Theorems 3.1 and 3.4 in Hordijk, 1977).

Theorem 9.3 (*The value and super-harmonic functions*)
(*i*) *Consider the transient framework, (Definition 7.1 and non-negative immediate cost). Let ϕ be a super-harmonic function. If for some optimal stationary policy g,*

$$\overline{\lim_{t \to \infty}} E_x^g \phi(X_t) 1\{T > t\} \leq 0, \qquad \forall x \in \mathbf{X} \tag{9.9}$$

then the value $C_{tc} \geq \phi$ (componentwise!).
(*ii*) *Consider an MDP with a uniform Lyapunov function and ν-bounded immediate cost. Then the value C_{tc} is the largest super-harmonic function among the μ-bounded functions. If the immediate cost is only ν-bounded from below, then C_{tc} is the largest super-harmonic function among those satisfying (9.9).*

Proof. (i) From Theorem 9.1 it follows that C_{tc} is a super-harmonic function. Choose a super-harmonic function ϕ and let g be an optimal stationary policy satisfying (9.9). Then

$$\begin{aligned}
\phi(x) &\leq c(x, g) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xgy}\phi(y) = c(x, g) + E_x^g \phi(X_2) 1\{T > 2\} \\
&\leq c(x, g) + E_x^g \left[c(X_2, A_2) 1\{T > 2\} + E_{X_2}^g \phi(X_3) 1\{T > 3\} \right] \\
&= c(x, g) + E_x^g c(X_2, A_2) 1\{T > 2\} + E_x^g \phi(X_3) 1\{T > 3\}
\end{aligned}$$

$$\leq \quad \ldots \leq \sum_{t=1}^{n} E_x^g c(X_t, A_t) 1\{T > t\} + E_x^g \phi(X_{n+1}) 1\{T > n+1\}.$$

(i) is established by taking the limit as $n \to \infty$. (A proof of a more general statement can be found in Lemma 3.6 in Feinberg and Sonin, 1983.)
(ii) follows as the proof of (i). In particular, for the case of ν-bounded immediate costs, this follows from the dominated convergence theorem, since for any policy g, $\lim_{t\to\infty} E_x^g \mu(X_t) 1\{T > t\} = 0$ (see property M1′ in Section 7.4). □

Motivated by Theorem 9.3, we introduce the following infinite Linear Program with decision variables $\phi(y), y \in \mathbf{X}$.

$$\begin{aligned} \mathbf{DP}(\beta): \qquad & \text{Find } \phi^* := \sup_\phi \langle \beta, \phi \rangle \text{ subject to} \\ \phi(x) \quad & \leq \quad c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \phi(y), \qquad x \in \mathbf{X}, a \in \mathsf{A}(x). \end{aligned}$$

For the transient case (where the immediate cost is assumed to be non-negative), we may further add non-negativity constraints on ϕ without loss of optimality. Indeed, if ϕ is feasible for $\mathbf{DP}(\beta)$, then it is easily seen that ϕ' given by $\phi'(y) = \max(\phi(y), 0)$, $y \in \mathbf{X}$, is also feasible, and it clearly dominates ϕ.

We begin by considering the MDP with a uniform Lyapunov function. Theorem 9.3 implies the following:

Theorem 9.4 (*Dual Linear Program for MDPs with uniform Lyapunov functions*)
Assume that the MDP has a uniform Lyapunov function and the immediate costs are ν-bounded. Consider $\mathbf{DP}(\beta)$ where the decision variables are restricted to the set $\phi \in \mathsf{F}^\mu$. Then for any initial distribution β (with $\langle \beta, \mu \rangle < \infty$), $\mathbf{DP}(\beta)$ is feasible; its value equals $C_{tc}(\beta)$ and $\phi(x) = C_{tc}(x), x \in \mathbf{X}$ is an optimal solution.

A similar statement could be obtained for other cases (the transient case with non-negative immediate costs and the case of uniform Lyapunov function with costs ν-bounded from below) when restricting to functions for which the condition (9.9) from Theorem 9.3 (i) holds. However, the above condition may be difficult to verify. We therefore adopt an alternative approach, and identify a simple subclass of functions satisfying that condition. Of course, if we restrict the LP to a subclass of functions, we risk obtaining only a lower bound to the optimal value. Our choice of functions ϕ will turn out, however, to be rich enough to obtain the same value as the one obtained by the richer class of policies satisfying (9.9). We begin by considering the absorbing case with *unbounded costs*.

Theorem 9.5 (*The dual Linear Program, absorbing case*)
Assume that the MDP is absorbing to $\mathcal{M}$, and the immediate costs are non-negative (not necessarily bounded). Consider $\mathbf{DP}(\beta)$ *where the decision variables ϕ are all non-negative bounded functions. Then for any initial distribution β,* $\mathbf{DP}(\beta)$ *is feasible and its value equals $C_{tc}(\beta)$.*

Proof. Denote by $C^1(\beta)$ the value of $\mathbf{DP}(\beta)$ restricted to bounded ϕ. Since the MDP is absorbing to $\mathcal{M}$, it follows that for any bounded function ϕ, (9.9) holds for all policies: $\lim_{t\to\infty} E_x^g \phi(X_t) 1\{T > t\} = 0$. Theorem 9.3 (i) implies that

$$C^1(x) \le C_{tc}(x) \tag{9.10}$$

for all x.

Let $\mathbf{X}_n$ be an increasing sequence of *finite* sets of states converging to $\mathbf{X}$. Consider **COP** with an immediate cost $c_n(x,a) = c(x,a)1\{x \in \mathbf{X}_n\}$; denote by $C_{tc}^n(\beta,u)$ the corresponding total expected cost, and by $C_{tc}^n(\beta)$ the corresponding optimal value. For any policy u and initial distribution β, $C_{tc}^n(\beta,u)$ is increasing in n, and hence also $C_{tc}^n(\beta)$. Denote by $C_{tc}^*(\beta)$ the limit of $C_{tc}^n(\beta)$ as n tends to infinity. By Theorem 9.1, we have

$$C_{tc}^n(x) = \min_{a\in\mathsf{A}(x)} \left[c_n(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y) \right], \qquad x \in \mathbf{X}, \tag{9.11}$$

which implies that for all $x \in \mathbf{X}$ and $a \in \mathsf{A}(x)$ we have

$$C_{tc}^n(x) \le c_n(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y) \le c(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y). \tag{9.12}$$

Thus C_{tc}^n is a bounded super-harmonic function. Hence

$$C_{tc}^*(x) \le C^1(x) \le C_{tc}(x). \tag{9.13}$$

Let a_n be a minimizing action in (9.11) (it is, of course, a function of x), and let a^* be some limit point (as $n \to \infty$) obtained by diagonalization. By Fatou's Lemma, applied to (9.11), we get

$$C_{tc}^*(x) \ge c(x,a^*) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xa^*y} C_{tc}^*(y), \qquad x \in \mathbf{X}.$$

It then follows from Theorem 9.1 (i) that $C_{tc}(x) \le C_{tc}^*(x)$, and hence, by (9.13), we have $C_{tc}(x) = C^1(x)$. □

Next, we introduce the transient case:

Theorem 9.6 (*The dual Linear Program, transient case*)
Assume that the MDP is $\mathbf{X}'$-transient, and the immediate costs are non-negative. Consider $\mathbf{DP}(\beta)$ *where the decision variables ϕ are all non-negative functions that vanish outside of some finite set of states. In other words,*

for each ϕ in this class, there exists some finite $\mathbf{Y} \subset \mathbf{X}$ such that

$$\phi(x) = 0 \text{ for all } x \notin \mathbf{Y}. \tag{9.14}$$

Then for any initial distribution β, $\mathbf{DP}(\beta)$ is feasible and its value equals $C_{tc}(\beta)$.

Proof. Denote again by $C^1(\beta)$ the value of $\mathbf{DP}(\beta)$ restricted to ϕ that satisfies (9.14). Since the MDP is $\mathbf{X}'$-transient, it follows that for any function ϕ satisfying (9.14), (9.9) holds for all policies: $\lim_{t\to\infty} E_x^g \phi(X_t) 1\{T > t\} = 0$. Theorem 9.3 (i) implies that

$$C^1(x) \le C_{tc}(x) \tag{9.15}$$

for all x.

Let $\mathbf{X}_n$ be a sequence of finite sets of states, increasing to $\mathbf{X}$. Consider a sequence $\mathbf{COP}_n$ of truncated problems where $\mathbf{COP}_n$ differs from the original $\mathbf{COP}$ by the fact that the process is restricted to $\mathbf{X}_n$. This is done by altering transition probabilities and the costs:

$$\mathcal{P}^n_{xay} = \begin{cases} \mathcal{P}_{xay} & \text{if } x, y \in \mathbf{X}_n, \\ 1 & \text{if } x \notin \mathbf{X}_n, y = 0, \\ 0 & \text{otherwise,} \end{cases} \qquad c_n(x,a) = c(x,a) 1\{x \in \mathbf{X}_n\}. \tag{9.16}$$

Here, 0 is some arbitrary (possibly new) state which is not in $\mathbf{X}'$. Denote by $C^n_{tc}(\beta, u)$ the corresponding total expected cost, and by $C^n_{tc}(\beta)$ the corresponding optimal value. For any policy u and initial distribution β, $C^n_{tc}(\beta, u)$ is increasing in n, and so too is $C^n_{tc}(\beta)$. By Theorem 9.1, we have

$$C^n_{tc}(x) = \min_{a \in \mathsf{A}(x)} \left[c_n(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}^n_{xay} C^n_{tc}(y) \right], \qquad x \in \mathbf{X}, \tag{9.17}$$

which implies that for all $x \in \mathbf{X}$ and $a \in \mathsf{A}(x)$, we have

$$C^n_{tc}(x) \le c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}^n_{xay} C^n_{tc}(y) \le c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C^n_{tc}(y). \tag{9.18}$$

Thus C^n_{tc} is a super-harmonic function that vanishes outside of $\mathbf{X}_n$. Denote by $C^*_{tc}(\beta)$ the (increasing) limit of $C^n_{tc}(\beta)$ as n tends to infinity. Hence

$$C^*_{tc}(x) \le C^1(x) \le C_{tc}(x). \tag{9.19}$$

Let a_n be a minimizing action in (9.17) and let a^* be some limit point obtained by diagonalization. By Fatou's Lemma, applied to (9.17), we get

$$C^*_{tc}(x) \ge c(x,a^*) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xa^*y} C^*_{tc}(y), \qquad x \in \mathbf{X}.$$

It then follows from Theorem 9.1 (i) that $C_{tc}(x) \le C^*_{tc}(x)$, and hence, by (9.19), we have $C_{tc}(x) = C^1(x)$. □

By arguments similar to the previous theorems we get the following:

Theorem 9.7 (*The dual Linear Program, uniform Lyapunov functions*) *Assume that the MDP has a uniform Lyapunov function, and the immediate costs are ν-bounded below. Consider* $\mathbf{DP}(\beta)$ *where the decision variables ϕ are functions that are bounded from above. Then for any initial distribution β,* $\mathbf{DP}(\beta)$ *is feasible and its value equals $C_{tc}(\beta)$.*

Proof. Denote by $C^1(\beta)$ the value of $\mathbf{DP}(\beta)$ restricted to ϕ bounded from above. Since the MDP is absorbing to $\mathcal{M}$, it follows that for any bounded function ϕ, (9.9) holds for all policies: $\overline{\lim}_{t\to\infty} E_x^g \phi(X_t) 1\{T > t\} \leq 0$. Theorem 9.3 (ii) implies (9.10) for all x.

Let $\mathbf{X}_n$ be as in the proof of Theorem 9.5, and consider **COP** with an immediate cost $c_n(x,a) = c^-(x,a) + c^+(x,a)1\{x \in \mathbf{X}_n\}$, where $c^+(x,a) = \max(c(x,a),0)$ and $c^-(x,a) = \min(c(x,a),0)$. Denote by $C_{tc}^n(\beta,u)$ the corresponding total expected cost, and by $C_{tc}^n(\beta)$ the corresponding optimal value. As in the proof of Theorem 9.5, $C_{tc}^n(\beta)$ is increasing in n. Denote by $C_{tc}^*(\beta)$ the limit of $C_{tc}^n(\beta)$ as n tends to infinity.

By Theorem 9.2, (9.17) holds, which implies again (9.12). Thus C_{tc}^n is a super-harmonic function which is bounded from above, which implies (9.13). Let a_n and a^* be as in the proof of Theorem 9.5. We rewrite (9.17) as

$$C_{tc}^n(x) = \min_{a\in\mathsf{A}(x)} \left[c_n(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}[C_{tc}^n(y) + \mu(y)] - \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}\mu(y) \right], \tag{9.20}$$

$x \in \mathbf{X}$. The second summation is continuous in a from the definition of the uniform Lyapunov function. Hence, applying Fatou's Lemma to the first summation, we obtain

$$\begin{aligned} C_{tc}^*(x) &\geq c(x,a^*) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xa^*y}[C_{tc}^*(y) + \mu(y)] - \sum_{y\in\mathbf{X}'} \mathcal{P}_{xa^*y}\mu(y) \\ &= c(x,a^*) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xa^*y} C_{tc}^*(y), \qquad x \in \mathbf{X}. \end{aligned}$$

It then follows from Theorem 9.2 (i) that $C_{tc}(x) \leq C_{tc}^*(x)$, and hence, by (9.13), we have $C_{tc}(x) = C^1(x)$. □

Remark 9.3 (*On the solvability of the dual program*)
Unlike the case of uniform Lyapunov function with ν-bounded cost, or the absorbing case with bounded costs, the sup in $\mathbf{DP}(\beta)$ need not be achieved in the setting of Theorem 9.5 or 9.6 (i.e., the dual LP need not be solvable); we cannot expect $\phi(x) = C_{tc}(x), x \in \mathbf{X}$ to be an optimal solution since there is no reason to expect $C_{tc}(x)$ to be bounded.

Remark 9.4 (*State truncation*)
We have in fact presented in the proof of Theorem 9.6 a state truncation procedure that enables us to approximate the value of a non-constrained MDP with countable state space by an MDP with a finite state space. The policy that chooses at state x the action $a^* = a^*(x)$ is optimal for the MDP, due to Theorem 9.1 (ii). Since this policy is the limit of policies optimal for the truncated MDPs, we also conclude that the policies converge. A different approach to state truncation will be presented in Chapter 16; it will be used for the contracting framework.

The restriction in the dual LP to bounded functions ϕ in the absorbing case (Theorem 9.5), or to functions ϕ converging to zero for the general transient case (Theorem 9.6), is indeed necessary. This can be seen from our Example 9.1. Any function among (9.6) is feasible for the dual LP. The supremum over all these unbounded functions is infinity, and not $M(\cdot)$.

9.3 Set of achievable costs

Define for any class of policies $\overline{U}$ the set of achievable vector performance measures:

$$\mathbf{M}^{tc}_{\overline{U}}(\beta) = \cup_{u\in\overline{U}}\{(C_{tc}(\beta,u), D^k_{tc}(\beta,u), k=1,\ldots,K)\}, \tag{9.21}$$

and set $\mathbf{M}^{tc}(\beta) := \mathbf{M}^{tc}_{U}(\beta) \cup \mathbf{M}^{tc}_{\overline{M(U_M)}}(\beta)$.

Define also $\mathbf{V}_{tc}(\beta), \mathbf{V}^{\nu}_{tc}(\beta), \mathbf{V}^{\mu}_{tc}(\beta)$ and $\mathbf{V}^{\mathrm{b}}_{tc}(\beta)$ by

$$\bigcup_{\rho}\{(\langle\rho,c\rangle, \langle\rho,d^1\rangle, \langle\rho,d^2\rangle, \ldots, \langle\rho,d^K\rangle)\}, \tag{9.22}$$

where the union is taken over ρ in $\mathbf{Q}_{tc}(\beta), \mathbf{Q}^{\nu}_{tc}(\beta), \mathbf{Q}^{\mu}_{tc}(\beta)$ and $\mathbf{Q}^{\mathrm{b}}_{tc}(\beta)$, respectively ($\mathbf{Q}_{tc}(\beta)$ is defined in (8.2)).

Recall the definition of $\min B$ from Section 8.1. The next characterization of achievable costs follows from Theorem 8.2, as well as the linear representation of the cost (Theorem 8.3).

Theorem 9.8 (*Characterization of the sets of achievable costs*)
(*i*) *For transient MDPs with non-negative immediate costs,* $\mathbf{M}^{tc}(\beta)$ *is convex, and*

$$\min \mathbf{V}_{tc}(\beta) = \mathbf{M}^{tc}_{U_S}(\beta) \prec \mathbf{M}^{tc}_{U_M}(\beta) = \mathbf{M}^{tc}(\beta) \subset \mathbf{V}_{tc}(\beta)$$

($\prec$ *is defined in Section* 8.1).
(*ii*) *For MDPs with a uniform Lyapunov function and immediate costs* ν*-bounded from below,* $\mathbf{M}^{tc}_{U_S}(\beta)$ *is convex and compact, and satisfies*

$$\begin{aligned}\mathbf{M}^{tc}_{\mathcal{U}}(\beta) &= \mathbf{M}^{tc}(\beta) = \mathbf{M}^{tc}_{U_S}(\beta) = \overline{co}\mathbf{M}^{tc}_{U_D}(\beta)\\ &= \mathbf{V}^{\nu}_{tc}(\beta) = \mathbf{V}^{\mathrm{b}}_{tc}(\beta) = \min \mathbf{V}_{tc}(\beta).\end{aligned}$$

For contracting MDPs, the above holds, in particular, with $\mathbf{V}^{\mu}_{tc}(\beta)$ *replacing* $\mathbf{V}^{\nu}_{tc}(\beta)$.

9.4 Constrained control: Lagrangian approach

We now go back to our constrained control problem. We follow the same steps as in Section 3.3, to show that

- **COP** is equivalent to solving a non-constrained sup-inf problem; the sup and inf can be interchanged under suitable conditions.
- The inf in the inf-sup problem is in fact achieved by some policy which is optimal for **COP**.
- Under the Slater conditions (9.32), the sup in the sup-inf problem is achieved by some Lagrange multiplier.

The main result is presented in the following theorem. Its derivation is independent of the theory we developed in the previous chapter for the achievable sets of occupation measure.

Theorem 9.9 (*The Lagrangian*)
Consider either
– the transient framework (*Definition 7.1 and non-negative immediate cost*), *or*
– MDPs with a uniform Lyapunov function and immediate costs ν*-bounded from below*
(*i*) *The value function satisfies*

$$C_{tc}(\beta) = \inf_{u\in U}\sup_{\lambda\geq 0} J^{\lambda}_{tc}(\beta,u) = \inf_{u\in U_M}\sup_{\lambda\geq 0} J^{\lambda}_{tc}(\beta,u) = \inf_{u\in \overline{M}(U_M)}\sup_{\lambda\geq 0} J^{\lambda}_{tc}(\beta,u), \tag{9.23}$$

where

$$\begin{aligned} J^{\lambda}_{tc}(\beta,u) &:= C_{tc}(\beta,u) + \langle \lambda, D_{tc}(\beta,u) - V\rangle \\ &= \sum_{t=1}^{\infty} E^{u}_{\beta} j^{\lambda}(X_t,A_t)1\{T>t\} - \langle \lambda, V\rangle \\ j^{\lambda}(x,a) &:= c(x,a) + \langle \lambda, d(x,a)\rangle. \end{aligned} \tag{9.24}$$

(*ii*) *A policy* u^* *is optimal for* **COP** *if and only if* $C_{tc}(\beta) = \sup_{\lambda\geq 0} J^{\lambda}_{tc}(\beta,u^*)$.
(*iii*) *The value satisfies*

$$C_{tc}(\beta) = \sup_{\lambda\geq 0}\min_{u\in\overline{M}(U_M)} J^{\lambda}_{tc}(\beta,u) = \sup_{\lambda\geq 0}\min_{u\in U_D} J^{\lambda}_{tc}(\beta,u). \tag{9.25}$$

Moreover, there exists some $u^*\in\mathcal{U}$ *such that*

$$C_{tc}(\beta) = \inf_{u\in\mathcal{U}}\sup_{\lambda\geq 0} J^{\lambda}_{tc}(\beta,u) = \sup_{\lambda\geq 0} J^{\lambda}_{tc}(\beta,u^*), \tag{9.26}$$

and u^* *is optimal for* **COP**.

In order to prove the theorem, we need the following (e.g., Aubin, 1993, p. 126):

Lemma 9.2 (*Minmax Theorem*)
Let G_1 and G_2 be convex subsets of linear topological spaces, and let G_1 be compact. Consider a function $\Psi : G_1 \times G_2 \to \mathbb{R}$ such that
– for each $g_2 \in G_2$, $g_1 \to \Psi(g_1, g_2)$ is convex and lower semi-continuous, and
– for each $g_1 \in G_1$, $g_2 \to \Psi(g_1, g_2)$ is concave.
Then there exists some $g_1^* \in G_1$ such that

$$\inf_{G_1} \sup_{G_2} \Psi(g_1, g_2) = \sup_{G_2} \Psi(g_1^*, g_2) = \sup_{G_2} \inf_{G_1} \Psi(g_1, g_2).$$

We are now ready to prove Theorem 9.9.
Proof of Theorem 9.9: (i) The first equality in (9.23) is standard: if $u \in U$ is feasible (i.e., it satisfies the constraints $D_{tc}(\beta, u) \leq V$), then

$$\sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) = C_{tc}(\beta, u). \tag{9.27}$$

If $u \in U$ is not feasible, then it is easily seen that

$$\sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) = \infty. \tag{9.28}$$

We conclude that

$$\begin{aligned} \inf_{u \in U} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) &= \inf_{u \in U, D_{tc}(\beta,u) \leq V} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) \\ &= \inf_{u \in U, D_{tc}(u) \leq V} C_{tc}(\beta, u) = C_{tc}(\beta). \end{aligned}$$

Similarly, let $C'_{tc}(\beta) := \inf C_{tc}(\beta, u)$ over the set $\{u \in U_M : D_{tc}(\beta, u) \leq V\}$. Then we have

$$C'_{tc}(\beta) = \inf_{u \in U_M} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u).$$

However, it follows from Section 6.4 that $C'_{tc}(\beta) = C_{tc}(\beta)$. This establishes the second equality. The third equality follows by the same arguments. This establishes (i). Moreover, (9.27) and (9.28) imply (ii).

(iii) We shall apply Lemma 9.2 where G_1 stands for the convex and compact set $\overline{M}(U_M)$ (for a discussion on the compactness, see Section 6.3), and G_2 stands for the convex set $\{\lambda \geq 0\}$. $J^{\lambda}(\beta, u) : G_1 \times G_2 \to \mathbb{R}$ is affine in both its arguments, and thus in particular, convex in its first argument and concave in the second. $J^{\lambda}(\beta, u)$ is lower semi-continuous in u (see Lemmas 8.4 and 8.5). Hence by Lemma 9.2, we have

$$\sup_{\lambda \geq 0} \min_{u \in \overline{M}(U_M)} J_{tc}^{\lambda}(\beta, u) = \inf_{u \in \overline{M}(U_M)} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) = C_{tc}(\beta) \tag{9.29}$$

where the last equality follows from (9.23); this establishes the first equality. For fixed λ, $J_{tc}^{\lambda}(\beta, u)$ is minimized by a policy in U_D (by Theorems 9.1 and

9.2), i.e.,

$$\min_{u\in\overline{U}} J^{\lambda}_{tc}(\beta,u) = \min_{u\in U_D} J^{\lambda}_{tc}(\beta,u) \tag{9.30}$$

for any class of policies $\overline{U}$ that contains U_D. The proof of (9.25) is established by combining this with (9.29).

By combining (9.29) with (9.30), we get

$$C_{tc}(\beta) = \sup_{\lambda\geq 0}\min_{u\in\mathcal{U}} J^{\lambda}_{tc}(\beta,u). \tag{9.31}$$

(9.26) is obtained by again applying Lemma 9.2 with G_1 as the convex and compact set $\mathcal{U}$, and G_2 as the convex set $\{\lambda \geq 0\}$. Here again, $J^{\lambda}_{tc}(\beta,u) : G_1 \times G_2 \to \mathbb{R}$ is affine in both its arguments, and thus in particular, convex in its first argument and concave in the second; $J^{\lambda}_{tc}(\beta,u)$ is lower semi-continuous in u (see Lemmas 8.4 and 8.5). This implies, in particular, the existence of of $u^* \in \mathcal{U}$ that minimizes the Lagrangian $J^{\lambda}_{tc}(\beta,u)$. By part (ii) of the theorem it is also optimal for **COP**. □

By the same type of arguments as in the proof of part (i) of Theorem 9.9, we obtain from (9.31) the following corollary for the transient framework (the case of uniform Lyapunov function was already established in Chapter 8).

Corollary 9.1 (*Dominance of* $\mathcal{U}$)
Consider the transient framework (Definition 7.1) with non-negative immediate costs. Then $\mathcal{U} = \overline{M}(U_D)$ *is a dominating class of policies.*

Proof. Choose an arbitrary policy v. Consider a new **COP** with the same state and action spaces, the same transition probabilities, and with $K+1$ constraints. The immediate costs $\{\tilde{c}, \tilde{d}^1, \ldots, \tilde{d}^{K+1}\}$ are given in terms of the immediate costs corresponding to the original **COP**:

$$\tilde{c} = 0, \qquad \tilde{d}^k = d^k, k = 1, \ldots, K, \qquad \tilde{d}^{K+1} = c.$$

We set

$$\tilde{V}_k := \tilde{D}_{tc}(\beta, v),\ k = 1, \ldots, K+1.$$

The new **COP** is feasible since the policy v is feasible (by definition of $\tilde{V}_k$). By the same arguments as in the first part of the proof of Theorem 9.9, it follows from (9.26) that there exists a feasible policy $\overline{u}$ among $\mathcal{U}$ for the new **COP**. This implies that $\overline{u}$ dominates v (for the original **COP**). □

Next, we consider the existence of maximizing Lagrangians.

Theorem 9.10 (*The Lagrangian: Slater condition*)
Consider either the transient framework with non-negative costs or MDPs with a uniform Lyapunov function and costs ν*-bounded from below. If there exists some policy* u *for which*

$$D_{tc}(\beta,u) < V, \tag{9.32}$$

then there exist non-negative Lagrange multipliers $\lambda^* = \{\lambda_1^*, \ldots, \lambda_K^*\}$ *such that*

$$C_{tc}(\beta) = \min_{u \in \mathcal{U}} J_{tc}^{\lambda^*}(\beta, u) = \min_{u \in U_D} J_{tc}^{\lambda^*}(\beta, u). \tag{9.33}$$

Moreover, any optimal policy u^* *satisfies the Kuhn-Tucker conditions:*

$$\lambda_k^*(D_{tc}^k(\beta, u^*) - V_k) = 0, \qquad k = 1, \ldots, K.$$

Proof. $J_{tc}^{\lambda}(\beta, u)$ is a convex function over the convex set $\mathcal{U}$, and $C_{tc}(\beta, u)$ and $D_{tc}^k(\beta, u)$ are lower semi-continuous in $\mathcal{U}$ (see Lemmas 8.4 and 8.5). By a standard minmax theorem (see e.g., Rockafellar, 1989 p. 45, and Theorems 17 and 18 on p. 41), it follows that there exist non-negative Lagrange multipliers $\lambda^* = \{\lambda_1^*, \ldots, \lambda_K^*\}$ such that

$$\min_{u \in \mathcal{U}} J_{tc}^{\lambda^*}(\beta, u) = \sup_{\lambda \geq 0} \min_{u \in \mathcal{U}} J_{tc}^{\lambda}(\beta, u) = \min_{u \in \mathcal{U}} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u),$$

which equals $C_{tc}(\beta)$, according to Theorem 9.9 (iii). The second equality in (9.33) follows from the fact that for fixed λ, $J^{\lambda}(\beta, u)$ is minimized by a policy in U_D (by Theorem 9.1 (ii) and and 9.2 (ii)). The Kuhn-Tucker conditions follow from standard arguments (similar to the proof of part (i) of Theorem 9.9, see Rockafellar, 1989, Theorem 15). □

Since stationary policies were shown to be dominating (Theorem 8.4) under suitable conditions, Theorem 9.9 implies the following corollary.

Corollary 9.2 (*Saddle-point*)
Consider either the transient framework (*Definition 7.1 and non-negative immediate cost*) *or an MDP with a uniform Lyapunov function and immediate costs* ν*-bounded from below. Then*

$$C_{tc}(\beta) = \sup_{\lambda \geq 0} \min_{u \in U_S} J_{tc}^{\lambda}(\beta, u) = \min_{u \in U_S} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) = \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u^*)$$

for some $u^* \in U_S$.

The above corollary relies on the fact, established in the previous chapter, that the stationary policies are dominating (Theorem 8.4). This is the only place in this section where we make use (indirectly) of the convex and compact properties of occupation measures (corresponding to stationary policies).

9.5 The Dual LP

Consider the DP with decision variables $\phi(y), y \in \mathbf{X}$ and $\lambda \in \mathbb{R}_+^K$ ($\mathbb{R}_+^K$ are vectors in $\mathbb{R}^K$ whose entries are non-negative).

$$\begin{aligned} \mathbf{DP_1}(\beta)\text{:} \qquad & \text{Find } \Theta^*(\beta) := \sup_{\phi, \lambda} \langle \beta, \phi \rangle - \langle \lambda, V \rangle \text{ subject to} \\ \phi(x) \quad \leq \quad & c(x,a) + \langle \lambda, d(x,a) \rangle + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \phi(y), \quad x \in \mathbf{X}, a \in \mathsf{A}(x). \end{aligned}$$

We begin by considering MDPs with a uniform Lyapunov function. Combining Theorem 9.4 and Theorem 9.7 with Theorem 9.9, we show that the value of $\mathbf{DP_1}(\beta)$ equals the value of **COP**. This, together with Theorem 8.6, implies that there is no duality gap between $\mathbf{LP}_\mathbf{1}^\mathrm{b}(\beta)$ and the dual program $\mathbf{DP_1}(\beta)$.

Theorem 9.11 (*The dual LP for MDPs with uniform Lyapunov function*) *Consider an MDP with a uniform Lyapunov function and ν-bounded immediate costs. Consider $\mathbf{DP_1}(\beta)$ restricted to $\psi \in \mathsf{F}^\mu$. $\mathbf{DP_1}(\beta)$ is feasible if and only if* **COP** *is feasible. The value of $\mathbf{DP_1}(\beta)$ equals $C_{tc}(\beta)$ and $\phi(x) = C_{tc}(x), x \in \mathbf{X}$ is an optimal solution.*

A similar result is obtained for the absorbing and the transient cases with non-negative immediate costs, and the case of uniform Lyapunov function with immediate costs ν-bounded from below. By combining Theorems 9.5, 9.6, 9.7 and 9.9, we have:

Theorem 9.12 (*The dual LP, unbounded immediate costs*)
Consider either
(*i*) *an MDP absorbing to $\mathcal{M}$ with non-negative immediate costs, with $\mathbf{DP_1}(\beta)$ restricted to non-negative bounded ϕ; or*
(*ii*) *an $\mathbf{X}'$-transient MDP with non-negative immediate costs, with $\mathbf{DP_1}(\beta)$ restricted to non-negative ϕ satisfying* (9.14) *(i.e., that vanish outside a finite set), or*
(*iii*) *an MDP with uniform Lyapunov function and with immediate costs ν-bounded from below, with $\mathbf{DP_1}(\beta)$ restricted to ϕ that are bounded from above.*
Then $\mathbf{DP_1}(\beta)$ is feasible ($\phi = 0$ is a feasible solution in cases (*i*) *and* (*ii*)*). Moreover, the value of $\mathbf{DP_1}(\beta)$ equals $C_{tc}(\beta)$.*

9.6 State truncation

We consider in this section transient MDPs with non-negative immediate costs. We already showed in Remark 9.4 that the value of a non-constrained MDP can be computed as the limit of the (increasing sequence of) values of the MDPs with truncated spaces, (as described in the proof of Theorem 9.6). In other words, we showed that

$$\lim_{n\to\infty} C_{tc}^n(\beta) = \sup_{n\in\mathbb{N}} C_{tc}^n(\beta) = C_{tc}(\beta),$$

where $C_{tc}^n(\beta)$ is the value of the MDP truncated to the finite set $\mathbf{X}_n$. Moreover, we showed that the optimal policies converge.

We show that a similar result holds for the constrained MDP. Indeed, for any $\lambda \geq 0$, we have by Remark 9.4:

$$\lim_{n\to\infty} J_{tc}^{\lambda,n}(\beta) = \sup_{n\in\mathbb{N}} J_{tc}^{\lambda,n}(\beta) = J_{tc}^{\lambda}(\beta) \tag{9.34}$$

where

$$J_{tc}^{\lambda,n}(\beta) = \inf_{u\in U} J_{tc}^{\lambda,n}(\beta,u), \qquad J_{tc}^{\lambda}(\beta) = \inf_{u\in U} J_{tc}^{\lambda}(\beta,u)$$

and where $J_{tc}^{\lambda,n}(\beta,u)$ is the Lagrangian defined above (9.24), corresponding to the nth-truncated MDP. According to Corollary 9.2, we have

$$C_{tc}^{n}(\beta) = \min_{u\in U}\sup_{\lambda\geq 0} J_{tc}^{\lambda,n}(\beta,u), \qquad C_{tc}(\beta) = \sup_{\lambda\geq 0}\min_{u\in U} J_{tc}^{\lambda}(\beta,u).$$

Combining this with (9.34), we have

$$C_{tc}(\beta) = \sup_{\lambda\geq 0}\sup_{n\in\mathbb{N}}\inf_{u\in U} J_{tc}^{\lambda,n}(\beta,u) = \sup_{n\in\mathbb{N}}\sup_{\lambda\geq 0}\inf_{u\in U} J_{tc}^{\lambda,n}(\beta,u) = \sup_{n\in\mathbb{N}} C_{tc}^{n}(\beta).$$

This establishes the convergence of the values for the state-truncated **COP** to the value of **COP**. Unlike previous approaches for state approximations for **COP** (most of which were derived for the contracting framework, see Chapter 16 and Altman 1993, 1994), we do not need here any Slater-type condition.

9.7 A second LP approach for optimal mixed policies

In this section we present an alternative LP formulation for **COP**. The decision variables will correspond to the probability measures over the space of all stationary deterministic policies; in particular, this will mean for the case that the state and action spaces are finite, that the number of decision variables will be equal to the number of stationary deterministic policies; this is in contrast to the previous LP approach for which the number of decision variables is typically much smaller: $\sum_{x\in\mathbf{X}} |\mathsf{A}(x)|$.

It follows from Corollary 9.1 (for the transient framework with non-negative immediate costs) and Theorem 8.4 (for the case of uniform Lyapunov function) that $C_{tc}(\beta)$ is the value of **COP** restricted to $\mathcal{U}$:

$$\min_{u\in\mathcal{U}} C_{tc}(\beta,u) \text{ subject to } D_{tc}(\beta,u) \leq V.$$

This can be rewritten as a Linear Program:

$$\begin{aligned} \mathbf{LP_2}(\beta)\text{:} \qquad & \min_{\gamma\in M_1(U_D)} \int C_{tc}(\beta,u)\gamma(du) \\ \text{subject to} \qquad & \int D_{tc}^{k}(\beta,u)\gamma(du) \leq V^k, \quad k=1,\ldots,K \end{aligned} \tag{9.35}$$

This yields the following:

Theorem 9.13 (*Relation between* **COP** *and* $\mathbf{LP_2}(\beta)$)
Consider either the case of uniform Lyapunov function and immediate costs that are ν-bounded from below, or the transient framework (*with non-negative immediate costs*). *Then*
(*i*) **COP** *is feasible if and only if* $\mathbf{LP_2}(\beta)$ *is feasible* (*i.e., the set satisfying*

(9.35) is non-empty). If $\mathbf{LP_2}(\beta)$ *is feasible, then there exists an optimal policy in* $\mathcal{U}$ *for* **COP**.
(*ii*) *The values of* **COP** *and of* $\mathbf{LP_2}(\beta)$ *are equal.*
(*iii*) *If* γ *is a solution of* $\mathbf{LP_2}(\beta)$*, then the policy* $\hat{\gamma} \in \mathcal{U}$ *is optimal for* **COP**.

9.8 More on unbounded costs

We consider all along this book immediate costs that are bounded below either by a constant (which is taken to be zero in the case of transient MDPs) or by some function (whose infimum may be $-\infty$) that satisfies the uniform Lyapunov conditions (Definition 7.5).

To illustrate the importance of these types of assumptions, we briefly describe some phenomena that arise in MDPs in which the boundedness assumption of the cost is dropped. The following example is due to Van Der Wal (1981a).

Example 9.2 (*Costs unbounded from below*)
Consider the following MDP:

- **State space:** $\mathbf{X} = \{0, 1, 2, \ldots\}$.
- **Action space:** $\mathsf{A} = \{1, 2\}$.
- **Transition probabilities:**

$$\mathcal{P}_{xay} = \begin{cases} 1 & \text{if } x = y = 0, \\ 1/2 & \text{if } y = 0, x > 0, a = 1, \\ 1/2 & \text{if } y = x + 1, x > 0, a = 1, \\ 1 & \text{if } a = 2, y = 0, \\ 0 & \text{otherwise.} \end{cases}$$

- **Immediate costs:**

$$c(x, a) = \begin{cases} 0 & \text{if } x = 0, \\ 0 & \text{if } a = 1, \\ -2^x + 1 & \text{if } a = 2. \end{cases}$$

Thus, at state x the system either goes to state 0 and the cost is $-2^x + 1$ as a result of action 2, or no cost is incurred and the system moves with equal probabilities to states 0 and $x + 1$, as a result of action 1.

Consider the problem of minimizing the total expected cost until state 0 is reached. Note that this MDP is absorbing to the set $\{0\}$.

Clearly, the value is $C(x) = -2^x$. However,

- The stationary deterministic policies are not optimal, nor even ε-optimal at states $x > 0$. (A set of policies $\overline{U}$ is called ε-optimal at x if for every $\varepsilon > 0$ there exists some policy $u \in \overline{U}$ such that $C_{tc}(x, u) \leq C_{tc}(x) + \varepsilon$). Indeed, if $g \in U_D$ chooses action 1 at all states, then $C(x, g) = 0$.

Otherwise, if for at least one state y it chooses 2 (i.e., $g(y) = 2$), then $C(y, g) = -2^y + 1 = C(y) + 1$.

- The value $C(x)$ satisfies the optimality equation (9.1) with equality. However, it follows from the previous point that the stationary policy that chooses the argmin in (9.1) is not optimal.

We conclude from the above example that the analog of Theorems 9.1 (ii) and 9.2 (ii) do not hold. In fact, *an optimal policy does not exist* for any initial state $x > 0$. Indeed, choose an arbitrary policy u. If it chooses with probability 1 action 1 at all times, then we know it is not optimal. If with positive probability it does not choose action 1 at all times, then consider the following policy v: at the first step, v chooses action 1. Then, if a transition to 0 does not occur, then for all the following steps t, the policy v behaves as follows. At step $t + 1$, v uses the action that policy u uses at time t. It can be shown that v has a cost strictly lower than the cost obtained by u.

Dynamic programming related to An MDP in which the immediate costs are non-negative but the total expected cost is maximized is called positive dynamic programming. Some interesting properties of such (non-constrained) problems are known (see e.g., Hordijk, 1974, and Van Der Wal, 1981a). We formulate these in terms of the equivalent problem of minimization of the total expected cost with non-positive immediate costs:

1. The stationary deterministic policies are not optimal (nor even ε-optimal).
2. Optimal policies need not exist.
3. For each state x, the stationary randomized policies U_S are ε-optimal.
4. We say that a set of policies $\overline{U}$ is ε-uniform optimal if for every $\varepsilon > 0$ there exists a policy $u \in \overline{U}$ such that for all states x, $C_{tc}(x, u) \leq C_{tc}(x) + \varepsilon$. Then neither the the stationary policies U_S nor the Markov policies U_M are uniformly optimal.
5. The stationary policies U_S are still "uniformly good" in the following sense: for every $\varepsilon > 0$ there exists some $g \in U_S$ such that for all states x, $C_{tc}(x, u) \leq C_{tc}(x)(1 + \varepsilon)$.

CHAPTER 10

The discounted cost

A simple and natural way to treat the discounted cost is to transform it into a total cost problem until some new dummy state is reached. We shall use this approach in this chapter to obtain the results corresponding to those obtained for the total cost problem. We shall illustrate the results by extending the model of Chapter 5 to the infinite buffer case (i.e., $L = \infty$).

10.1 The equivalent total cost model

We consider a discounted cost criterion for an MDP with a state space $\mathbf{X}_\alpha$, transition probabilities $\mathcal{P}^\alpha$, and a discount factor α. The equivalent total cost model is obtained by adding an extra state x^o; we are then interested in the total cost until x^o is reached. The probability to move from any state in $\mathbf{X}_\alpha$ to x^o is equal to $1-\alpha$ for any action. We summarize this in a formal way:

- The state space is given by $\mathbf{X} = \mathbf{X}_\alpha \cup \{x^o\}$, where x^o is some additional dummy state; and $\mathbf{X}' := \mathbf{X}_\alpha$, $\mathcal{M} = \{x^o\}$. The action space is unchanged.
- The transition probabilities are

$$\mathcal{P}_{xay} = \begin{cases} \alpha \mathcal{P}^\alpha_{xay} & \text{if } x, y \in \mathbf{X}_\alpha \\ 1-\alpha & \text{if } x \in \mathbf{X}_\alpha, y = x^o \\ 1 & \text{if } x = y = x^o \\ 0 & \text{otherwise .} \end{cases}$$

- There is only one dummy action a^o available at state x^o, i.e., $\mathsf{A}(x^o) = \{a^o\}$, and $c(x^o, a^o) = d^k(x^o, a^o) = 0,\ k = 1, \ldots, K$. (The immediate costs in other states are unchanged.)
- The normalization is obtained by setting the new initial distribution

$$\beta_\alpha(y) = \begin{cases} (1-\alpha)\beta & \text{for } y \neq x^o \\ \alpha & \text{for } y = x^o. \end{cases}$$

Next, we observe that in the equivalent total cost model,

$$\sum_{t=1}^{\infty} p^u_\beta(t, \mathbf{X}_\alpha) = 1$$

for any policy u and initial distribution β on $\mathbf{X}_\alpha$, so that the equivalent MDP is $\mathbf{X}_\alpha$-absorbing. One can also check that

Lemma 10.1 *$\mu(x) = 1/(1-\alpha), x \in \mathbf{X}$, is a uniform Lyapunov function for the total expected life-time (Definition 7.4) for the new MDP.*

10.2 Occupation measure and LP

The occupation measure for the discounted cost is defined as

$$f_\alpha(\beta, u; x, \mathcal{A}) := (1-\alpha)\sum_{t=1}^{\infty} \alpha^{t-1} p_\beta^u(t; x, \mathcal{A}), \qquad x \in \mathbf{X}_\alpha, \mathcal{A} \subset \mathsf{A}(x).$$

Let $f_\alpha(\beta, u)$ be the probability measure on $(\mathcal{K}', \mathbb{K}')$ given by $f_\alpha(\beta, u)(x, \mathcal{A})$ $:= f_\alpha(\beta, u; x, \mathcal{A})$. (Here $\mathcal{K}' := \{(x, a) : x \in \mathbf{X}_\alpha, a \in \mathsf{A}(x)\}$, and $\mathbb{K}'$ is its Borel σ-field.)

As for the total cost model, define for any class of policies $\overline{U}$

$$\mathbf{L}_{\overline{U}}^\alpha(\beta) = \bigcup_{u \in \overline{U}} f_\alpha(\beta, u), \tag{10.1}$$

and define $\mathbf{L}^\alpha := \mathbf{L}_U^\alpha \cup \mathbf{L}_{\overline{M}(U_M)}^\alpha$. A class of policies $\overline{U}$ is said to be complete with respect to the discounted cost problem if $\mathbf{L}^\alpha = \mathbf{L}_{\overline{U}}^\alpha$. Define

$$\mathbf{Q}^\alpha(\beta) = \tag{10.2}$$
$$\left\{ \begin{array}{c} \rho \in M(\mathcal{K}') : \sum_{y \in \mathbf{X}_\alpha} \int_{\mathcal{A}(y)} \rho(y, da)(\delta_x(y) - \alpha \mathcal{P}_{yax}) = (1-\alpha)\beta(x), \\ \\ x \in \mathbf{X}_\alpha, \qquad \rho(x, \mathsf{A}(x)) < \infty \text{ for } x \in \mathbf{X}_\alpha. \end{array} \right\}$$

Define $\mathbf{Q}^{\alpha,\mathrm{b}}(\beta) \stackrel{\mathrm{def}}{=}$ the subset of finite measures among $\mathbf{Q}^\alpha(\beta)$. Using the above equivalent absorbing total cost model, we may apply Theorems 8.1 (ii), Theorem 8.2 and Lemma 10.1 to conclude that

Corollary 10.1 (*Properties of occupation measures*)
The set of stationary policies is complete. Moreover, $\mathbf{L}_{U_S}^\alpha(\beta)$ is convex and compact, and satisfies

$$\mathbf{L}^\alpha(\beta) = \mathbf{L}_{\mathcal{U}}^\alpha(\beta) = \mathbf{L}_U^\alpha(\beta) = \mathbf{L}_{U_S}^\alpha(\beta) = \overline{co}\mathbf{L}_{U_D}^\alpha(\beta) = \min \mathbf{Q}^\alpha(\beta) = \mathbf{Q}^{\alpha,\mathrm{b}}(\beta).$$

Since the equivalent total cost MDP (obtained from the original discounted cost one) is $\mathbf{X}_\alpha$-absorbing, all results obtained for the total cost under the assumption that the immediate costs are non-negative hold for the discounted cost as well.

10.3 Non-negative immediate cost

Since the equivalent MDP is absorbing (and thus transient), we recover for non-negative immediate cost all the results from Lemma 8.4, as well as Theorems 8.3, 8.4 and 8.5. (Everywhere in these lemmas and theorems,

the subscript tc should be replaced by the subscript α.) In particular, the equivalent LP (that corresponds to the one in (8.18)) becomes
$\mathbf{LP}_{\mathbf{1}}^{\alpha}(\beta)$: Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle \rho, c \rangle$ subject to:

$$\mathcal{D}^k(\rho) := \langle \rho, d^k \rangle \leq V_k, k = 1, \ldots, K, \qquad \rho \in \mathbf{Q}^{\alpha}(\beta). \tag{10.3}$$

We may again obtain a Lagrangian formulation and recover the corresponding saddle point and optimality results from Section 9.4. This leads us again to results on the dual LP and on the duality gap, as in Section 9.5, and to state truncation techniques, as in Section 9.6.

For the dual LP, the decision variables are $\phi : \mathbf{X}_{\alpha} \to \mathbb{R}$, which are restricted to be non-negative, and the K-dimensional non-negative vectors $\lambda \in \mathbb{R}_{+}^{K}$. We have:

$$\begin{aligned} \mathbf{DP}_{\mathbf{1}}^{\alpha}(\beta): \qquad & \text{Find } \Theta^* := \sup_{\phi,\lambda} \langle \beta, \phi \rangle - \langle \lambda, V \rangle \text{ subject to} \\ \phi(x) \quad \leq \quad & (1-\alpha)(c(x,a) + \langle \lambda, d(x,a) \rangle) + \alpha \sum_{y \in \mathbf{X}_{\alpha}} \mathcal{P}_{xay} \phi(y), \\ & x \in \mathbf{X}_{\alpha}, a \in \mathsf{A}(x). \end{aligned}$$

10.4 Weak contracting assumptions and Lyapunov functions

We formulate the conditions for the equivalent total cost MDP to be contracting in terms of the original discounted MDP, and thus to have a uniform Lyapunov function for the total expected cost. We then present the conditions for the immediate costs to be ν-bounded (or ν-bounded from below) for $\nu = \mu$.

By using (10.2), the following condition on the discounted MDP will imply that the equivalent total cost MDP is $\mathbf{X}_{\alpha}$-contracting (i.e., (7.34) will hold): some scalar $\xi \in [0, 1)$, a vector $\mu : \mathbf{X}_{\alpha} \to [1, \infty)$, and a finite set $\mathcal{M}_{\alpha}$ exist, such that for all $x \in \mathbf{X}, a \in \mathsf{A}$,

$$\alpha \sum_{y \notin \mathcal{M}_{\alpha}} \mathcal{P}_{xay}^{\alpha} \mu(y) \leq \xi \mu(x). \tag{10.4}$$

We call this the *weak contracting condition*. In that case, the equivalent total cost MDP is contracting with the same ξ and μ, and with $\mathcal{M} := \mathcal{M}_{\alpha} \cup \{x^o\}$. In many applications, the set $\mathcal{M}_{\alpha}$ can be chosen to be an empty set.

A sufficient condition for (10.4) to hold is that the original MDP has a uniform Lyapunov function (ULF) μ with respect to the sets $\mathcal{M}, \mathbf{X}'$, with $\mathcal{M} \cup \mathbf{X}' = \mathbf{X}_{\alpha}$. Indeed, since μ is a ULF, it follows that it remains a ULF if we replace $\mu(x)$ by 1 for all $x \in \mathcal{M}$. So we shall assume without loss of generality that $\mu(x) = 1$ for $x \in \mathcal{M}$. We have

$$\alpha \sum_{y \in \mathbf{X}_{\alpha}} \mathcal{P}_{xay} \mu(y) \quad = \quad \alpha \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \mu(y) + \alpha \sum_{y \in \mathcal{M}} \mathcal{P}_{xay} \mu(y)$$

$$\begin{aligned} &\leq \alpha(\sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}\mu(y) + 1) \\ &\leq \alpha\mu(x). \end{aligned}$$

The last inequality follows since μ is a uniform Lyapunov function. We thus conclude that the weak contracting condition indeed holds if the original MDP has a uniform Lyapunov function, with $\xi = \alpha$, $\mathcal{M}_\alpha = \mathcal{M}$, and $\mu_\alpha(x) = \mu(x)1\{x \in \mathbf{X}'\} + 1\{x \in \mathcal{M}\}$.

We conclude:

Theorem 10.1 (*Conditions for contracting and uniform Lyapunov function*)
Assume that either one of the following assumptions hold:
(*i*) *The original MDP has a uniform Lyapunov function for the total expected time (Definition 7.4) and the immediate costs are μ-bounded,*
(*ii*) *The original MDP is weakly contracting, i.e., it satisfies Definition 7.9 of contracting MDPs with (10.4) replacing (7.34).*
Then the equivalent MDP is contracting. If
(*iii*) *Condition (ii) holds with the immediate costs only μ-bounded below, then the new MDP has a uniform Lyapunov function with ν proportional to μ, and the immediate costs are μ-bounded below.*

Under assumptions (i) or (ii) or (iii) of the theorem, it now follows that Lemma 8.5 as well as Theorems 8.3, 8.4 and 8.6, hold. The equivalent LP (that corresponds to the one in (8.18)) is again (10.3), where the decision variables ρ are constrained to lie in $\mathbf{Q}^\alpha(\beta)$ The corresponding results for the Lagrangian approach and dual LP follow too. In particular, Theorems 9.2, 9.3, 9.4 and 9.7 hold for the case of no constraints. The relation to the Lagrangian is given in Theorem 9.9 and in the other theorems in that section. The corresponding dual LP is again $\mathbf{DP}_1^\alpha(\beta)$, where ϕ is restricted to F^μ under conditions (i) or (ii) of Theorem 10.1, and is restricted to functions that are bounded from above, under condition (iii).

Remark 10.1 (*Equivalence between contracting and discounted CMDPs*)
We established in this section the theory of discounted cost problem as a special case of the total cost problem. It turns out that the converse is also true for the special case of contracting MDPs. Indeed, Van Der Wal (1981a) has shown (p. 101) that the contracting total cost problem with μ-bounded immediate cost is equivalent to a discounted cost problem with bounded cost.

10.5 Example: flow and service control

We consider again the model and notation of Chapter 5, this time with $L = \infty$. The model is unchanged, except for the following:

- Unlike the finite case, we do not make here the assumption that $0 \in \mathsf{B}(x)$.

(This assumption was needed to check that some properties hold at the boundary $x = L$.)

- The immediate cost c is polynomially bounded.
- $\mathcal{N}$ is defined here to be the set of polynomially bounded functions on $\mathbf{X}$.

Theorem 10.2 (*Structure of optimal policies*)
Under the above conditions, all statements of Theorem 5.1 hold for the case $L = \infty$ *as well.*

Proof. We only point where the proof changes with respect to that of Theorem 5.1.

(5.1) follows from Corollary 9.2, which is used together with Theorem 9.9 whenever we used Theorem 3.6 in Chapter 5.

We now establish Lemma 5.2 for $L = \infty$. We shall show that the weak contracting condition (10.4) holds with respect to some function μ, where we let $\mathcal{N}$ correspond to functions that are μ-bounded. This implies that the discounted cost is equivalent to a total cost *contracting* MDP. (i) then follows from Theorem 9.2 (i), and (ii) follows from Theorem 9.2 (ii). Finally, (iii) follows from a well-known value iteration theorem, see e.g., Wessels (1977) (for more details, see Chapter 15).

It remains to check the weak contraction. Choose some ξ, $\alpha < \xi < 1$, and denote $\hat{\xi} := \alpha/\xi$. We have to show that for all x, a, b,

$$\hat{\xi} R(x, a, b, \mu)(x) \leq \mu(x).$$

Define

$$r \stackrel{\text{def}}{=} 1 + \frac{\hat{\xi}^{-1} - 1}{\bar{a}b}, \qquad \mu(x) \stackrel{\text{def}}{=} r^x.$$

Then

$$\hat{\xi} R(x, a, b, \mu)(x) - \mu(x) =$$

$$\begin{cases} \left[\bar{a}br^2 + (ab + \bar{a}\bar{b} - \hat{\xi}^{-1})r + a\bar{b}\right] \hat{\xi} r^{x-1} = -\hat{\xi} r^{x-1} a\bar{b}(r-1) < 0, & x > 0 \\ \\ \hat{\xi}[\bar{a}br + (1 - \bar{a}b - \hat{\xi}^{-1})] = 0, & x = 0. \end{cases}$$

This establishes Lemma 5.2.

The proof of the theorem is now the same as that of Theorem 5.1 except that we use Theorem 9.2 (ii) instead of Theorem 3.4. □

CHAPTER 11

The expected average cost

We study in this chapter the expected average cost. Just as for the total cost, we shall be especially interested in the following frameworks:
(i) The case for which the costs are bounded below (known as negative dynamic programming)

$$\begin{array}{l} c \text{ and } d^k, k = 1, \ldots, K, \text{ are bounded below, i.e.,} \\ \inf_{\kappa \in \mathcal{K}} c(\kappa) \geq \underline{b} \text{ and } \inf_{k, \kappa \in \mathcal{K}} d^k(\kappa) \geq \underline{b} \text{ for some constant } \underline{b}. \end{array} \tag{11.1}$$

An additional growth condition on the cost will often be made. This case exhibits features similar to those in the total cost with non-negative immediate cost (and a transient framework).
(ii) The case for which the occupation measures are tight and some uniform integrability conditions of the immediate costs hold. This will be directly related to the uniform Lyapunov framework (Section 11.9), and will exhibit features similar to the corresponding ones in the total cost criterion.

We shall assume throughout this chapter that

$$\begin{array}{ll} (B1) & \text{Under any } w \in U_S, \mathbf{X} \text{ contains a single (aperiodic) ergodic} \\ & \text{class, and absorption into the positive recurrent class} \\ & \text{takes place in a finite expected time (that may depend on } w). \end{array} \tag{11.2}$$

Remark 11.1 (i) Note that this assumption may restrict the choice of the initial distribution β. To see that, fix some stationary policy w; even if for any fixed initial state x, absorption into the positive recurrent class takes place in a finite expected time (that may depend on the initial state x), the choice of β may render the expected absorption time infinite.
(ii) Sufficient and necessary conditions for (11.2) in terms of policies in U_D can be found in Fisher (1968).

11.1 Occupation measure

For any given initial distribution β and policy u, define the finite horizon occupation measure $f^t_{ea}(\beta, u; x, \cdot)$

$$f^t_{ea}(\beta, u; x, \mathcal{A}) = \frac{1}{t} \sum_{s=1}^{t} P^u_\beta(X_s = x, A_s \in \mathcal{A}), \qquad \mathcal{A} \subset \mathsf{A}(x). \tag{11.3}$$

Let $f^t_{ea}(\beta, u)$ be the probability measure on $(\mathcal{K}, \mathbb{K})$ generated by the rectangles $(x, \mathcal{A})$, $(\mathcal{A} \subset \mathsf{A}(x))$ such that $f^t_{ea}(\beta, u)(x, \mathcal{A}) := f^t_{ea}(\beta, u; x, \mathcal{A})$. With some abuse of notation, we define $f^t_{ea}(\beta, u; x) = f^t_{ea}(\beta, u; x, \mathsf{A}(x))$. The subscript *ea* stands for *expected average*.

We denote by $F_{ea}(\beta, u)$ the non-empty compact set obtained as all the limits, in the sense of vague convergence of measures, of $\{f^t_{ea}(\beta, u)\}$ (see Definition 17.1 in the appendix). Any subprobability measure on $\mathcal{K}$ can be written as

$$f = \Delta_f f' \tag{11.4}$$

where $\Delta_f \in [0, 1]$, and where f' is a probability measure. Define,

$$\mathcal{L}_{\overline{U}}(\beta) = \bigcup_{u \in \overline{U}} \{F_{ea}(\beta, u)\} \text{ for any set of policies } \overline{U},$$

$$\mathbf{Q}_{ea}(\beta) = \left\{ \rho \in M_1(\mathcal{K}) : \sum_{y \in \mathbf{X}} \int_{\mathsf{A}(x)} \rho(y, da)(\delta_x(\{y\}) - \mathcal{P}_{yax}) = 0, x \in \mathbf{X} \right\}, \tag{11.5}$$

where $M_1(\mathcal{K})$ are the set of probability measures over $\mathcal{K}$, and δ_x is the Dirac probability measure concentrated on x. We set $\mathcal{L}(\beta) = \mathcal{L}_U(\beta) \cup \mathcal{L}_{\overline{M}(U_M)}(\beta)$. $\mathcal{L}_{\overline{U}}(\beta)$ is called the set of expected occupation measures achievable by $\overline{U}$.

Definition 11.1 *(Completeness and weak completeness, expected average cost)*
A class of policies $\overline{U}$ is called complete for the expected average cost criterion (for a given initial distribution β) if

$$\mathcal{L}(\beta) = \mathcal{L}_{\overline{U}}(\beta) \text{ and } \forall u \in \overline{U},\ F_{ea}(\beta, u) \text{ is a singleton.}$$

It is called weakly complete if

$$\mathcal{L}(\beta) \cap M_1(\mathcal{K}) = \mathcal{L}_{\overline{U}}(\beta)$$

$$\text{and } \forall u \in \overline{U},\ F_{ea}(\beta, u) \text{ is a singleton.}$$

Thus a complete class of policies $\overline{U}$ has the property that the achievable expected occupation measures under $\overline{U}$ is the same as under all policies. A weakly complete class of policies achieves all those expected occupation for which the measure of $\mathcal{K}$ is one.

Definition 11.2 *For any sets B_1, B_2 of subprobability measures on $\mathcal{K}$, define $B_1 \propto B_2$ if $\forall f_1 \in B_1$ there exists $f_2 \in B_2$ such that $f'_1 = f'_2$ and $\Delta_{f_1} \leq \Delta_{f_2}$ (where f' and Δ_f are defined in (11.4)).*

Theorem 11.1 (*Weak completeness of stationary policies*)
The stationary policies are weakly complete and $\mathcal{L}_U(\beta) \propto \mathcal{L}_{U_S}(\beta)$.

Proof. Choose a policy $u \in U$. Let t_n be some increasing sequence of times along which $f^t_{ea}(\beta, u)$ converges vaguely to some limit $f \in F_{ea}(\beta, u)$. Define γ that maps states y to measures over $\mathsf{A}(y)$:

$$\gamma_y(\mathcal{A}) = \frac{f(y, \mathcal{A})}{f(y, \mathsf{A}(y))}, \ \mathcal{A} \subset \mathsf{A}(y)$$

whenever the denominator is non-zero. When it is zero, $\gamma_y(\cdot)$ is chosen arbitrarily. Define the stationary policy w as $w_x(\mathcal{A}) = \gamma_x(\mathcal{A})$. It follows from assumption (11.2) that the Markov chain with transition probabilities $P(w)$ has a unique invariant probability measure $\pi(w)$, independent of the initial distribution β, that satisfies

$$\pi_y(w) = \lim_{t\to\infty} f^t_{ea}(\beta, w; y),$$

and hence, $F_{ea}(\beta, w) = \{f^w\}$ is a singleton and it satisfies

$$f^w(y, \mathcal{A}) = w_y(\mathcal{A})\pi_w(y). \tag{11.6}$$

We show that $f^w = \Delta f$ for some $\Delta \in [0, 1]$. It follows from (6.4) (when setting $\mathcal{M} = \emptyset\}$) that for any $x \in \mathbf{X}$,

$$f^t_{ea}(\beta, u; x) - \frac{\beta(x)}{t} = \int_{\mathcal{K}} f^t_{ea}(\beta, u; d\kappa)\mathcal{P}_{\kappa x} - \frac{\int_{\mathcal{K}} p^u_\beta(t; d\kappa)\mathcal{P}_{\kappa x}}{t}. \tag{11.7}$$

By applying Lemma 17.2(i) in the appendix, we get from (11.7)

$$f(x, \mathsf{A}(x)) = \lim_{n\to\infty} f^{t_n}_{ea}(\beta, u; x) = \lim_{n\to\infty} \int_{\mathcal{K}} f^{t_n}_{ea}(\beta, u; d\kappa)\mathcal{P}_{\kappa x} \geq \int_{\mathcal{K}} f(d\kappa)\mathcal{P}_{\kappa x}. \tag{11.8}$$

By definition of γ and of $P_{xy}(w)$,

$$\int_{\mathcal{K}} f(d\kappa)\mathcal{P}_{\kappa x} = \sum_y f(y, \mathsf{A}(y)) \int_{\mathsf{A}(y)} \gamma_y(da)\mathcal{P}_{yax} = \sum_y f(y, \mathsf{A}(y))P_{yx}(w), \tag{11.9}$$

which, together with (11.8), leads to

$$f(x, \mathsf{A}(x)) \geq \sum_y f(y, \mathsf{A}(y))P_{yx}(w). \tag{11.10}$$

Measures satisfying (11.10) are called excessive measures; $\pi(w)$ is known to be the unique probability measure over $\mathbf{X}$ satisfying the inequality (11.10), (this is a straightforward extension of Proposition 6.4 in Kemeney, Snell and Knapp (1976), see Altman and Shwartz (1991a)). This, together with the definition of γ, implies that $\{f\} \propto \{f^w\}$, which establishes the proof. □

Lemma 11.1 (*Tightness of occupation measure*)
$\{f^t_{ea}(\beta, u)\}_{t\in\mathbb{N}}$ are tight for any $u \in U_S$ and any $u \in \mathcal{U}$.

Proof. Assumption (11.2) implies that the stationary state probabilities $\pi(u)$ exist under any $u \in U_S$ and $P^n(u)$ converge to $\pi(u)$ in total variation. Hence $f^t_{ea}(\beta, u; y, \mathcal{A})$ converge weakly to $\pi_u(y)u_y(\mathcal{A})$. Lemma 17.2(ii) (in the appendix) then implies the tightness of $\{f^t_{ea}(\beta, u)\}_{t\in\mathbb{N}}$.

The claim for $\mathcal{U}$ follows from the bounded convergence theorem (Royden, 1988, Proposition 11.18): for any $y \in \mathbf{X}$, measurable $\mathcal{A} \subset \mathsf{A}(y)$ and $u \in U_S$, $f^t_{ea}(\beta, u; y, \mathcal{A})$ is bounded by 1 and converges weakly to $\pi_u(y)u_y(\mathcal{A})$. Hence for any $\gamma \in M_1(U_S)$, the following limit exists:

$$\begin{aligned} & f_{ea}(\beta, \hat{\gamma}; y, \mathcal{A}) \qquad (11.11)\\ = \; & \lim_{t\to\infty} f^t_{ea}(\beta, \hat{\gamma}; y, \mathcal{A}) = \lim_{t\to\infty} \langle \gamma, f^t_{ea}(\beta, \cdot; y, \mathcal{A})\rangle = \langle \gamma, f_{ea}(\beta, \cdot; y, \mathcal{A})\rangle. \end{aligned}$$

This implies that $f_{ea}(\beta, \hat{\gamma})$ is a probability measure, and the result follows from Lemma 17.2(ii) (in the appendix). □

Theorem 11.2 (*Continuity of occupation measures*)
(*i*) *Let* $\overline{U}$ *be some subset of* U_S. *Assume that* $\mathcal{L}_{\overline{U}}(\beta)$ *are tight. Then* $f : \mathcal{L}_{\overline{U}} \to \mathbb{R}$ *are weakly continuous over* $\overline{U}$.
(*ii*) *Assume that* $\mathcal{L}_{U_D}(\beta)$ *are tight. Then* $f : \mathcal{L}_{\mathcal{U}} \to \mathbb{R}$ *are weakly continuous over* $\mathcal{U}$.

Proof. (i) For any $u \in U_S$ we use $f_{ea}(\beta, u)$ to denote the singleton $F_{ea}(\beta, u) = \{f_{ea}(\beta, u)\}$. Let u^n be a sequence of stationary policies converging to u (i.e., $u^n(\cdot \mid x)$ converge weakly to $u(\cdot \mid x)$ for all $x \in \mathbf{X}$).
Let $n(i)$ be an increasing sequence of integers. Since $\mathcal{L}_{U_S}(\beta)$ are tight, it follows that there is a subsequence $n(i(j))$ of $n(i)$ along which a limit (in the sense of weak convergence of probability measures)

$$f' = \lim_{j\to\infty} f_{ea}(\beta, u^{n(i(j))})$$

exists (Prohorov's Theorem, see Billingsley, 1968, p. 37). Since u^n converges to u, it also follows that the transition probabilities $P(u^n)$ converge to $P(u)$ (pointwise). Since all entries of $P(u^n)$ are bounded by 1, it follows from Fatou's Lemma that for all $x \in \mathbf{X}$,

$$\lim_{j\to\infty} \sum_{y\in\mathbf{X}} f_{ea}(\beta, u^{n(i(j))}; y)[P(u^n(i(j)))]_{yx} \geq \sum_{y\in\mathbf{X}} f'(y)[P(u)]_{yx}.$$

Since for each n and $y \in \mathbf{X}$,

$$f_{ea}(\beta, u^n; x) = \sum_{y\in\mathbf{X}} f_{ea}(\beta, u^n; y)[P(u^n)]_{yx},$$

we obtain

$$f'(x, \mathsf{A}(x)) \geq \sum_{y\in\mathbf{X}} f'(y, \mathsf{A}(y))[P(u)]_{yx}, \quad f' \in M_1(\mathcal{K}). \qquad (11.12)$$

Due to assumption (11.2), π_u is the unique solution to $\pi \geq \pi P(u)$, $\pi \in$

$M_1\mathbf{X}$. Hence $f'(y, \mathsf{A}(y)) = \pi_u$, (see Theorem 1.10 of Revuz, 1975, p. 67, or Lemma 3.1 in Altman and Shwartz, 1991a).

Choose any bounded continuous function $c' : \mathcal{K} \to \mathbb{R}$. In order to establish the proof, we have to show that

$$\lim_{j\to\infty} \langle f_{ea}(\beta, u^{n(i(j))}), c' \rangle = \langle f_{ea}(\beta, u), c' \rangle. \tag{11.13}$$

Since $u^n \to u$, it follows that for every y,

$$\lim_{j\to\infty} \int_{\mathsf{A}(y)} \pi_{u^{n(i(j))}}(da \mid y) c'(y, a) = \int_{\mathsf{A}(y)} \pi_u(da \mid y) c'(y, a).$$

This implies (11.13) by the bounded convergence theorem, since

$$\langle f_{ea}(\beta, u^n), c' \rangle = \sum_{y\in\mathbf{X}} f_{ea}(\beta, u^n; y) \int_{\mathsf{A}(y)} \pi_{u^n}(da \mid y) c'(y, a).$$

(ii) Follows as the proof of the second part of Lemma 8.1 (ii). □

The following set of assumptions will turn out to be especially useful for the case that the immediate costs are not bounded below.

- **(B2(u))** Given a policy u, the expected occupation measures $\{f^t_{ea}(\beta, u)\}_t$ are tight.
- **(B2)** Assumption B2(u) holds for all policies u.
- **(B2*)** The family of stationary probabilities $\{\pi_u, u \in U_{SD}\}$ is tight.

There are many known sufficient conditions for assumptions (B2), (B2*), see e.g., Section 4 of Altman and Shwartz (1991a). We note also that (B2*) is equivalent to $\mathcal{L}_{U_S}(\beta)$ being tight; this follows from Lemma 17.1 and Lemma 17.2(ii) (in the appendix): For any $u \in U_S$, $y \in \mathbf{X}, \mathcal{A} \subset \mathsf{A}(y)$, we have $f^t_{ea}(\beta, u; y, \mathcal{A}) = f_{ea}(\beta, u; y) u_y(\mathcal{A})$. Hence the following weak limit exists and satisfies

$$f_{ea}(\beta, u, y\mathcal{A}) \stackrel{\text{def}}{=} \lim_{t\to\infty} f^t_{ea}(\beta, u; y, \mathcal{A}) = \pi_u(y) u_y(\mathcal{A}). \tag{11.14}$$

Assume that (B2*) holds. Let $u(n)$ be an arbitrary sequence of stationary policies. Let $n[j]$ be a subsequence along which $\pi_{u(n[j])}$ weakly converge to some probability measure f (that such a subsequence can be chosen follows from Lemma 17.1 in the appendix). (11.14) implies that $\lim_{j\to\infty} f_{ea}(\beta, u(n[j]); \mathcal{K}) = f(\mathcal{K}) = 1$, and hence, by applying again Lemma 17.2(ii) (in the appendix), we see that $\{f_{ea}(\beta, u(n))\}_n$ (and hence $\mathcal{L}_{U_S}(\beta)$) are tight. The converse follows from the same argument.

11.2 Completeness properties of stationary policies

Theorem 11.3 (*Completeness of stationary policies*)
(*i*) $\mathcal{L}_{U_S}(\beta)$ *and* $\mathcal{L}(\beta)$ *are convex and satisfy*

$$\mathcal{L}(\beta) = \mathcal{L}_{U_M}(\beta) \propto \mathcal{L}_{U_S}(\beta) = \mathbf{Q}_{ea}(\beta).$$

(ii) Under Assumption (B2), $\mathcal{L}_{U_S}(\beta)$ is convex, compact and tight, and satisfies

$$\mathcal{L}_{\mathcal{U}}(\beta) = \mathcal{L}(\beta) = \mathcal{L}_{U_S}(\beta) = \overline{co}\mathcal{L}_{U_D}(\beta) = \mathbf{Q}_{ea}(\beta).$$

Hence the stationary policies are complete.

In order to prove the theorem, we need the following lemma (that corresponds to Lemma 8.2 in the case of total cost). Its proof is a straightforward extension of the proof that we presented for the finite case (in the proof of Theorem 4.2).

Lemma 11.2 (*Splitting in a state*)
Choose $w \in U_S$ and a state y. Define $w^a \in U_S$ to be the policy that always chooses action a when in state y, and otherwise behaves exactly like w. Then, there exists a probability measure γ over $\mathsf{A}(y)$ such that

$$f_{ea}(\beta, w) = \int_{\mathsf{A}(y)} \gamma(da) f_{ea}(\beta, w^a)$$

(where $f_{ea}(\beta, w^a) = \lim_{t\to\infty} f^t_{ea}(\beta, w^a)$).

Proof of Theorem 11.3: (i) Theorem 6.1 implies that $\mathcal{L}(\beta)$ is convex, and that $\mathcal{L}(\beta) = \mathcal{L}_{U_M}(\beta)$. Theorem 11.1 implies that $\mathcal{L}_{U_M}(\beta) \propto \mathcal{L}_{U_S}(\beta)$. Since for each $w \in U_S$, (11.10) is obtained with equality (see paragraph below (11.10)), it follows from (11.9) and (11.10) that $f_{ea}(\beta, w) \in \mathbf{Q}_{ea}(\beta)$. It remains to show the converse. For any $\rho \in \mathbf{Q}_{ea}(\beta)$, define again γ that maps states y to measures over $\mathsf{A}(y)$:

$$\gamma_y(\mathcal{A}) = \frac{\rho(y, \mathcal{A})}{\rho(y, \mathsf{A}(y))},\ \mathcal{A} \subset \mathsf{A}(y)$$

whenever the denominator is non-zero. When it is zero, $\gamma_y(\cdot)$ is chosen arbitrarily. Define the stationary policy w as $w_x(\mathcal{A}) = \gamma_x(\mathcal{A})$. It follows from the definition of $\mathbf{Q}_{ea}(\beta)$ and of γ that for all $x \in \mathbf{X}$,

$$\rho(x, \mathsf{A}(x)) = \sum_{y\in\mathbf{X}} \rho(y, \mathsf{A}(y)) \int_{\mathsf{A}(y)} \gamma_y(da)\mathcal{P}_{yax} = \sum_{y\in\mathbf{X}} \rho(y, \mathsf{A}(y)) P_{yx}(w).$$

Since $\pi_y(w) = f_{ea}(\beta, w; y),\ y \in \mathbf{X}$ is the unique solution to $\pi = \pi P(w)$ that satisfies $\pi(\mathbf{X}) = 1, \pi \geq 0$, it follows that $\rho(x, \mathsf{A}(x)) = f_{ea}(\beta, u; x)$ for all $x \in \mathbf{X}$, and by the definition of γ, $\rho = f_{ea}(\beta, u)$. This establishes $\mathcal{L}_{U_S}(\beta) = \mathbf{Q}_{ea}(\beta)$.

Next, we show that $\mathcal{L}_{U_S}(\beta)$ is convex. Choose an arbitrary constant $0 < \alpha < 1$ and choose $f_1, f_2 \in \mathcal{L}_{U_S}$. Since $\mathcal{L}(\beta)$ is convex, $f := \alpha f_1 + (1 - \alpha) f_2 \in \mathcal{L}(\beta)$. Moreover, since $f_1, f_2 \in M_1(\mathcal{K})$, it follows that $f \in M_1(\mathcal{K})$. Theorem 11.1 then implies that there exists a stationary policy w such that $F_{ea}(\beta, w) = \{f\}$, and thus, $f \in \mathcal{L}_{U_S}(\beta)$. This establishes the convexity of $\mathcal{L}_{U_S}(\beta)$.
(ii) We first show that $\mathcal{L}(\beta) = \mathcal{L}_{U_S}(\beta)$. Choose some policy u and initial

distribution β, and some $f \in F_{ea}(\beta, u)$. By Theorem 11.1, there is some $\Delta \in [0,1]$, a stationary policy w such that $F_{ea}(\beta, w) = \{f^w\}$ for some f^w, and $f = \Delta f^w$. It follows from Lemma 17.2(ii) (in the appendix) that $f(\mathcal{K}) = 1$. This implies that $\Delta = 1$, so that $f^w = f$. Consequently, $\mathcal{L}(\beta) = \mathcal{L}_{U_S}(\beta)$.

Next, we show that $\mathcal{L}(\beta)$ is compact. Let $f_i \in \mathcal{L}(\beta), i \in \mathbb{N}$. Let f be some limit point of f_i in the sense of weak convergence of measures over $\mathcal{K}$ (its existence follows from Lemma 17.1 in the appendix). Our aim is to find a policy u such that $F_{ea}(\beta, u) = \{f\}$.

By Theorem 11.1, there exists a stationary policy g_i that achieves f_i, i.e., $F_{ea}(\beta, g_i) = \{f_i\}$. Let $\varepsilon_i := d(f, f_i)$, so that $\lim_{n\to\infty} \varepsilon_n = 0$, where d is a metric compatible with the weak convergence topology, i.e., f_i weakly converges to f if and only if $d(f, f_i)$ converges to 0 (see e.g., the Prohorov metric in Daley and Vere-Jones, 1988, p. 624). Consider the non-stationary policy u, that uses g_1 until the time $t_1 := \min\{t : d(f_1, f^t_{ea}(\beta, u)) \le \varepsilon_1\}$, and uses g_i between $t_{i-1}+1$ and t_i, where $t_i := \min\{t > t_{i-1} : d(f_i, f^t_{ea}(\beta, u)) \le \varepsilon_i\}$. The fact that $t_n < \infty$ is proved by contradiction as follows. Suppose the policy u uses g_n from time s onward forever. Then

$$\begin{aligned} & f^t_{ea}(\beta, u; y, \mathcal{A}) \\ = \quad & \frac{s}{t} f^s_{ea}(\beta, u; y, \mathcal{A}) + [g_n]_y(\mathcal{A}) \frac{t-s}{t} \sum_{z \in \mathbf{X}} P^u_\beta(X_s = z) \left(\sum_{r=1}^{t-s} [P^r(g_n)]_{zy} \right) \end{aligned}$$

(where $P(g_n)$ is the transition probabilities matrix under g_n). It then follows easily that $f^t_{ea}(\beta, u)$ weakly converges to f_n (as $t \to \infty$). Thus t_n is indeed finite (which contradicts the fact that g_n is used forever after s, which once more confirms that t_n is indeed finite). Now,

$$d(f, f^{t_n}_{ea}(\beta, u)) \le d(f, f_n) + d(f_n, f^{t_n}_{ea}(\beta, u)) \le 2\varepsilon_n.$$

Hence

$$\lim_{n\to\infty} f^{t_n}_{ea}(\beta, u) = f, \tag{11.15}$$

which establishes the compactness of $\mathcal{L}$.

Next we show that $\mathcal{L}_{U_S}(\beta)$ is equal to the closed convex hull of $\mathcal{L}_{U_D}(\beta)$. Since it is compact, by the Krein-Milman Theorem (Krein and Milman, 1940), it is the closed convex hull of its extreme points. Choose some extreme point ρ of $\mathcal{L}_{U_S}(\beta)$. Define $w(\rho)$ to be a stationary policy such that $w_y(\mathcal{A}) = \rho(y, \mathcal{A})[\rho(y, \mathsf{A}(y))]^{-1}$ whenever the denominator is non-zero. When it is zero, we let w_y be concentrated on $a(y)$, where $a(y)$ is some arbitrary action in $\mathsf{A}(y)$. We have $f_{ea}(\beta, w) = \rho$ (by the proof of Theorem 11.1). Assume that $w \notin U_D$. Then there exists some $y \in \mathbf{X}$ such that $\rho(y, \mathsf{A}(y)) > 0$. But then by Lemma 11.2, $f_{ea}(\beta, w)$ is not an extreme point of $\mathcal{L}_{U_S}(\beta)$, as it can be expressed as a convex combination of the distinct points $f_{tc}(\beta, w^a)$ (where w^a are given in Lemma 8.2).

Tightness of $\mathcal{L}_{U_S}$ follows from Lemma 17.2(ii) in the appendix. □

Remark 11.2 (*The importance of being tight*)
In order for the stationary policies to be complete, assumption (B2) is indeed necessary, as can be shown using the Fisher and Ross example (Example 8.4) which satisfies the unichain assumption. Indeed, Spieksma (1990, Theorem 11.11) showed that there exists a non-stationary policy u for which $f_{ea}(x,u;\mathcal{K}) = 0.7$. Hence, $f_{ea}(x,u;\cdot)$ is not a probability measure, and therefore Assumption (B2(u)) does not hold (see Lemma 17.1 in the appendix).

11.3 Relation between cost and occupation measure

We shall assume in the rest of this chapter that the immediate costs are either bounded below or that (B2) and the uniform integrability condition (B3), introduced below, hold. For the former case (cost bounded below), we shall use either the additional assumption (B2), or alternatively, assume some growth condition.

- **(B3(u))** Given a policy u, the expected occupation measures $\{f^t_{ea}(\beta,u)\}_t$ are integrable with respect to the absolute value of the immediate costs $|c|, |d^1|, \ldots, |d^K|$, uniformly in t.
- **(B3)** Assumption (B3(u)) holds for all $u \in U$.

(B3) implies the following:

- **(B3*)** The set of measures $\mathcal{L}(\beta)$ is integrable with respect to the absolute value of the immediate costs $|c|, |d^1|, \ldots, |d^K|$, uniformly in $u \in U_M$ (and hence over U and $\mathcal{U}$ and $\overline{M}(U_M)$).

The proof that (B3) implies (B3*) is postponed to Lemma 11.6.

Assumptions (B2) and (B3) will suffice to obtain a similar linear representation of the cost as was obtained for the total cost case (Section 8.5). When (B3) does not hold, we shall use assumption (11.1) to show that U_S and $\mathcal{U}$ have these properties. For other policies, that representation will not hold in general. To illustrate that, let $c(\kappa) = 1$ for all $\kappa \in \mathcal{K}$, and consider a policy u for which the occupation measures are not tight (e.g., the Fisher and Ross (1968) Counter-example, see Remark 11.2). Then we have

$$1 = C_{ea}(\beta,u) = C^t_{ea}(\beta,u) > \langle f, c\rangle, \qquad \forall t \in \mathbb{N}.$$

We have the following properties of the expected average costs (see Altman and Shwartz, 1991a, Lemma 2.3):

Theorem 11.4 (*Linear representation of the cost*)
(*i*) *Assume* (B2)-(B3). *Then for any* β, $u \in U \cup \overline{M}(U_M)$ *and* $f \in F_{ea}(\beta,u)$,

$$C_{ea}(\beta,u) \geq \langle f, c\rangle := \int_{\mathcal{K}} c(\kappa) f(d\kappa) \tag{11.16}$$

with equality holding for some $f \in F_{ea}(\beta, u)$; the expected average costs are uniformly bounded over all policies:

$$\sup_u |C_{ea}(\beta, u)| < \infty. \tag{11.17}$$

(ii) Assume that the cost is bounded below, i.e., (11.1) holds. Fix some β and $u \in U_S$ or $u \in \mathcal{U}$.
– If either (ii.1) The total expected cost to reach some recurrent state z is finite, or (ii.2) $C_{ea}(z, u) = \infty$, then (11.16) holds with equality.
– For any u, if c is non-negative or if $f \in M_1(\mathcal{K})$ for some $f \in F_{ea}(\beta, u)$, then (11.16) *holds.*

Proof. (i) Choose any $u \in U$ and let t_n be some sequence along which the weak limit $f = \lim_{n\to\infty} f^{t_n}_{ea}(\beta, u)$ exists. Then

$$\begin{aligned} C_{ea}(\beta, u) &= \overline{\lim_{t\to\infty}} C^t_{ea}(\beta, u) = \overline{\lim_{t\to\infty}} \int f^t_{ea}(\beta, u; d\kappa)c(\kappa) \\ &\geq \overline{\lim_{n\to\infty}} \int f^{t_n}_{ea}(\beta, u; d\kappa)c(\kappa). \end{aligned}$$

Due to the uniform integrability of $f^{t_n}_{ea}(\beta, u; d\kappa)$ w.r.t. the cost c, the integration and limit may be interchanged, see Lemma 17.4 in the appendix. This establishes (11.16). Equality is obtained in (11.16) by choosing t_n so as to achieve the limsup:

$$\overline{\lim_{t\to\infty}} \int f^t_{ea}(\beta, u; d\kappa)c(\kappa) = \lim_{n\to\infty} \int f^{t_n}_{ea}(\beta, u; d\kappa)c(\kappa)$$

and so that a weak limit $f =_w \lim_{n\to\infty} f^{t_n}_{ea}(\beta, u)$ exists. (Such a choice is possible due to Lemma 17.1(i) in the appendix.)

Finally, (11.17) follows from (B3) and from the fact that

$$\sup_u |C_{ea}(\beta, u)| \leq \sup_{f\in\mathcal{L}(\beta)} \langle f, |c|\rangle.$$

This establishes (i).

(ii) Consider $u \in U_S$. Assume (ii.1). Call a 'cycle' the period between two consecutive visits to some state z. If the total expected cost per cycle is finite, then the result follows from standard theory of Markov chains, see e.g., Chung (1967, p. 91-92). Note that this cost is always well defined since the immediate cost is bounded below. If $C_{ea}(z, u) = \infty$, then the expected cost per cycle is infinite, since the expected average cost equals the expected cost per cycle divided by the expected cycle duration (which is finite due to assumption (11.2)). In that case, one may replace the immediate cost c by the truncated cost $c^B(\kappa) = \min(c(\kappa), B)$. For every finite B, the corresponding total expected cost per cycle is finite, as well as the total expected cost until state z is first reached. Hence, by the first part

of the proof, (11.16) holds. The result is then obtained by the monotone convergence theorem.

To establish (11.16) for the case of non-negative cost and arbitrary u, we choose some $f \in F_{ea}(\beta, u)$, and a sequence t_n such that f is the vague limit $f =_v \lim_{n\to\infty} f_{ea}^{t_n}(\beta, u)$. Following Lemma 17.2(i) (in the appendix), we then have

$$\begin{aligned} C_{ea}(\beta, u) &= \overline{\lim_{t\to\infty}} \int f_{ea}^{t}(\beta, u; d\kappa)c(d\kappa) \geq \overline{\lim_{n\to\infty}} \int f_{ea}^{t_n}(\beta, u; d\kappa)c(d\kappa) \\ &\geq \int \underline{\lim_{n\to\infty}} f_{ea}^{t_n}(\beta, u; d\kappa)c(d\kappa) = \langle f, c\rangle. \end{aligned} \tag{11.18}$$

To establish (11.16) for the case of $f \in M_1(\mathcal{K})$, we define $c'(x, a) = c(x, a) - \underline{b}$, where $\underline{b}$ is as in (11.1). By (11.18) we obtain

$$C_{ea}(\beta, u) = \underline{b} + \overline{\lim_{t\to\infty}} \int f_{ea}^{t}(\beta, u; d\kappa)c'(d\kappa) \geq \underline{b} + \langle f, c'\rangle = \langle f, c\rangle.$$

Next, establish the first statement in (ii) for $\hat{\gamma} \in \mathcal{U}$. The expected average cost corresponding to the immediate cost c^B (as defined above) satisfies:

$$\begin{aligned} C_{ea}^{B}(\beta, \hat{\gamma}) &= \overline{\lim_{t\to\infty}} C_{ea}^{t,B}(\beta, \hat{\gamma}) = \overline{\lim_{t\to\infty}} \int_{\mathcal{U}} \gamma(du) C_{ea}^{t,B}(\beta, u) \\ &= \int_{\mathcal{U}} \gamma(du) C_{ea}^{B}(\beta, u), \end{aligned}$$

due to the bounded convergence theorem. For stationary u, we have

$$\lim_{B\to\infty} C_{ea}^{B}(\beta, u) = C_{ea}(\beta, u)$$

(due to the first part of (ii)). Hence, by the monotone convergence theorem, the right-hand side converges to $\int_{\mathcal{U}} \gamma(du) C_{ea}^{B}(\beta, u)$. Since for every positive B, $C_{ea}^{B}(\beta, \hat{\gamma}) \leq C_{ea}(\beta, \hat{\gamma})$, this implies that

$$C_{ea}(\beta, \hat{\gamma}) \leq \int_{\mathcal{U}} \gamma(du) C_{ea}(\beta, u).$$

Since $f_{ea}(\beta, \hat{\gamma}) = \int_{\mathcal{U}} \gamma(du) f_{ea}(\beta, u)$ (see (11.11), it follows by Fubbini's Theorem that

$$C_{ea}(\beta, \hat{\gamma}) \leq \langle f_{ea}(\beta, \hat{\gamma}), c\rangle.$$

The proof now follows from the last statement of part (ii) of the theorem (which we have already proved), since, by Lemma 11.1, $f_{ea}(\beta, \hat{\gamma})$ is a probability measure. □

As we did for the total cost, we may relax the assumptions in Theorem 11.4 (i) and combine them with those of part (ii). We obtain the following:

Theorem 11.5 (*Relaxing the assumptions in Theorem 11.4 (i)*)
(*i*) *Define* $c^+(x, a) = \max(c(x, a), 0)$ *and* $c^-(x, a) = \min(c(x, a), 0)$ *(so that*

$c = c^+ + c^-$). Assume that (B2) *holds, and that assumption* (B3) *applies for c^-. Then (11.16) holds for any u.*
(ii) If, moreover, $u \in U_S$ or $u \in \mathcal{U}$ and conditions (ii.1) or (ii.2) of Theorem 11.4 hold, then (11.16) *is obtained with equality.*

Remark 11.3 (11.11), together with Theorem 11.4 (i), implies that

$$C_{ea}(\beta, \hat{\gamma}) = \langle \gamma, C_{ea}(\beta, \cdot) \rangle, \quad \forall \hat{\gamma} \in \mathcal{U}$$

under assumptions (B2)-(B3), or under the assumptions (ii.1) or (ii.2) of Theorem 11.4 (ii), or under the assumptions of Theorem 11.5 (ii).

Next we describe continuity properties of the expected average cost.

Lemma 11.3 (*Continuity of the cost, uniform integrability assumptions*) *Assume* (B2)-(B3). *Then $C_{ea}(\beta, u)$ is continuous on U_S and on $\mathcal{U}$.*

Proof. This follows from the continuity of the occupation measures over U_S and $\mathcal{U}$ (Theorem 11.2), from the fact that (11.16) holds with equality for $u \in U_S$ and $u \in \mathcal{U}$ (since $F_{ea}(\beta, u)$ is then a singleton), and from (B3*). Indeed, let $\overline{U}$ stand for either U_S or $\mathcal{U}$. Let u_n be a sequence of policies in $\overline{U}$ converging to a policy $u \in \overline{U}$. Then

$$\lim_{n\to\infty} C_{ea}(\beta, u_n) = \lim_{n\to\infty} \langle f^n_{ea}(\beta, u_n), c) \rangle = \langle f_{ea}(\beta, u), c) \rangle = C_{ea}(\beta, u). \tag{11.19}$$

(The integration and limit may be interchanged due to (B3*), see Lemma 17.4(i) in the appendix. □

Next, we obtain continuity properties for the case of cost bounded below. We shall assume in addition either (B2) or another growth condition, due to Borkar (1983), and adapted to constrained MDPs in Altman and Shwartz (1991a):
There exists a sequence of increasing compact subsets $\mathcal{K}_i$ of $\mathcal{K}$ such that $\cup_i \mathcal{K}_i = \mathcal{K}$ and such that the immediate cost functions c satisfies

$$\underline{\lim}_{i\to\infty} \{c(\kappa); \kappa \notin \mathcal{K}_i\} = \infty. \tag{11.20}$$

Note that the so-called 'moment condition' (11.20) implies that c is bounded below by some $\underline{b}$. (Note that c and d achieve their minima over each compact set $\mathcal{K}_i$, since they are continuous on $\mathcal{K}$, by (6.1)).

A sufficient condition for (11.20), which is frequently used in the literature (e.g., Cavazos-Cadena, 1989, Cavazos-Cadena and Sennott, 1992), is the following moment condition:

$$\forall \ell \in \mathbb{R}, \text{ the set } \{x \in \mathbf{X} : \inf_a c(x, a) < \ell\} \text{ is finite .} \tag{11.21}$$

Lemma 11.4 (*Lower semi-continuity of the cost*)
Assume that either
(i) (11.20) *holds for the immediate cost c, or*
(ii) c is bounded below and that Assumption (B2*) *holds.*
Then $C_{ea}(\beta, u)$ is l.s.c. over U_S and $\mathcal{U}$.

Proof. Let u^n be a sequence of stationary policies converging to u (i.e., $u^n(\cdot \mid x)$ converges weakly to $u(\cdot \mid x)$ for all $x \in \mathbf{X}$).

Assume that $f_{ea}(\beta, u^n)$ are tight.
Theorem 11.2 (i) implies that $f_{ea}(\beta, u^n)$ converges weakly to $f_{ea}(\beta, u)$. We have

$$\underline{\lim}_{n\to\infty} C_{ea}(\beta, u^n) = \underline{\lim}_{n\to\infty} \langle f_{ea}(\beta, u^n), c\rangle \geq \langle f_{ea}(\beta, u), c\rangle,$$

where the first equality follows from Theorem 11.4, and the inequality follows from Doob (1994, p. 133) (in the related theorem in Doob, the cost is assumed to be bounded; However, it can easily be seen that only boundedness from below is used in the proof of that theorem). Thus, (ii) is established, and (i) is established for the case that $f_{ea}(\beta, u^n)$ are tight.

Assume now that (11.20) holds. Let $n(i)$ be some sequence of integers along which $\underline{\lim}_n C_{ea}(\beta, u^n)$ is obtained as a limit. To establish lower semi-continuity, it clearly suffices to consider the case that $\lim_i C_{ea}(\beta, u^{n(i)})$ is finite. But in that case, it follows that $f_{ea}(\beta, u^{n(i)})$ are tight, so the result is obtained from the first part of the proof.

Indeed, assume that $\{f_{ea}(\beta, u^{n(i)})\}_i$ are not tight. Let $\mathcal{K}_n$ be a sequence of compact subsets of $\mathcal{K}$, increasing to $\mathcal{K}$. Since $f_{ea}(\beta, u^n)$ are not tight, there exists some δ and a subsequence $n(i(j))$ of $n(i)$ such that

$$f_{ea}(\beta, u^{n(i(j))}; \mathcal{K}_j^c) > \delta$$

for all j. Hence

$$C_{ea}(\beta, u^{n(i(j))}) = \langle f_{ea}(\beta, u^{n(i(j))}), c\rangle \geq \delta \inf_{\kappa \notin \mathcal{K}_j} c(\kappa),$$

which tends to infinity as $j \to \infty$ (the first equality follows from Theorem 11.4). This contradicts the fact that t $\lim_i C_{ea}(\beta u^{n(i)})$ is finite, and establishes the proof. □

Remark 11.4 (*Other conditions for lower semi-continuity*)
The results of Lemma 11.4 can be obtained by using a combination of conditions of Lemmas 11.3 and 11.4, as in Theorem 11.5 (ii).

11.4 Dominating classes of policies

Theorem 11.6 (*Dominance under tightness and uniform integrability*)
Assume (B2)-(B3). *Then, any complete class of policies is a dominating class. If* **COP** *is feasible, then there exist optimal policies in* U_S *and in* $\mathcal{U}$.

Proof. The proof is the same as the one for the total cost, i.e., the proof of Theorem 8.4. (The basic steps can be found in Altman and Shwartz, 1991a, Theorem 2.8 and Corollary 5.4.) □

Next, we relate weakly complete classes of policies to dominating policies. In general, the fact that a class of policy is dominant does not ensure

the fact that an optimal policy exists; it only implies that we may restrict our search for such a policy to that dominating class. The existence of optimal stationary policies for **COP** was established by Altman and Shwartz (1991a) Corollary 5.4, and in Altman (1994) Theorem 4.2 (under suitable conditions). This implies the existence of optimal policies within any dominating class of policies.

Theorem 11.7 *(Dominant policies: a growth condition on the cost) Assume that the growth condition (11.20) holds for the immediate cost c or for d^k, for some $k = 1, \ldots, K$.*
(i) Let $\overline{U}$ be a class of policies that is weakly complete and for which (11.16) holds with equality for the immediate costs c as well as for d^k $(k = 1, \ldots, K)$ for all $u' \in \overline{U}$. Then $\overline{U}$ is a dominating class. In particular, $\overline{U}$ can be taken to be U_S.
(ii) $\mathcal{U}$ is a dominating class of policies.
(iii) If **COP** *is feasible, then there exist optimal policies for* **COP** *within $\overline{U}$, and in particular, within U_S and $\mathcal{U}$.*

Proof. (i) The proof is related to Altman and Shwartz (1991a, p. 800, Theorem 4.4). It clearly suffices to show that for any u for which $C_{ea}(\beta, u) < \infty$ and $D^k_{ea}(\beta, u) < \infty,\ k = 1, \ldots, K$, there exists some $u' \in \overline{U}$ such that

$$C_{ea}(\beta, u') \le C_{ea}(\beta, u) \qquad D_{ea}(\beta, u') \le D_{ea}(\beta, u).$$

Thus, assume without loss of generality that (11.20) holds for c. Choose some policy u, and $f \in F_{ea}(\beta, u)$. Assume that $C_{ea}(\beta, u) < \infty$ and that $\{f^t_{ea}(\beta, u)\}_t$ are not tight. Then there exists some $\varepsilon > 0$ and an increasing sequence $\{t_l\}$ such that $f^{t_l}_{ea}(\mathcal{K}^c_l) > \varepsilon$. Denote $c_j := \inf\{c(\kappa) : \kappa \notin \mathcal{K}_j\}$. It follows that

$$C^{t_j}_{ea}(\beta, u) \ge c_j \varepsilon + \max(\underline{b}, 0), \quad j \in \mathbb{N}.$$

Since by (11.20), $\lim_{j\to\infty} c_j = \infty$, it follows that $C_{ea}(\beta, u) = \infty$, which contradicts our assumption. Hence, if $C_{ea}(\beta, u) < \infty$, then $\{f^t_{ea}(\beta, u)\}_t$ are tight and f is a probability measure (this follows from Lemma 17.1 in the appendix). If $\overline{U}$ is a weakly complete class of policies, then there exists some $u' \in \overline{U}$ such that $f_{ea}(\beta, u') = f$. The last part of Theorem 11.4 (ii) and the assumption that (11.16) holds with equality for u' implies that u' dominates u. The statement for U_S follows since by Theorem 11.1 it is weakly complete and since by Theorem 11.4 (ii) it satisfies (11.16) with equality.
(ii) We postpone the proof of $\mathcal{U}$ to the next chapter (Corollary 12.1).
(iii) Assume that **COP** is feasible, and let g_i be a sequence of stationary policies which are ε_i optimal, where $\lim_{i\to\infty} \varepsilon_i = 0$. Assume moreover that $f_{ea}(\beta, g_i)$ converges to some limit f (in the sense of weak convergence of measures over $\mathcal{K}$). We may repeat now the argument in the part of the proof of compactness in Theorem 11.3 above (11.15); we may choose an increasing sequence t_n and construct a Markov policy u that uses policy g_i

during time $[t_i, t_{i+1})$, such that

$$\lim_{n\to\infty} f_{ea}^{t_n}(\beta, u) =_v f$$

(f is not necessarily a probability measure) and moreover,

$$\overline{C}_{ea}(\beta, u) := \lim_{n\to\infty} C_{ea}^{t_n}(\beta, u) = \lim_{n\to\infty} C_{ea}(\beta, g_n) = C_{ea}(\beta),$$

$$\overline{D}_{ea}^{k}(\beta, u) := \lim_{n\to\infty} D_{ea}^{k,t_n}(\beta, u) = \lim_{n\to\infty} D_{ea}^{k}(\beta, g_n) \le V_k \ k = 1, \ldots, K.$$

But then, by the first part of the theorem, there exists an optimal policy in $\overline{U}$, and in particular, among U_S. The statement for $\mathcal{U}$ follows from Theorem 11.7 (ii). □

Theorem 11.8 *(Dominant policies: tightness and cost bounded below)*
Assume (B2) and that either
(i) the immediate cost c or for d^k, $k = 1, \ldots, K$ are bounded below, or
(ii) (B3) holds for the negative *part c^- and $d^{k,-}$ of the immediate costs c and d^k, $k = 1, \ldots, K$.*
Then all claims of Theorem 11.7 hold.

Proof. The proof of (i) is almost the same as that of Theorem 11.7: the only place we used condition (11.20) there was to exclude policies for which (B2(u)) does not hold. Now these policies are excluded due to assumption (B2). The proof of (ii) is obtained similarly. □

We show below that Theorem 11.7 holds when the growth condition (11.20) is replaced by a weaker condition, due to Borkar (1983), which was applied to constrained MDPs in Altman and Shwartz (1991a, Section 4).

Definition 11.3 *(Almost monotone costs)*
$c : \mathcal{K} \to \mathbb{R}$ is called V-almost monotone if there exists a sequence of increasing compact subsets $\mathcal{K}_i$ of $\mathcal{K}$ such that $\cup_i \mathcal{K}_i = \mathcal{K}$ and

$$\underline{\lim}_{i\to\infty} \{c(\kappa); \kappa \notin \mathcal{K}_i\} > V. \tag{11.22}$$

Theorem 11.9 *(Weak completeness and dominance)*
Assume that $\overline{U}$ is weakly complete and that (11.16) holds with equality for both immediate costs c as well as d^k for all $u' \in \overline{U}$. Assume that there exists some feasible policy $u' \in \overline{U}$ i.e., $D_{ea}^k(\beta, u') \le V_k$, and define $V_0 := C_{ea}(\beta, u')$. If c is V_0-almost monotone and d^k are V_k-almost monotone for all $k = 1, \ldots, K$, then
(i) $\overline{U}$ is a dominating class of policies. Moreover, $\overline{U}$ can be taken as U_S or $\mathcal{U}$.
(ii) If **COP** *is feasible, then there exist optimal policies for* **COP** *within $\overline{U}$, and in particular, within U_S and $\mathcal{U}$.*

Proof. We do not present the detailed proof. The proof of (i) follows from ideas similar to those in Theorem 11.7. The exact proof of the dominance

of $\overline{U}$ can be found in Altman and Shwartz (1991a, Lemma 4.6). The existence of optimal policies within U_S was established in Altman (1994, Theorem 4.2). This, together with the dominance of $\mathcal{U}$, implies the existence of an optimal policy within $\mathcal{U}$. □

11.5 Equivalent Linear Program

We now obtain an LP formulation similar to the one we obtained for the total cost; we show again that the **COP** is equivalent to an LP with an infinite set of decision variables and a countable set of constraints. Consider the following LP:

LP$_3$(β): Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle \rho, c \rangle$ subject to:

$$\mathcal{D}^k(\rho) := \langle \rho, d^k \rangle \le V_k, k = 1, \ldots, K, \qquad \rho \in \mathbf{Q}_{ea}(\beta),$$

where $\mathbf{Q}_{ea}(\beta)$ was defined in (11.5).

Define $w(\rho)$ to be any stationary policy such that

$$w_y(\mathcal{A}) = \rho(y, \mathcal{A})[\rho(y, \mathsf{A}(y))]^{-1}$$

whenever the denominator is non-zero. We show that there is a one-to-one correspondence between feasible (and optimal) solutions of the LP, and the feasible (and optimal) solutions of **COP**.

Theorem 11.10 *(Equivalence between COP and* **LP$_3$**(β)*)*
Assume that one of the following three conditions holds:

- *The immediate cost is bounded below (11.1) and satisfies the growth condition (11.20) or (11.22); moreover, for any stationary policy u, the total expected cost to reach some recurrent state z is either finite, or $C_{ea}(z, u) = \infty$.*
- *(B2) holds and the costs are bounded below; moreover, for any stationary policy u, the total expected cost to reach some recurrent state z is either finite, or $C_{ea}(z, u) = \infty$.*
- *(B2) holds, and (B3) holds for the the* negative *part c^- and $d^{k,-}$ of immediate costs c and d^k, $k = 1, \ldots, K$.*

Then
(i) $\mathcal{C}^* = C_{ea}(\beta)$.
(ii) For any $u' \in U$, there exists a dominating stationary policy $u \in U_S$ such that $\rho(u) := f_{ea}(\beta, u) \in \mathbf{Q}_{ea}(\beta)$, $C_{ea}(\beta, u) = \mathcal{C}(\rho(u))$ and $D_{ea}(\beta, u) = \mathcal{D}(\rho(u))$; conversely, for any $\rho \in \mathbf{Q}_{ea}(\beta)$, the stationary policy $w(\rho)$ satisfies $C_{ea}(\beta, w(\rho)) = \mathcal{C}(\rho)$ and $D_{ea}(\beta, w(\rho)) = \mathcal{D}(\rho)$.
(iii) **LP$_3$**(β) *is feasible if and only if* **COP** *is feasible. Assume that* **COP** *is feasible. Then there exists an optimal solution ρ^* for* **LP$_3$**(β)*, and the stationary policy $w(\rho^*)$ is optimal for* **COP**.

Proof. We start from (ii). The first claim follows from the fact that it holds

for stationary policies (as is shown in the first paragraph of the proof of Theorem 11.3), by combining Theorem 11.3 with Theorems 11.4 (ii), 11.6, 11.7, 11.8 and 11.9. The claims on the costs follow from Theorem 11.4. The converse part follows by noting that for any $\rho \in \mathbf{Q}_{ea}(\beta)$, $\rho = f_{ea}(\beta, w(\rho))$ (this follows from the first paragraph of the proof of Theorem 11.3), and by applying again Theorem 11.4. This establishes (ii), and thus implies (i). (iii) now follows from (ii) and Theorems 11.6, 11.7, 11.8 and 11.9. □

11.6 The Dual Program

Next, we present the formal dual program DP for the LP above. The decision variables are $\psi \in \mathbb{R}$, $\phi : \mathbf{X} \to \mathbb{R}$ and the K-dimensional non-negative vectors $\lambda \in \mathbb{R}_+^K$.

$\mathbf{DP_3}(\beta)$: Find $\Theta^*(\beta) := \sup_{\psi,\phi,\lambda} \psi - \langle \lambda, V \rangle$ subject to

$$\phi(x) + \psi \quad \leq \quad \left[c(x,a) + \langle \lambda, d(x,a) \rangle + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y) \right], \quad x \in \mathbf{X}, a \in \mathsf{A}(x).$$

We shall show in the next chapter that when choosing the decision variables ϕ to be in the appropriate linear space, then there is no duality gap, and

$$\Theta^* = \mathcal{C}^* = C_{ea}(\beta), \tag{11.23}$$

for both the case that (B2)-(B3) hold, as well as the framework of costs bounded below.

11.7 The contracting framework

We present below the contracting framework, which provides simple sufficient conditions for assumptions (B2)-(B3).

The cost is assumed to be μ-bounded (7.36), the transition probabilities are μ-continuous (Assumption (7.35)), and the initial distribution satisfies $\langle \beta, \mu \rangle < \infty$.

In the context of expected average cost, we shall define the 'contracting framework' to be the μ-uniformly geometrically recurrent MDP, together with the above assumptions on the transition probabilities, immediate costs and initial distribution.

Definition 11.4 *(μ-uniformly geometric recurrence)*
For a given vector $\mu : \mathbf{X} \to [1, \infty)$, an MDP is μ-uniformly geometrically recurrent if it satisfies Definition 7.9 (contracting MDP) with the set $\mathcal{M}$ being finite.

Definition 11.5 *(μ-uniformly geometric ergodicity)*
The MDP is said to be μ-uniformly geometrically ergodic if there exist

constants $\sigma > 0$ and $\tilde{\xi} < 1$ such that for all $u \in U_S$,

$$\begin{cases} \|P^n(u) - \Pi(u)\|_\mu \leq \sigma \tilde{\xi}^n, & \forall n \in \mathbb{N}, \\ \|P(u)\|_\mu \leq \sigma \end{cases} \tag{11.24}$$

where $P^n(u)$ is the matrix of n-step transition probabilities under policy u, and $\Pi(u)$ is the matrix whose rows are equal to the steady-state probabilities under u.

We present below a part of the remarkable equivalence relation established in Spieksma (1990, Key Theorem II, p. 108, and Lemma 5.3 (ii)) and Dekker *et al.* (1994).

Theorem 11.11 (*μ-uniformly geometric ergodicity and recurrence*)
Assume that (11.2) *holds. Then,*
(i) if the MDP is μ-uniform geometrically recurrent, then it is μ-uniformly geometrically ergodic.
(ii) if the MDP is μ-uniform ergodic, then it is $\tilde{\mu}$-uniformly geometrically recurrent with

$$\tilde{\mu} = \sup_{w \in U_S} \sum_{n=0}^{\infty} {}_{\mathcal{M}}P^n(w)\mu.$$

Note that $\nu \leq \tilde{\mu}$; on the other hand, if the MDP is $\tilde{\mu}$-uniformly geometrically recurrent, then $\tilde{\mu}$ is μ-bounded. Indeed,

$$\tilde{\mu} \leq \mu + \sup_{w \in U_S} {}_{\mathcal{M}}P(w)\tilde{\mu} \leq \mu + \xi\tilde{\mu}$$

so that $\tilde{\mu} \leq \mu/(1-\xi)$.

Next we present uniform tightness and integrability properties of contracting MDPs.

Lemma 11.5 (*Tightness and uniform integrability*)
Under the contracting framework, the sets $\{f^t_{ea}(\beta, u)\}_{t \in \mathbb{N}, u \in \mathrm{U}}$ are tight and are, moreover, uniformly integrable with respect to the cost c and to $\mu(y, a) := \mu(y)$. Hence (B2) *and* (B3) *hold.*

Proof. The uniform integrability follows directly from Corollary 6.2 in Spieksma (1990), who restricted herself to Markov policies, to fixed initial states, and to uniform integrability with respect to μ. The generalization to any policy follows from Theorem 6.1. The proof in Spieksma (1990) extends in a straightforward way to any initial distribution (satisfying of course $\langle \beta, \mu \rangle < \infty$). Tightness follows from the uniform integrability (see Lemma 6.5 in Spieksma (1990)). □

Finally, we present a result by Spieksma (1990, Proposition 5.1, p. 97) that establishes a stronger version of the continuity of the occupation measure over the set of stationary policies.

Theorem 11.12 *(Continuity of occupation measures)*
Consider the contracting framework. Then the state occupation measures $f_{ea}(\beta, \bullet; \bullet)$ are μ-continuous over U_S.

11.8 Other conditions for the uniform integrability

We present in this section other conditions for the uniform integrability conditions (B3) and (B3*).

Lemma 11.6 *Assume* (B2). *Then assumption* (B3) *implies* (B3*).

Proof. Assume that (B3) holds and (B3*) does not hold. Then there is a sequence $f_i \in \mathcal{L}(\beta), i \in \mathbb{N}$ converging to some f that are not uniformly integrable with respect to, say, $|c|$. Let g be a stationary policy that achieves f, i.e., $F_{ea}(\beta, g) = \{f\}$. Its existence follows from Theorem 11.3 (ii). We have $\lim_{i\to\infty} \langle f_i, |c|\rangle < \langle f, |c|\rangle$, see Lemma 17.4(ii) in the appendix. But this contradicts the linear representation of the cost (Theorem 11.4) for the policy g, or the continuity of the cost in the policies (see (11.19)). □

Lemma 11.7 (*Linear representation of the cost implies* B3(u))
Let $\overline{C}_{ea}(\beta, u)$ be the expected average cost corresponding to the immediate cost $|c|$. Assume that for some u and $f \in F_{ea}(\beta, u)$,

$$\overline{C}_{ea}(\beta, u) = \langle f, |c|\rangle. \tag{11.25}$$

Then $\{f^t_{ea}(\beta, u)\}$ are uniformly integrable with respect to $|c|$.

Proof. Assume that B3(u) does not hold. Let $t(n)$ be a subsequence along which $\lim_{n\to\infty} f^{t(n)}_{ea}(\beta, u) = f$. Then

$$\overline{C}_{ea}(\beta, u) = \varlimsup_{t\to\infty} \langle f^t_{ea}(\beta, u), |c|\rangle \geq \varlimsup_{n\to\infty} \langle f^{t(n)}_{ea}(\beta, u), |c|\rangle > \langle f, |c|\rangle$$

according to Lemma 17.4(ii) in the appendix, which contradicts the linear representation of the cost (11.25). □

Lemma 11.8 *Assume*

- (a1) *$f^t_{ea}(\beta, u)$ are integrable with respect to $|c|, |d^1|, \ldots, |d^K|$, for all t and u;*
- (a2) *This integrability is uniformly in $\{t, u \in U_S\}$;*
- (a3) *The stationary policies are dominating in the following sense. For any non-negative immediate cost r,*

$$\sup_{u\in U} R(\beta, u) = \sup_{u\in U_S} R(\beta, u),$$

where $R(\beta, u)$ is the expected average cost corresponding to the immediate cost r.

Then (B3) *holds.*

Proof. Assume that B3(u) does not hold for some Markov policy u, with respect to, say, $|c|$. Let $\mathcal{K}_n \subset \mathcal{K}$ be a sequence of sets converging to $\mathcal{K}$. There is some $\varepsilon > 0$ and a strictly increasing sequence $t(n)$ such that

$$\int 1\{\kappa \notin \mathcal{K}_n\}|c(\kappa)| f^{t(n)}_{ea}(\beta, u; d\kappa) > \varepsilon.$$

(That $t(n)$ can be chosen to be strictly increasing follows from the integrability of f^t.) This implies, in particular, that for every n,

$$\overline{\lim_{t\to\infty}} \int 1\{\kappa \notin \mathcal{K}_n\}|c(\kappa)| f^t_{ea}(\beta, u; d\kappa) > \varepsilon. \tag{11.26}$$

(a2) implies that there exists some n such that for the immediate cost $r(x,a) = |c(x,a)|1\{x \notin \mathcal{K}_n\}$, we have

$$\sup_{u\in U_S} R(\beta, u) < \varepsilon/2. \tag{11.27}$$

However, (11.26) and (11.27) are not compatible with Assumption (a3). This establishes the lemma by contradiction. □

11.9 The case of uniform Lyapunov conditions

We establish in this section the equivalence between conditions (B2)-(B3) and MDPs with uniform Lyapunov functions.

We first show that MDPs with uniform Lyapunov functions (as defined in Definition 7.5) for which $\mathcal{M}$ is a finite set, satisfy conditions (B2)-(B3).

The proof of the following theorem uses ideas from Spieksma (1990) Lemma 2.3.

Theorem 11.13 *(Tightness and uniform integrability under a uniform Lyapunov function)*
Assume that the MDP has a uniform Lyapunov function μ, that the set $\mathcal{M}$ is finite, and that $\langle \beta, \mu\rangle < \infty$. Then the sets $\{f^t_{ea}(\beta, u)\}_{t,u}$ are
(i) Integrable with respect to the immediate costs $c, d^1, \ldots, d^K$, uniformly in t and u. Hence (B3) holds.
(ii) Are tight, i.e., (B2) holds.

Proof. Without loss of generality, we restrict ourselves to Markov policies. Denote $c'(x,a) := 1 + \nu(x,a)$, where ν (given in Definition 7.5) is a bound on the immediate costs. Both (i) and (ii) follow by showing that $f^t_{ea}(\beta, u)$ are integrable w.r.t. c', uniformly in $u \in U_M$ and $t \in \mathbb{N}$.

We use the last exit time decomposition. Define

$$T^{[s]} \stackrel{\text{def}}{=} \min\{n > s : X_n \in \mathcal{M}\}, \tag{11.28}$$

$$E^u_\beta c'(X_t, A_t) = E^u_\beta c'(X_t, A_t)1\{T > t\} \tag{11.29}$$
$$+ \sum_{z\in\mathcal{M}} \sum_{s=1}^{t} E^u_\beta [c'(X_t, A_t)1\{\mathcal{T}^{[s]} > t\}1\{X_s = z\}].$$

Taking the time average until t of the first term on the right-hand side of (11.29), we get

$$\langle f^t_{ea}(\beta, u), c'\rangle \le \frac{1}{t}\langle f_{tc}(\beta, u), c'\rangle.$$

Next we consider the time average of the second term in (11.29). For any Markov policy u, define $\theta^s u$ to be that policy, shifted by s steps, i.e., $[\theta^s u]_t(\cdot \mid x) = u_{t+s}(\cdot \mid x)$.

$$\begin{aligned}
&\frac{1}{r}\sum_{t=1}^{r}\sum_{s=1}^{t} E_\beta^u[c'(X_t,A_t)1\{T^{[s]}>t\}1\{X_s=z\}] \\
&= \frac{1}{r}\sum_{s=1}^{r}\sum_{t=s}^{r} E_\beta^u[c'(X_t,A_t)1\{T^{[s]}>t\}1\{X_s=z\}] \\
&\le \frac{1}{r}\sum_{s=1}^{r}\sum_{t=s}^{\infty} E_\beta^u[c'(X_t,A_t)1\{T^{[s]}>t\}1\{X_s=z\}] \\
&\le \frac{1}{r}\sum_{s=1}^{r}\langle f_{tc}(z,\theta^s u),c'\rangle.
\end{aligned}$$

We conclude that

$$\langle f_{ea}^t(\beta,u),c'\rangle \le \frac{1}{t}\langle f_{tc}(\beta,u),c'\rangle + \sum_{z\in\mathcal{M}}\frac{1}{r}\sum_{s=1}^{r}\langle f_{tc}(z,\theta^s u),c'\rangle. \qquad (11.30)$$

$f_{tc}(\beta,u)$ are integrable w.r.t. c', uniformly in $u \in U_M$; this follows from Lemma 17.4 in the appendix, since $\hat{M}(\beta,u) = \langle f_{tc}(\beta,u),c'\rangle$ are continuous in u (relation M1 $\Leftrightarrow$M5 in Theorem 7.3). Thus, the first term on the right-hand side of (11.30) is uniformly integrable w.r.t. c'. Similarly, $f_{tc}(z,\theta^s u)$ are integrable w.r.t. c' uniformly over U_M and $s \in \mathbb{N}$. It then follows that $r^{-1}\sum_{s=1}^{r}\langle f_{tc}(z,\theta^s u),c'\rangle$ are also integrable w.r.t. c' uniformly over U_M and $s \in \mathbb{N}$. Since all terms on the right-hand side of (11.30) are integrable w.r.t. c' uniformly over U_M and $s \in \mathbb{N}$, then so is their sum, and this implies (see Remark 17.1) that $\langle f_{ea}^t(\beta,u),c'\rangle$ are integrable w.r.t. c' uniformly over U_M and t. □

Next we establish the converse:

Theorem 11.14 *(Relation between (B2), (B3) and uniform Lyapunov functions)*
Assume that conditions (B2) and (B3) hold. Then the MDP has a uniform Lyapunov function where $\mathcal{M} = \{0\}$, where 0 is any arbitrary state and where $\nu(x,a) \stackrel{\text{def}}{=} |c(x,a)| + \sum_{i=1}^{K}|d^k(x,a)|$.

Proof. Choose some arbitrary state 0. It follows from Assumption (B2) that the steady-state $\pi_x(u), x \in \mathbf{X}$ probabilities are continuous over U_S (Theorem 11.2). Since the total expected time $M(0,u) \stackrel{\text{def}}{=} E_0^u T_{\{0\}}$ between two visits of state 0 is given by $M(0,u) = [\pi_x(u)]^{-1}$, and since $\pi_x(u) > 0$ for all x and $u \in U_S$ due to the unichain assumption, it follows that $M(0,u)$ is continuous over U_S.

Consider the expected average cost and total expected cost corresponding

to the immediate costs $|c|$ instead of c, and denote these by $\overline{C}_{ea}(\beta, u)$ and $\overline{C}_{tc}(\beta, u)$.

Assumptions (B2) and (B3) imply that $\overline{C}_{ea}(\beta, u)$ is continuous over U_S (Theorem 11.3).

For any $u \in U_S$, we have

$$\overline{C}_{ea}(0, u) = \frac{\overline{C}_{tc}(0, u)}{E_0^u T_{\{0\}}}.$$

From the continuity of $E_0^u T_{\{0\}}$ and $\overline{C}_{ea}(0, u)$ it then follows that $\overline{C}_{tc}(0, u)$ is continuous over U_S.

Defining $\overline{D}_{tc}(0, u)$, we obtain the continuity of $\overline{D}_{tc}(0, u)$ over U_S. For any sequence $u(n) \in U_S$ converging to some $u \in U_S$,

$$\begin{aligned}
\lim_{n\to\infty} \hat{M}(0, u(n)) &= \lim_{n\to\infty} E_0^{u(n)} \sum_{n=1}^{\infty} [1 + \nu(X_n, A_n)] 1\{T > n\} \\
&= \lim_{n\to\infty} E_0^{u(n)} \sum_{n=1}^{\infty} [1 + |c(X_n, A_n)| + \sum_{i=1}^{K} |d^k(X_n, A_n)|] 1\{T > n\} \\
&= \lim_{n\to\infty} [M(0, u(n)) + \overline{C}_{tc}(0, u(n)) + \sum_{i=1}^{K} \overline{D}_{tc}^K(0, u(n))] \\
&= M(0, u) + \overline{C}_{tc}(0, u) + \sum_{i=1}^{K} \overline{D}_{tc}^K(0, u) \\
&= \hat{M}(0, u).
\end{aligned}$$

(Changing the order of summation is possible as the summands are all non-negative.) We conclude that Assumption N3 (in Section 7.5) holds. Due to the unichain assumption, this implies property M1, i.e., the existence of a uniform Lyapunov function, see Corollary 7.1. □

CHAPTER 12

Expected average cost: Dynamic Programming and LP

We present in this chapter dynamic programming, similar to that in Chapter 9, for the unconstrained control problem, and then, using Lagrangian and duality methods, derive the linear program $\mathbf{DP_3}(\beta)$, which is the dual of that obtained in the previous chapter. We show again that there is no duality gap both for the case of the uniform Lyapunov function, as well as the case of costs bounded below. As in Chapter 9, we conclude by presenting a different LP approach for computing the optimal values and optimal mixed strategies.

The uniform Lyapunov function that we consider will be with respect to a set $\mathcal{M} = \{0\}$ where 0 is some arbitrary state. We recall from Theorem 11.13 that the uniform Lyapunov conditions imply the conditions (B2) and (B3) of Chapter 11.

12.1 The non-constrained case: optimality inequality

Introduce the (expected) Average Cost Optimality Inequality (ACOI):

$$\textbf{ACOI}: \qquad \phi(x) + \psi \geq \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y) \right], \qquad (12.1)$$

where ψ is some constant, and $\phi : \mathbf{X} \to \mathbb{R}$. This type of equation is closely related to the optimal value and the computation of optimal policies, as will be established in details in the following two sections. Before going into details, we motivate the above optimality inequality in the following lemmas that hold under general cost and ergodic structure. They provide in particular lower and upper bounds for the expected average cost. The ideas below can be found in Yushkevich (1973), Dynkin and Yushkevich (1979), Hernández-Lerma and Lasserre (1995) and Arapostathis *et al.* (1993) and references therein.

Lemma 12.1 (*Upper bound on the value*)
Let (ψ, ϕ) *be a solution of* (12.1) *and let* w *be a stationary policy that chooses at state* x *an action that achieves the inf of* $[c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y)]$ *up*

to $\varepsilon \geq 0$, i.e.,

$$c(x,w) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xwy}\phi(y) \leq c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay}\phi(y) + \varepsilon, \quad \forall a \in \mathsf{A}(x).$$

Assume that ϕ satisfies

$$\underline{\lim}_{n\to\infty} \frac{E_x^w \phi(X_n)}{n} \geq 0. \tag{12.2}$$

Then $\psi \geq C_{ea}(x,w) - \varepsilon$, and hence $\psi \geq C_{ea}(x) - \varepsilon$.

Proof. We iterate (12.1) and obtain:

$$\begin{aligned}
\phi(x) &\geq -\psi - \varepsilon + c(x,w) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xwy}\phi(y) \\
&= -\psi - \varepsilon + c(x,w) + E_x^w \phi(X_2) \\
&\geq -2\psi - 2\varepsilon + c(x,w) + E_x^w \left[c(X_2,A_2) + E_{X_2}^w \phi(X_3)\right] \\
&= -2\psi - 2\varepsilon + c(x,w) + E_x^w c(X_2,A_2) + E_x^w \phi(X_3) \\
&\geq \ldots \geq -n(\psi+\varepsilon) + \sum_{t=1}^{n} E_x^w c(X_t,A_t) + E_x^w \phi(X_{n+1}). \qquad (12.3)
\end{aligned}$$

Dividing by n in (12.3) and going to the limit as n tends to infinity, we conclude that $\psi \geq C_{ea}(x,w) - \varepsilon \geq C_{ea}(x) - \varepsilon$. □

Remark 12.1 Clearly, a sufficient condition for (12.2) to hold is that ϕ is bounded from below.

Definition 12.1 *(Superharmonic pair)*
A pair (ψ, ϕ) (where ψ is a constant and $\phi : \mathbf{X} \to \mathbb{R}$) is called superharmonic (for the expected average cost criterion) if it satisfies for all $x \in \mathbf{X}$ and $a \in \mathsf{A}(x)$:

$$\phi(x) + \psi \leq c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay}\phi(y). \tag{12.4}$$

Lemma 12.2 (*Lower bound on the value*)
Assume that there exists some superharmonic pair (ψ, ϕ) such that

$$\overline{\lim}_{n\to\infty} \frac{E_x^u \phi(X_n)}{n} \leq 0 \tag{12.5}$$

for some Markov policy u. Then $\psi \leq C_{ea}(x,u)$.

Proof. Iterating (12.5), we get

$$\begin{aligned}
\phi(x) &\leq c(x,u_1) - \psi + \sum_{y \in \mathbf{X}} \mathcal{P}_{xu_1y}\phi(y) = c(x,u_1) - \psi + E_x^u \phi(X_2) \\
&\leq c(x,u_1) - 2\psi + E_x^u \left[c(X_2,A_2) + E_{X_2}^u(\phi(X_3))\right] \\
&= c(x,u_1) - 2\psi + E_x^u c(X_2,A_2) + E_x^u(\phi(X_3))
\end{aligned}$$

$$\leq \quad \ldots \leq \sum_{t=1}^{n} E_x^u c(X_t, A_t) - n\psi + E_x^u \phi(X_{n+1}).$$

The lemma follows by dividing by n and taking the limsup as n tends to infinity. □

Next we consider the case where the optimality inequality (12.1) holds in fact with equality. Consider the (expected) Average Cost Optimality Equation (ACOE):

$$\textbf{ACOE}: \qquad \phi(x) + \psi = \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y) \right], \quad (12.6)$$

where ψ is some constant, and $\phi : \mathbf{X} \to \mathbb{R}$. We note that if ACOE holds, then the pair (ψ, ϕ) is superharmonic. This allows us to combine both Lemmas 12.1 and 12.2 to get the following optimality results:

Lemma 12.3 (*Characterization of optimal value and policy*)
Assume that there exists a pair (ψ, ϕ) *satisfying the ACOE* (12.6)*, and that* (12.5) *holds for any Markov policy* u*. Let* w *be the stationary policy that chooses at state* x *an action that achieves the minimum of* $[c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y)]$*. Assume that*

$$\lim_{n \to \infty} \frac{E_x^w \phi(X_n)}{n} = 0.$$

Then (i) ψ *is the optimal value and* w *an optimal stationary policy, i.e.,* $\psi = C_{ea}(x,w) = C_{ea}(x)$*.*
(ii) C_{ea} *is the largest constant for which there exists a function* ϕ' *such that* $(ii.1)$ *the pair* (C_{ea}, ϕ') *is superharmonic and for which* $(ii.2)$ *for any Markov policy* u*,* (12.5) *holds.*

The following converse can be found in Arapostathis *et al.* (1993):

Lemma 12.4 (*The converse*)
Assume that there exists a pair (ψ, ϕ) *satisfying the ACOE* (12.6)*, and that*

$$\lim_{n \to \infty} \frac{E_x^u \phi(X_n)}{n} = 0$$

for all $u \in U_S$*. Then any optimal stationary policy* g *for which the state is irreducible and positive recurrent satisfies*

$$c(x,g) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xgy} \phi(y) = \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y) \right]. \quad (12.7)$$

Proof. Let $g \in U_S$ be optimal and assume that the state is irreducible and positive recurrent and that (12.7) does not hold. Then there exists some

state x_0 and action $a_0 \in \mathsf{A}(x_0)$ such that

$$\begin{aligned} c(x_0,g) + \sum_{y\in\mathbf{X}} \mathcal{P}_{x_0gy}\phi(y) &= \min_{a\in\mathsf{A}(x_0)} \left[c(x_0,a) + \sum_{y\in\mathbf{X}} \mathcal{P}_{x_0ay}\phi(y) \right] + \delta \\ &> c(x_0,a_0) + \sum_{y\in\mathbf{X}} \mathcal{P}_{x_0a_0y}\phi(y) \end{aligned}$$

for some $\delta > 0$. Let $g' \in U_S$ be the policy given by

$$g'(x) = \begin{cases} g(x) & \text{if } x \neq x_0 \\ a_0 & \text{if } x = x_0. \end{cases}$$

Using the ACOE, it follows from the irreducibility and positive recurrence that $C_{ea}(x_0,g') < C_{ea}(x,g)$, which contradicts the fact that g is optimal. □

We now introduce candidates to serve as the pair (ψ,ϕ) in ACOI or ACOE, and candidates for the optimal value and optimal stationary policies. In the following sections we shall establish for either the bounded cost assumptions or the Lyapunov assumptions, that these candidates are indeed an appropriate choice.

Assume that for any α in a neighborhood of 1, there exists an optimal stationary policy $g(\alpha)$ for the α-discount problem. Let α_n be some arbitrary sequence of discount factors converging to 1, along which the following limits exist:

$$g^* \stackrel{\text{def}}{=} \lim_{n\to\infty} g(\alpha_n) \tag{12.8}$$

$$h(x) = \lim_{n\to\infty} h_{\alpha_n}(x) \text{ where } h_\alpha(x) \stackrel{\text{def}}{=} \frac{C_\alpha(x) - C_\alpha(0)}{1-\alpha}, \quad \forall x \tag{12.9}$$

$$\psi^* \stackrel{\text{def}}{=} \lim_{n\to\infty} C_{\alpha_n}(0), \tag{12.10}$$

where 0 is some state. The pair (ψ^*, h) is the candidate for the functions that will satisfy the ACOI and ACOE, ψ^* is the candidate for the optimal value, and g^* for an optimal policy. We shall need some properties of h_α. Assume that $c \geq 0$. Let

$$T \stackrel{\text{def}}{=} \inf_{t>1}\{X_t = 0\}, \qquad W_\alpha := \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t). \tag{12.11}$$

Then $C_\alpha(x)$ can be written as

$$C_\alpha(x) = \min_{u\in U_M} \left[(1-\alpha) E_x^u W_\alpha + E_x^u \alpha^{T-1} C_\alpha(0) \right]. \tag{12.12}$$

Let $C_{tc}(x,u)$ be the total expected cost until hitting the set $\mathcal{M} = \{0\}$. It follows from (12.12) that $C_\alpha(0) \leq C_{tc}(0,u)$. If $C_{tc}(x,u) < \infty$ for all $x \in \mathbf{X}$,

then it follows from (12.12) that

$$C_\alpha(x) \le (1-\alpha)C_{tc}(x,u) + \alpha C_{tc}(0,u) < \infty,$$

and then

$$\begin{aligned} h_\alpha(x) &= \min_{u\in U_M}\left[E_x^u W_\alpha - \frac{1-E_x^u\alpha^{T-1}}{1-\alpha}C_\alpha(0)\right] \\ &= \min_{u\in U_M}\left[E_x^u W_\alpha - E_x^u\sum_{s=1}^{T-1}\alpha^{t-1}C_\alpha(0)\right]. \end{aligned} \tag{12.13}$$

Thus for any α,

$$h_\alpha(x) \le \min_{u\in U_M} E_x^u W_\alpha \le \min_{u\in U_M} C_{tc}(x,u). \tag{12.14}$$

Hence, if there exists some policy u for which $C_{tc}(x,u)$ is finite for all x, then for each x, $h_\alpha(x)$ is uniformly bounded over $\alpha \in (0,1)$.

12.2 Non-constrained control: cost bounded below

We assume that (11.1) holds, i.e., that the costs are bounded below. Without loss of generality, we shall assume that the costs are non-negative (since the optimality of a policy for the expected average cost is not affected by adding constants to the costs and to the corresponding bounds V).

Following Sennott (1989), we present below conditions for optimality of some stationary policies, and relate the values to the dynamic programming equation (12.1). We then present some sufficient conditions that are simpler to verify. The approach that we pursue is based on relating the expected average cost to the limit of discounted cost control problems.

Introduce some assumptions on the model:

- $\mathcal{S}$**1:** For every state $x \in \mathbf{X}$, and discount factor α, the value $C_\alpha(x)$ of the non-constrained MDP is finite.
- $\mathcal{S}$**2:** There exists a non-negative constant $\underline{h}$ such that

  $$-\underline{h} \le h_\alpha(x) := \frac{C_\alpha(x) - C_\alpha(0)}{1-\alpha}$$

  for all $x \in \mathbf{X}$ and discount factors α, and for some state $0 \in \mathbf{X}$.
- $\mathcal{S}$**3:** There exists some non-negative $\overline{m}(x)$ such that $h_\alpha(x) \le \overline{m}(x)$ for every x and α; moreover, for every x, there exists an action $a(x)$ such that

  $$\sum_{y\in\mathbf{X}} \mathcal{P}_{xa(x)y}\overline{m}(y) < \infty.$$

- $\mathcal{S}$**3*****:** There exists some non-negative $\overline{m}(x)$ such that $h_\alpha(x) \le \overline{m}(x)$ for every x and α, and, for every x and $a \in \mathsf{A}(x)$, $\sum_{y\in\mathbf{X}} \mathcal{P}_{xay}\overline{m}(y) < \infty$.

Remark 12.2 (*Sufficient conditions for $\mathcal{S}1$-$\mathcal{S}3$*)
If there exists a $g \in U_D$ under which the process is ergodic and irreducible with an invariant probability measure $\pi(g)$, and $\sum_{x\in\mathbf{X}} c(x,g)\pi(g) < \infty$, then Assumptions $\mathcal{S}1$ and $\mathcal{S}3$ hold. If $\mathbf{X}$ is fully ordered and $C_\alpha(x)$ is increasing in x, then Assumption $\mathcal{S}2$ holds. (See Arapostathis *et al.*, 1993, and Cavazos-Cadena and Sennott, 1992, for these results and for references to other sufficient conditions).

As is seen at the end of Section 12.2, if there exists some policy u for which $C_{tc}(x,u)$ is finite for all x, then for each x, the first part of $\mathcal{S}3$ holds, and one can choose $\overline{m}(x) = \inf_u C_{tc}(x,u)$. Moreover, it is easily seen from (12.13) that if the growth condition (11.20) holds, then $\mathcal{S}2$ is satisfied.

The following well-known Tauberian Theorem will turn to be very useful. For its proof, we refer e.g., to Sznadjer and Filar (1992).

Lemma 12.5 (*Tauberian Theorem*)
Let $\{a_n\}$ be a sequence of non-negative real numbers. Then

$$\underline{\lim}_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} a_t \le \underline{\lim}_{\alpha\to 1}(1-\alpha)\sum_{t=0}^{\infty}\alpha^t a_t \le \overline{\lim}_{\alpha\to 1}(1-\alpha)\sum_{t=0}^{\infty}\alpha^t a_t \le \overline{\lim}_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1} a_t.$$

The following is due to Sennott (1989):

Theorem 12.1 (*Existence of optimal values and stationary policies*)
Assume $\mathcal{S}1$-$\mathcal{S}3$, and consider non-negative immediate cost. Then
(*i*) *The value of the expected average control problem does not depend on the initial state x and is given as the limit of the discounted value*

$$C_{ea} = \lim_{\alpha\to 1} C_\alpha(x)$$

(this limit is independent of the sequence α_n in (12.8)).
(*ii*) *Any stationary policy g^* that is obtained as the limit of α-discount optimal policies $g(\alpha_n)$ (as in (12.8)) is optimal.*
(*iii*) *The pair (ψ^*, h) given in* (12.9) – (12.10) *satisfies the ACOI* (12.1). *If moreover, $\mathcal{S}3^*$ holds, then it satisfies the ACOE* (12.6).

Proof. For each α_n, the following holds for any fixed $x \in \mathbf{X}$:

$$C_{\alpha_n}(x) = (1-\alpha_n)c(x, g(\alpha_n)) + \alpha_n \sum_{y\in\mathbf{X}} \mathcal{P}_{xg(\alpha_n)y} C_{\alpha_n}(y).$$

By subtracting $C_{\alpha_n}(0)$ from both sides and dividing by $1-\alpha_n$, we get

$$C_{\alpha_n}(0) + h_{\alpha_n}(x) = c(x, g(\alpha_n)) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xg(\alpha_n)y} h_{\alpha_n}(y). \tag{12.15}$$

We now take the liminf in both sides and apply Fatou's Lemma (as h_{α_n} are bounded below by Assumption $\mathcal{S}2$), and obtain

$$\psi^* + h(x) \ge c(x, g^*) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xg^*y} h(y),$$

so that

$$\psi^* + h(x) \geq \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} h(y) \right].$$

This concludes the first part of (iii). The second part of (iii) follows by applying the dominated convergence theorem.

It follows from $\mathcal{S}1$ that for all x and $a \in \mathsf{A}(x)$,

$$C_{\alpha_n}(0) + h_{\alpha_n}(x) \leq c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} h_{\alpha_n}(y). \tag{12.16}$$

Thus, we get using $\mathcal{S}3$ and applying the dominated convergence theorem,

$$C_{\alpha_n}(0) + h(x) \leq c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} h(y). \tag{12.17}$$

We conclude that g^* minimizes the right-hand side of ACOI (12.1); so by Lemma 12.1, g^* satisfies $C_{ea}(x, g^*) \leq \psi^*$. On the other hand, it follows from Lemma 12.5 that for any policy u,

$$C_{ea}(x,u) \geq \overline{\lim_{\alpha_n \to 1}} C_{\alpha_n}(x,u) \geq \overline{\lim_{\alpha_n \to 1}} C_{\alpha_n}(x) = \psi^*. \tag{12.18}$$

We thus conclude that (i) and (ii) hold. □

12.3 Dynamic programming and uniform Lyapunov function

Next, we consider MDPs with uniform Lyapunov function with ν-bounded costs. We consider in their definition the set $\mathcal{M} = \{0\}$ where 0 is some arbitrary state. Let $T = T_0$ be the time to hit state 0 (see definition in (6.3)). In order to evaluate $C_\alpha(0)$ and h_α, we note that the discounted cost satisfies the following for $u \in U_S$:

$$C_\alpha(x,u) = (1-\alpha) E_x^u \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t) + E_x^u \alpha^{T-1} C_\alpha(0,u).$$

Hence,

$$\begin{aligned} C_\alpha(0,u) &= \frac{(1-\alpha) E_0^u \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t)}{1 - E_\beta^u \alpha^{T-1}} \\ &= \frac{E_0^u \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t)}{E_0^u \sum_{t=1}^{T-1} \alpha^{t-1}}. \end{aligned} \tag{12.19}$$

This implies

$$\begin{aligned} &\frac{C_\alpha(x,u) - C_\alpha(0,u)}{1-\alpha} \\ &= E_x^u \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t) + \frac{E_x^u \alpha^{T-1} - 1}{1-\alpha} C_\alpha(0,u) \end{aligned}$$

$$\begin{aligned} &= E_x^u \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t) - C_\alpha(0,u) E_x^u \sum_{t=1}^{T-1} \alpha^{t-1} \\ &= E_x^u \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t) - E_0^u \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t). \end{aligned}$$

Hence,

$$\begin{aligned} |h_\alpha(x,u)| &\stackrel{\text{def}}{=} \left| \frac{C_\alpha(x,u) - C_\alpha(0,u)}{1-\alpha} \right| \\ &\leq E_x^u \sum_{t=1}^{T-1} \alpha^{t-1} \nu(X_t, A_t) + E_0^u \sum_{t=1}^{T-1} \alpha^{t-1} \nu(X_t, A_t) \\ &\leq \hat{M}(x) + \hat{M}(0) \leq \mu(x) + \mu(0) \end{aligned}$$

(the last inequality follows from Lemma 7.5 (ii)). Since we know that there exists a uniformly optimal stationary policy u_α for the discounted cost, the above implies that

$$|h_\alpha(x)| \leq \hat{M}(x) + \hat{M}(0) \leq \mu(x) + \mu(0). \tag{12.20}$$

Theorem 12.2 (*Existence of optimal values and stationary policies*) *Consider an MDP with a uniform Lyapunov function and ν-bounded immediate costs. Then*
(*i*) *The value of the expected average control problem does not depend on the initial distribution β and is given as the limit of the value of the discounted problem*

$$C_{ea}(\beta) = \lim_{\alpha \to 1} C_\alpha(\beta)$$

(*this limit is independent of the sequence α_n in* (12.8)).
(*ii*) *Any stationary policy g^* that is obtained as limit of α-discount optimal policies $g(\alpha_n)$ (as in* (12.8)) *is optimal.*
(*iii*) *The pair (ψ^*, h) given in* (12.9) – (12.10) *satisfies the ACOE* (12.6).

Proof. We begin with (iii). For each α_n, the following holds for any fixed $x \in \mathbf{X}$:

$$\begin{aligned} C_{\alpha_n}(0) + h_{\alpha_n}(x) &= c(x, g(\alpha_n)) + \sum_{y \in \mathbf{X}} P_{xy}(g(\alpha_n)) h_{\alpha_n}(y) \\ &= c(x, g(\alpha_n)) + \sum_{y \in \mathbf{X}} {}_0P_{xy}(g(\alpha_n)) h_{\alpha_n}(y) \end{aligned} \tag{12.21}$$

(see (12.15)). h_α are μ-bounded, uniformly in α (this follows from (12.20)); there exists some constant $\overline{m}$ such that

$$|h_\alpha(x)| \leq \overline{m} \mu(x),$$

for all α smaller than 1.

Since $P(u)\mu$ is continuous over U_D (property M1(ii) in the definition of

the uniform Lyapunov function), we may take the liminf in both sides of (12.21) and apply a dominated convergence theorem, to obtain

$$\psi^* + h(x) = c(x, g^*) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xg^*y} h(y).$$

This establishes (iii). Since $h(x) \leq \mu(x) + \mu(0)$, it follows by property M6 (and Lemma 7.5 (i)) that

$$\lim_{n \to \infty} \frac{E_x^u h(X_n)}{n} = 0 \tag{12.22}$$

for all policies u and states x. Lemma 12.3(i) now implies that ψ^* is the optimal value, from which statement (i) follows, and also implies statement (ii). □

An alternative way to show part (i) of the theorem is by establishing that $C_\alpha(\beta, u)$ converges to $C_{ea}(\beta, u)$ uniformly over $u \in U_S$. Since stationary policies are dominant for both the discounted and the expected average cost, this implies the convergence of the values. (All details of the above statements are presented in the proof of Lemma 14.1.)

12.4 Superharmonic functions and linear programming

Fix an initial distribution β.

Theorem 12.3 (*The value and superharmonic functions: MDPs with uniform Lyapunov function*)
Consider an MDP with a uniform Lyapunov function and ν-bounded immediate costs. Then
(*i*) *The pair $(C_{ea}(\beta), h)$ is superharmonic, and h is μ-bounded (h is as in (12.9)).*
(*ii*) *For any other superharmonic pair (ψ, ϕ) for which $\phi : \mathbf{X} \to \mathbb{R}$ is μ-bounded, we have $C_{ea}(\beta) \geq \psi$.*
(*In other words, consider the class of superharmonic pairs (ψ, ϕ) for which $\phi : \mathbf{X} \to \mathbb{R}$ are μ-bounded. The value $C_{ea}(\beta)$ is the largest constant for which there exists a μ-bounded function $\overline{\phi} : \mathbf{X} \to \mathbb{R}$ such that $(C_{ea}(\beta), \overline{\phi})$ is within the above class.*)

Proof. (i) follows from Theorem 12.2 (i) and (iii).
(ii) follows from Lemma 12.3(ii), since for any μ-bounded ϕ,

$$\lim_{n \to \infty} \frac{E_x^u \phi(X_n)}{n} = 0$$

by the same arguments as in the proof of Theorem 12.2, and hence (12.5) holds. □

Motivated by Theorem 12.3, we introduce the following infinite Linear Program with decision variables $\psi \in \mathbb{R}$ and $\phi(y), y \in \mathbf{X}$, which may be

used to obtain the optimal expected average value of **COP**.

$$\begin{array}{lll} \mathbf{DP}(\beta): & \text{Find } \Theta^* := \sup_{\psi,\phi} \psi \text{ subject to} & \\ \phi(x) + \psi & \leq & c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y) \qquad \forall x \in \mathbf{X}, a \in \mathsf{A}(x), \psi \in \mathbb{R}\,. \end{array}$$

Theorem 12.3 implies the following:

Theorem 12.4 (*The dual linear program, uniform Lyapunov function*) *Consider an MDP with a uniform Lyapunov function and ν-bounded immediate costs. Consider the dual program* $\mathbf{DP}(\beta)$*, where ϕ is restricted to the linear space* F^{μ}*. Then* $\mathbf{DP}(\beta)$ *is feasible; its value equals* $C_{ea}(\beta)$*, and* $(C_{ea}(\beta), \phi)$*, with $\phi = h$, is an optimal solution.*

We now obtain a similar statement for the case of non-negative costs, when restricting to bounded functions (for which (12.5) clearly holds under any policy). The fact that we restrict to the subclass of functions satisfying the conditions of Lemma 12.3(ii) might lead to a value which is only a lower bound on the original value (without the restriction). However, it will turn out that the family of bounded functions ϕ is rich enough to yield the same value as the one obtained by the richer class of policies satisfying (12.5).

Theorem 12.5 (*The dual linear program, non-negative immediate costs*) *Assume that the immediate costs are non-negative, and the standard moment condition* (11.21) *holds. Assume further that there exists a policy for which the total expected cost from any state to state 0 is finite. Consider* $\mathbf{DP}(\beta)$ *where the decision variables ϕ are bounded functions. Then for any initial distribution β,* $\mathbf{DP}(\beta)$ *is feasible and its value equals* $C_{ea}(\beta)$*.*

Proof. Denote by $C^1(\beta)$ the value of $\mathbf{DP}(\beta)$ restricted to bounded ϕ. Since for any bounded function ϕ eq. (12.5) holds for all policies, we have by Lemma 12.2

$$C^1(\beta) \leq C_{ea}(\beta). \tag{12.23}$$

Consider a set of approximating **COP**s with an immediate cost $c_n(x,a) = \min\{n, c(x,a)\}$; denote by $C^n_\alpha(x,u)$ the corresponding infinite horizon expected discounted cost. Denote by $C^n_{tc}(\beta,u)$ the corresponding total expected cost until state 0 is reached. Denote

$$h^n_\alpha(x) := \frac{C^n_\alpha(x) - C^n_\alpha(0)}{1-\alpha_n}.$$

The pair $(C^n_\alpha(0), h^n_\alpha)$ is superharmonic, since, by the same arguments as those that yield (12.15),

$$\begin{aligned} C^n_\alpha(0) + h^n_\alpha(x) &\leq c_n(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xg(\alpha)y} h^n_\alpha(y) \\ &\leq c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xg(\alpha)y} h^n_\alpha(y), \end{aligned} \tag{12.24}$$

where $g(\alpha)$ is an optimal stationary policy for the α-discounted MDP. Consider an arbitrary sequence α_n converging to 1, along which the following limits exist:

$$C^* = \lim_{n\to\infty} C^n_{\alpha_n}(0), \qquad h^*(x) = \lim_{n\to\infty} h^n_{\alpha_n}(x), \quad \forall x, \qquad g^* = \lim_{n\to\infty} g^*(n),$$

where $g^*(n)$ is an optimal stationary policy for the α_n-discounted MDP. Since c_n are bounded by n, we have $C^n_\alpha(x) \le n/(1-\alpha)$. Hence $h^n_\alpha(x)$ are bounded (in x) by $n/(1-\alpha)$. Since, for any fixed $\alpha \in (0,1)$ and n, the pair $(C^n_\alpha(0), h^n_\alpha)$ is feasible for $\mathbf{DP}(\beta)$, and since h^n_α are bounded in x, we have

$$C^* \le C^1(\beta). \tag{12.25}$$

For any α and n, we have (as follows from (12.14))

$$h^n_\alpha(x) \le \inf_u C^n_{tc}(x,u) \le \inf_u C_{tc}(x,u),$$

and thus, in particular, $h^*(x) \le \inf_u C_{tc}(x,u)$ is finite. For each $x \in \mathbf{X}$ and n, we have

$$C^n_{\alpha_n}(0) + h^n_{\alpha_n}(x) = c_n(x, g(n)) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xg(n)y} h^n_{\alpha_n}(y),$$

as in (12.15).

One may verify from the growth condition (11.21) and from (12.13) that $h^n_\alpha(x)$ are uniformly bounded from below by some constant independent of α and n (which implies in particular condition $\mathcal{S}2$). Taking the limit as n tends to infinity, we get, by Fatou's Lemma,

$$C^* + h^*(x) \ge c(x, g^*) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xg^*y} h^*(y).$$

We conclude that (C^*, h^*) satisfy ACOI. Since h^* is bounded below, it satisfies (12.2), so by Lemma 12.1, $C_{ea}(\beta) \le C^*$. This, together with (12.23) and (12.25), establishes the proof. □

Remark 12.3 (*Non-negativity of the immediate cost*)
It is in (12.24) that we made use of the non-negativity of the immediate cost. One can relax the non-negativity assumption by assuming that the immediate costs are bounded below, since the optimal value and optimal policies can always be computed by shifting all costs by some constant, so that they can be non-negative.

Remark 12.4 (*Relaxing the growth condition*)
In the above theorem, the growth condition was only needed in order to ensure that Condition $\mathcal{S}2$ holds in a slightly stronger version: $h^n_\alpha(x)$ should be bounded below, uniformly in n and α. It can thus be relaxed by other weaker sufficient conditions.

Remark 12.5 (*On the methodology of approximation*)
The method used to establish the convergence of the approximation scheme in the proof of Theorem 12.5 is similar in spirit to the method used by Sennott (1995) to obtain finite state approximation.

Remark 12.6 (*Initial distributions and infinite costs*)
Note that we allowed for arbitrary β. It may happen, however, that $C_{ea}(x)$ is finite for all x, but β is chosen such that $C_{ea}(\beta)$ is infinite.

12.5 Set of achievable costs

Define for any subset $\overline{U}$ of policies the set of achievable vector performance measures:

$$\mathbf{M}^{ea}_{\overline{U}}(\beta) = \cup_{u\in\overline{U}}\{(C_{ea}(\beta,u), D^k_{ea}(\beta,u), k=1,\ldots,K)\}, \tag{12.26}$$

and set $\mathbf{M}^{ea}(\beta) := \mathbf{M}^{ea}_{U}(\beta) \cup \mathbf{M}^{ea}_{\overline{M}(U_M)}(\beta)$. Define also

$$\mathbf{V}_{ea}(\beta) := \bigcup_{\rho\in\ \mathbf{Q}_{ea}(\beta)} \{(\langle\rho,c\rangle, \langle\rho,d^1\rangle, \langle\rho,d^2\rangle, \ldots, \langle\rho,d^K\rangle)\}, \tag{12.27}$$

where $\mathbf{Q}_{ea}(\beta)$ is given in (11.5).

The next characterization of achievable costs follows by combining Theorems 11.3, 11.4, 11.7 and 11.9.

Theorem 12.6 (*Characterization of the sets of achievable costs*)
(*i*) *Assume that the immediate cost is bounded below* (11.1) *and satisfies the growth condition* (11.20) *or* (11.22). *Then*
$\mathbf{M}^{ea}_{U_S}(\beta)$ *and* $\mathbf{M}^{ea}(\beta)$ *are convex, and*

$$\mathbf{V}_{ea}(\beta) = \mathbf{M}^{ea}_{U_S}(\beta) \prec \mathbf{M}^{ea}_{U_M}(\beta) = \mathbf{M}^{ea}(\beta)$$

($\prec$ *is defined in Section* 8.1).
(*ii*) *In the case of uniform Lyapunov function and* ν*-bounded immediate costs,* $\mathbf{M}^{ea}_{U_S}(\beta)$ *is convex and compact, and satisfies*

$$\mathbf{M}^{ea}_{\mathcal{U}}(\beta) = \mathbf{M}^{ea}_{U_S}(\beta) = \overline{co}\mathbf{M}^{ea}_{U_D}(\beta) \prec \mathbf{M}^{ea}(\beta) = \mathbf{V}_{ea}(\beta).$$

12.6 Constrained control: Lagrangian approach

By the same arguments as the ones used to establish Theorems 9.9 and 9.10, we now obtain:

Theorem 12.7 (*The Lagrangian*)
Consider either (1) *the immediate cost bounded below* (11.1) *and satisfying the growth condition* (11.20) *or* (11.22), *or* (2) *the MDP has a uniform Lyapunov function and* ν*-bounded immediate costs.*
(*i*) *Let* $\overline{U}$ *be any class of policies containing* U_s. *Then the value function*

satisfies

$$C_{ea}(\beta) = \inf_{u\in\overline{U}} \sup_{\lambda\geq 0} J^{\lambda}_{ea}(\beta, u), \tag{12.28}$$

where

$$J^{\lambda}_{ea}(\beta, u) \quad := \quad C_{ea}(\beta, u) + \langle \lambda, D_{ea}(\beta, u) - V \rangle. \tag{12.29}$$

(ii) A policy u^* is optimal for **COP** *if and only if* $C_{ea}(\beta) = \sup_{\lambda\geq 0} J^{\lambda}_{ea}(\beta, u^*)$.
(iii) For any class $\overline{U}$ containing U_D, the value function satisfies

$$C_{ea}(\beta) = \sup_{\lambda\geq 0} \min_{u\in\overline{M}(U_S)} J^{\lambda}_{ea}(\beta, u) = \sup_{\lambda\geq 0} \min_{u\in\overline{U}} J^{\lambda}_{ea}(\beta, u). \tag{12.30}$$

Moreover, there exist some $u^* \in \mathcal{U}$ such that

$$C_{ea}(\beta) = \inf_{u\in\mathcal{U}} \sup_{\lambda\geq 0} J^{\lambda}_{ea}(\beta, u) = \sup_{\lambda\geq 0} J^{\lambda}_{ea}(\beta, u^*), \tag{12.31}$$

and u^* is optimal for **COP**.

Proof. For the proof, we make use of the fact that the stationary policies are dominant (see Theorems 11.6, 11.7 and 11.9). The proof of (i) and (ii) is exactly the same as the proof of the similar results in Theorem 12.7. The proof of (iii) is the same, except that $\overline{M}(U_S)$ replaces $\overline{M}(U_M)$. □

Remark 12.7 (*Comparing Theorem 9.9 and Theorem 12.7*)
In the above proof, we used from the beginning the fact that the stationary policies are dominant. Hence, the proof makes heavy use of the results obtained in the last chapter, concerning the properties of occupation measures achieved by stationary policies. This was not necessary for the proof in the analogous Theorem 9.9 for the case of total expected cost. There we could work directly with the policies U_M and $\overline{M}(U_M)$, since the Lagrangian, for the total expected cost, is lower semi-continuous in U_M (and hence in $\overline{M}(U_M)$). For the expected average case, this is generally not true.

By the same type of arguments as in the proof of Corollary 9.1, we obtain from (12.31) the following:

Corollary 12.1 (*Dominance of $\mathcal{U}$*)
Let the immediate costs be bounded below (11.1) *and satisfy the growth condition* (11.20) *or* (11.22). *Then $\mathcal{U}$ is a dominating class of policies.*

Corollary 12.2 (*Saddle point*)
Consider either the case of immediate cost bounded below (11.1) *and satisfying the growth condition* (11.20) *or* (11.22), *or the case of a uniform Lyapunov function with ν-bounded immediate costs. Then for any class of policies $\overline{U}$ that contains either U_S or $\mathcal{U}$, we have*

$$C_{ea}(\beta) = \sup_{\lambda\geq 0} \min_{u\in\overline{U}} J^{\lambda}_{ea}(\beta, u) = \min_{u\in\overline{U}} \sup_{\lambda\geq 0} J^{\lambda}_{ea}(\beta, u) = \sup_{\lambda\geq 0} J^{\lambda}_{ea}(\beta, u^*)$$

for some $u^* \in \overline{U}$.

The existence of minimizing Lagrange multipliers is summarized in the following theorem, whose proof follows the same lines as the proof of Theorem 9.10.

Theorem 12.8 (*The Lagrangian: Slater condition*)
Under the conditions of Corollary 12.2*, if there exists some policy* u *for which* $D_{ea}(\beta, u) < V$*, then there exist non-negative Lagrange multipliers* $\lambda^* = \{\lambda_1^*, \ldots, \lambda_K^*\}$ *such that*

$$C_{ea}(\beta) = \min_{u \in \mathcal{U}} J_{ea}^{\lambda^*}(\beta, u) = \min_{u \in U_D} J_{ea}^{\lambda^*}(\beta, u).$$

Moreover, any optimal policy u^* *satisfies the Kuhn-Tucker conditions:*

$$\lambda_k^*(D_{ea}^k(\beta, u^*) - V_k) = 0, \qquad k = 1, \ldots, K.$$

12.7 The dual LP

For any $u \in U_S$,

$$J_{ea}^{\lambda}(\beta, u) = \lim_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} E_\beta^u j^\lambda(X_s, A_s)$$

where

$$j^\lambda(x, a) := c(x, a) + \langle \lambda, d(x, a) \rangle.$$

This, together with the results of Section 12.4, suggests that the following LP can be used to compute the optimal value of **COP**, with decision variables $\psi \in \mathbb{R}$, $\phi \in \mathbf{X} \to \mathbb{R}$ and $\lambda \in \mathbb{R}_+^K$.

$$\begin{array}{lll} \mathbf{DP_3}(\beta)\text{:} & \text{Find } \Theta^*(\beta) := \sup_{\psi,\phi,\lambda} \psi - \langle \lambda, V \rangle \text{ subject to} & \\ \phi(x) + \psi & \leq & c(x, a) + \langle \lambda, d(x, a) \rangle + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y), \quad x \in \mathbf{X}, a \in \mathsf{A}(x). \end{array}$$

Combining Theorem 12.7 with the results of Section 12.4, we get:

Theorem 12.9 (*The dual LP, the case of a uniform Lyapunov function*)
Consider the case of a uniform Lyapunov functions and ν*-bounded costs. Consider the dual program* $\mathbf{DP}(\beta)$*, where* ϕ *is restricted to the linear space* F^μ*. Then* $\mathbf{DP_3}(\beta)$ *is feasible if and only if* **COP** *is feasible. The value of* $\mathbf{DP_3}(\beta)$ *equals* $C_{ea}(\beta)$ *and* $\phi(x) = h(x), x \in \mathbf{X}$ *(where* h *is given in* (12.9)*) is an optimal solution.*

Theorem 12.10 (*The dual LP, cost bounded below*)
Assume that for each one of the immediate cost functions $c(\cdot, \cdot)$*, and* $d^k(\cdot, \cdot)$*,* $k = 1, \ldots, K$*, the following hold:*

- *The immediate costs are non-negative.*
- *The standard moment condition (11.21) holds.*
- *Conditions* $\mathcal{S}1$ *and* $\mathcal{S}3$ *hold for the (non-constrained) MDP with the corresponding immediate cost.*

- *There exists some policy for which the total corresponding expected cost from any state to state 0 is finite.*

Consider $\mathbf{DP_3}(\beta)$ *where the decision variables* ϕ *are bounded functions. Then for any initial state* β, $\mathbf{DP}(\beta)$ *is feasible and its value equals* $C_{ea}(\beta)$.

$\mathbf{DP_3}(\beta)$ is the dual LP to $\mathbf{LP_3}(\beta)$. By comparing Theorems 11.10 to 12.9 and 12.10, we see that there is no duality gap between $\mathbf{LP_3}(\beta)$ and $\mathbf{DP_3}(\beta)$.

12.8 A second LP approach for optimal mixed policies

In this section we present an alternative LP formulation for **COP**. The decision variables will correspond to the probability measures over the space of all stationary deterministic policies.

If the immediate cost is bounded below and satisfies the growth condition (11.20) or (11.22), or if the MDP has a uniform Lyapunov function with ν-bounded immediate cost, then we know by Corollary 12.1 and by Theorem 11.7 (i) that $C_{ea}(\beta)$ is equal to the value of **COP** restricted to $\mathcal{U}$:

$$\min_{u \in \mathcal{U}} C_{ea}(\beta, u) \text{ subject to } D_{ea}(\beta, u) \leq V.$$

This can be rewritten as a Linear Program:

$$\mathbf{LP_4}(\beta): \qquad \min_{\gamma \in M_1(U_D)} \int C_{ea}(\beta, u)\gamma(du)$$

$$\text{subject to} \qquad \int D^k_{ea}(\beta, u)\gamma(du) \leq V^k, \quad k = 1, \ldots, K. \quad (12.32)$$

This yields the following:

Theorem 12.11 (*Relation between* **COP** *and* $\mathbf{LP_4}(\beta)$)
Consider either the case of immediate cost bounded below and satisfying the growth condition (11.20) *or* (11.22)*, or the case of a uniform Lyapunov function with* ν*-bounded immediate costs. Then*
(*i*) **COP** *is feasible if and only if* $\mathbf{LP_4}(\beta)$ *is feasible* (*i.e., the set satisfying* (12.32) *is non-empty*) *If* $\mathbf{LP_4}(\beta)$ *is feasible, then there exists an optimal policy in* $\mathcal{U}$ *for* **COP**.
(*ii*) *The values of* **COP** *and of* $\mathbf{LP_4}(\beta)$ *are equal.*
(*iii*) *If* γ *is a solution of* $\mathbf{LP_4}(\beta)$*, then the policy* $\hat{\gamma} \in \mathcal{U}$ *is optimal for* **COP**.

PART III

Part Three: Asymptotic methods and approximations

CHAPTER 13

Sensitivity analysis

13.1 Introduction

We consider in this chapter a sequence $\mathbf{COP}_n$, $n = 1, 2, \ldots$ of CMDPs and a 'limit' CMDP, denoted by $\mathbf{COP}_\infty$, or simply by **COP**. **COP** is assumed to be feasible, and therefore, under the standard conditions developed in the previous chapters, to have an optimal solution. However, for any given n, $\mathbf{COP}_n$ need not be feasible, and even if it is, it need not possess an optimal solution (i.e., it may only have ε-optimal solutions). We are interested in the following questions:

- (i) Do the values of $\mathbf{COP}_n$ converge to the value of **COP**? If so, then at what rate?
- (ii) Do optimal (or almost optimal) policies converge in some sense?
- (iii) Given an (almost) optimal policy for $\mathbf{COP}_n$, will it be an almost optimal policy for **COP** if n is sufficiently large?
- (iv) Conversely, given an optimal policy for **COP**, will it be an almost optimal policy for $\mathbf{COP}_n$ for all n sufficiently large?

We shall proceed as follows. We first introduce a general framework for approximations that will provide sufficient conditions for obtaining convergence in the sense of (i) and (ii) above, and will provide also the rate of convergence. It turns out that the answers to (iii) and to (iv) are in general negative, unlike the unconstrained case. The reason is that an optimal policy for $\mathbf{COP}_n$ may be infeasible for **COP**, and vice versa. We shall, however, establish sufficient conditions for the following slightly weaker version of (iii) and (iv):

- (iii′) Given an optimal policy for $\mathbf{COP}_n$, can we perturb it 'slightly' so that it becomes almost optimal for **COP** if n is sufficiently large?
- (iv′) Given an optimal policy for **COP**, can we perturb it 'slightly' so that it becomes almost optimal for $\mathbf{COP}_n$ for all n sufficiently large?

As applications of the general framework, we shall examine in the next chapters the convergence of values and policies in the discount factor, including the case when it converges to one and the convergence in the horizon as it tends to infinity. In Chapter 16 we further use the results below to obtain algorithms based on finite-state truncation, for computing optimal policies and values of MDPs with a countable state space.

To illustrate the usefulness of the results for approximations, we note that finite horizon CMDPs have, in general, Markov optimal policies, and their computation is very costly for large horizon. Infinite horizon CMDPs, on the other hand, have optimal stationary (or mixed stationary) policies, and their computation is much less costly. A constructive answer to question (iv′) will thus provide us with an efficient method for obtaining almost optimal stationary policies for CMDPs with finite (but large) horizon.

Another application of the approximation results is adaptive CMDPs. It is assumed that the transition probabilities are unknown to the controller. The controller thus has to design a policy whose role combines estimation and control. Under suitable conditions, an efficient estimation can be guaranteed, i.e., the estimated transition probabilities converge to the true value almost surely. The controls are updated according to the 'Certainty Equivalence' rule: at any given time, the policy that is used imitates the one that would be optimal for a CMDP whose transition probabilities are those given by the current estimations. The asymptotic results of the current chapter can be used to prove the optimality of that policy for the countable state space. For the precise formulation and solution of adaptive control of CMDPs in the finite state and action spaces, see Altman and Shwartz (1991a, 1991b).

We briefly mention some related work on the continuity and sensitivity analysis of mathematical programs, and of control problems. Many papers and books are devoted to the continuity of mathematical programs in the case of the finite-dimensional state, e.g., Dantzig *et al.* (1967), Pervozvanskii and Gaitsgory (1986, 1988). Several special issues of scientific journals have focused on such questions, as well as other related sensitivity, stability and parametric analysis: *Mathematical Programming* **21**, 1984, *Annals of Operations Research* **27**, 1990. Similar questions to those addressed in this chapter were studied in Fiacco (1974) and in Schochetman (1990), and some of the results there are close to those in the first part of the chapter. Some other related references are Birge and Wets (1996), Kanniappan and Sastry (1974), Lignotat and Morgan (1992), Lucchetti and Wets (1993), Schochetman (1990), and Schochetman and Smith (1991).

Convergence results for constrained dynamic control problems have been obtained by Altman and Shwartz (1991b, 1991c), Altman and Gaitsgory (1993), Altman (1993, 1994), and Tidball and Altman (1995). Conditions were obtained there for the convergence in the transition probabilities, in the horizon and in the immediate cost. Conditions for the non-continuity, and the analysis of the limiting behavior for these cases have been obtained by Altman and Gaitsgory (1993).

Our approach below to obtain convergence conditions is based on Lagrangian techniques, and they are related to the techniques in Rockafellar (1989).

We begin by developing Key Theorems for approximating a **COP** by a sequence $\mathbf{COP}_n$. **COP** is called the limit problem, and will stand for either the finite horizon problem, or the infinite horizon discounted problem, total cost problem, or the infinite horizon expected average problem. In fact, the results of this section hold for any constrained optimization problem where some costs are defined over some topological space (of policies) $\overline{U} \subset U$: $C(\cdot) : U \to \mathbb{R}$, $D(\cdot) : U \to \mathbb{R}^K$. These costs may stand for the finite horizon, infinite horizon discounted costs, total cost, or expected average cost. We consider $\mathbf{COP}(\overline{U})$:

$$\inf_{u \in \overline{U}} C(u) \quad \text{subject to} \quad D(u) \leq V.$$

Denote by $C^{\overline{U}}$ the value of $\mathbf{COP}(\overline{U})$. Assume that

$$|C(u)| < \overline{B} \tag{13.1}$$

for all $u \in \overline{U}$. We shall use below e to denote a K-dimensional vector whose components are all 1.

We consider next a sequence $\mathbf{COP}_n(\overline{U})$, also called the approximating problems, defined as follows. Consider a sequence of cost functions $C_n : \overline{U} \to \mathbb{R}$, $D_n : \overline{U} \to \mathbb{R}^K$, $n = 1, 2, \ldots$; $\mathbf{COP}_n(\overline{U})$ is defined by:

$$\inf_{u \in \overline{U}} C_n(u) \quad \text{subject to} \quad D_n(u) \leq V.$$

Denote by $C_n^{\overline{U}}$ the value of $\mathbf{COP}_n(\overline{U})$.

Remark 13.1 (*Set of policies that depend on n*)
The sets of policies in the above setting do not depend on n. There are cases, however, where it is desirable to allow such a dependence. An example is the finite approximation scheme III in Section 16.4. All the results we present here generalize to this case, using the same types of arguments, see Tidball and Altman (1995). However, for simplicity of presentation we restrict ourselves to the simpler model.

We introduce the following assumptions.

- (**S**1): Slater-type condition:

$$\exists v \in \overline{U} \text{ such that } D(v) < V. \tag{13.2}$$

- (**S**2): Saddle-point condition: For any value of right-hand side constraints V for which (**S**1) holds, there exists $u^* \in \overline{U}$ and $\lambda^* \in R^K$ with $\lambda^* \geq 0$, (which depend on V) such that

$$\begin{aligned} C^{\overline{U}} &= C(u^*) = \min_{u \in \overline{U}} \max_{\lambda \geq 0} [C(u) + \langle \lambda, D(u) - V \rangle] \\ &= \max_{\lambda \geq 0} \min_{u \in \overline{U}} [C(u) + \langle \lambda, D(u) - V \rangle] \\ &= \max_{\lambda \geq 0} [C(u^*) + \langle \lambda, D(u^*) - V \rangle] \end{aligned}$$

$$= \min_{u\in\overline{U}} \left[C(u) + \langle \lambda^*, D(u) - V\rangle\right].$$

We shall sometimes use the notation u_V^* and λ_V^* to express the dependence on V.

- (**S**3): $C_n(u)$ and $D_n(u)$ converge to $C(u)$ and $D(u)$ uniformly over $u \in \overline{U}$, i.e., there exists some sequence $\varepsilon_1(n) \in \mathbb{R}$, $n = 1, 2, \ldots$ such that for all $u \in \overline{U}$,

$$\lim_{n\to\infty} \varepsilon_1(n) = 0,$$

and for all $n \in \mathbb{N}$,

$$|C_n(u) - C(u)| < \varepsilon_1(n), \qquad |D_n^k(u) - D^k(u)| < \varepsilon_1(n),\ k = 1, \ldots, K.$$

Remark 13.2 (*The unconstrained case*)
Our results will be applicable even for unconstrained MDPs. In that case, (**S**1) and (**S**2) hold trivially.

13.2 Approximation of the values

The following theorem establishes the convergence of the values, and the rate of convergence.

Theorem 13.1 (*Convergence of the values*)
Denote $\eta(V) := \min_{k=1,\ldots,K}[V_k - D^k(v)]$. *Assume* (**S**1) − (**S**3). *Then the values converge, i.e.,*

$$\lim_{n\to\infty} C_n^{\overline{U}} = C^{\overline{U}}.$$

Moreover, for all n *large enough,* $|C^{\overline{U}} - C_n^{\overline{U}}|$ *is of the order of* $\varepsilon_1(n)$, *i.e.,*

$$\overline{\lim_{n\to\infty}} \frac{|C^{\overline{U}} - C_n^{\overline{U}}|}{\varepsilon_1(n)} \le \left(1 + \frac{2\overline{B}}{\eta(V)}\right). \tag{13.3}$$

In order to establish the theorem, we need the following lemmas.

Lemma 13.1 (*Bound on the sum of Lagrange multipliers*)
Assume (**S**2) *and* (**S**1). *Then*

$$\langle \lambda^*, e\rangle \le \frac{2\overline{B}}{\eta(V)}. \tag{13.4}$$

Proof.

$$\begin{aligned} -\overline{B} &\le C^{\overline{U}} \\ &= \min_{u\in\overline{U}} \left[C(u) + \langle \lambda^*, D(u) - V\rangle\right] \\ &\le C(v) + \langle \lambda^*, D(v) - V\rangle \\ &\le \overline{B} + \langle \lambda^*, D(v) - V\rangle. \end{aligned}$$

Hence,

$$\langle \lambda^*, V - D(v) \rangle \leq 2\overline{B}. \tag{13.5}$$

We then obtain (13.4) by noting that $\eta(V)\langle \lambda^*, e \rangle \leq \langle \lambda^*, V - D(v) \rangle$. □

The following lemma shows that a property similar to (**S**2) holds also for $\mathbf{COP}_n(\overline{U})$, for n large enough.

Lemma 13.2 (*Asymptotic properties of the approximating problems*) *Assume* (**S**1) − (**S**3). *Fix some* δ_0 *with* $0 < \delta_0 < \eta(V)$, *and denote*

$$k_1 = 1 + \frac{2\overline{B}}{\eta(V) - \delta_0}.$$

For all n *large enough,* $\mathbf{COP}_n(\overline{U})$ *is feasible, and*

$$\begin{aligned} C_n^{\overline{U}} &= \inf_{u \in \overline{U}} \sup_{\lambda \geq 0} [C_n(u) + \langle \lambda, D_n(u) - V \rangle] \\ &\leq \sup_{\lambda \geq 0} \inf_{u \in \overline{U}} [C_n(u) + \langle \lambda, D_n(u) - V \rangle] + 2\varepsilon_1(n)k_1. \end{aligned} \tag{13.6}$$

Moreover, there exists $u_n^* \in \overline{U}$ *and* $\lambda_n^* \in R^K$ *with* $\lambda_n^* \geq 0$, *and*

$$\langle \lambda_n^*, e \rangle \leq \frac{2\overline{B}}{\eta(V) - \delta_0} \tag{13.7}$$

such that

$$\begin{aligned} &\inf_{u \in \overline{U}} \sup_{\lambda \geq 0} [C_n(u) + \langle \lambda, D_n(u) - V \rangle] \\ &\quad \leq \inf_{u \in \overline{U}} [C_n(u) + \langle \lambda_n^*, D_n(u) - V \rangle] + 2\varepsilon_1(n)k_1 \end{aligned} \tag{13.8}$$

and

$$\begin{aligned} &\sup_{\lambda \geq 0} \inf_{u \in \overline{U}} [C_n(u) + \langle \lambda, D_n(u) - V \rangle] \\ &\quad \geq \sup_{\lambda \geq 0} [C_n(u_n^*) + \langle \lambda, D_n(u_n^*) - V \rangle] - 2\varepsilon_1(n)k_1. \end{aligned} \tag{13.9}$$

Proof. The upper bound of λ_n^* follows by applying Lemma 13.1 to $\mathbf{COP}(\overline{U})$ with $V - \varepsilon_1(n)e$ replacing V.

We shall prove the lemma by using for u_n^*, λ_n^* the pair $(u^*_{V-\varepsilon_1(n)e}, \lambda^*_{V-\varepsilon_1(n)e})$ defined in (**S**2) (corresponding to $\mathbf{COP}(\overline{U})$ with the right-hand side constraint V replaced by $V - \varepsilon_1(n)e$). Consider n sufficiently large so that $\varepsilon_1(n) < \delta_0$.

$$\begin{aligned} &\inf_{u \in \overline{U}} \sup_{\lambda \geq 0} [C_n(u) + \langle \lambda, D_n(u) - V \rangle] \\ &\quad \leq \sup_{\lambda \geq 0} \left[C_n(u^*_{V-\varepsilon_1(n)e}) + \langle \lambda, D_n(u^*_{V-\varepsilon_1(n)e}) - V \rangle \right] \\ &\quad \leq \sup_{\lambda \geq 0} \left[C(u^*_{V-\varepsilon_1(n)e}) + \varepsilon_1(n) + \langle \lambda, D(u^*_{V-\varepsilon_1(n)e}) - (V - \varepsilon_1(n)) \rangle \right] \end{aligned} \tag{13.10}$$

$$= \inf_{u\in\overline{U}} \Big[C(u) + \langle \lambda^*_{V-\varepsilon_1(n)e}, D(u) - (V - \varepsilon_1(n))\rangle\Big] + \varepsilon_1(n) \tag{13.11}$$

$$\leq \inf_{u\in\overline{U}} \Big[C(u) + \langle \lambda^*_{V-\varepsilon_1(n)e}, D(u) - V\rangle\Big] + \varepsilon_1(n)k_1 \tag{13.12}$$

and from (13.11) we have

$$\begin{aligned}
&\inf_{u\in\overline{U}} \sup_{\lambda\geq 0} [C_n(u) + \langle \lambda, D_n(u) - V\rangle] \\
&\quad\leq \inf_{u\in\overline{U}} \Big[C(u) + \langle \lambda^*_{V-\varepsilon_1(n)e}, D(u) - V\rangle\Big] + \varepsilon_1(n)k_1 \qquad (13.13)\\
&\quad\leq \inf_{u\in\overline{U}} \Big[C_n(u) + \varepsilon_1(n) + \langle \lambda^*_{V-\varepsilon_1(n)e}, D_n(u) + \varepsilon_1(n) - V\rangle\Big] + \varepsilon_1(n)k_1 \\
&\quad\leq \inf_{u\in\overline{U}} \Big[C_n(u) + \langle \lambda^*_{V-\varepsilon_1(n)e}, D_n(u) - V\rangle\Big] + 2\varepsilon_1(n)k_1 \\
&\quad\leq \sup_{\lambda\geq 0} \inf_{u\in\overline{U}} [C_n(u) + \langle \lambda, D_n(u) - V\rangle] + 2\varepsilon_1(n)k_1. \qquad (13.14)
\end{aligned}$$

(13.14) implies (13.6). The feasibility of $\mathbf{COP}_n(\overline{U})$ follows from the fact that (13.11) is finite for all n large, as the first term in (13.11) equals the value of $\mathbf{COP}(\overline{U})$ with $V - \varepsilon_1(n)e$ replacing V; the latter is bounded by $\overline{B}$ since (**S**1) implies that for all n large enough, $\mathbf{COP}(\overline{U})$ with $V - \varepsilon_1(n)e$ replacing V is feasible.

The other assertions of the lemma follow from the above inequalities. In particular, (13.8) follows from (13.11), where

$$\lambda^*_n \stackrel{\text{def}}{=} \lambda^*_{V-\varepsilon(n)e}. \tag{13.15}$$

(13.9) follows from (13.10), where $u^*_n \stackrel{\text{def}}{=} u^*_{V-\varepsilon(n)e}$, since (13.10) $\leq$ (13.12) = (13.13) $\leq$ (13.14). □

Proof of Theorem 13.1: Choose some small $\delta_0 > 0$ as in Lemma 13.2. Recall the definition of λ^*_n in (13.8) and (13.15). It follows from Lemma 13.2, the bound (13.7) and (13.13), that for all n large enough,

$$\begin{aligned}
C^{\overline{U}}_n - C^{\overline{U}} &= \sup_{\lambda\geq 0} \inf_{u\in\overline{U}} [C_n(u) + \langle \lambda, D_n(u) - V\rangle] \\
&\quad - \max_{\lambda\geq 0} \min_{u\in\overline{U}} [C(u) + \langle \lambda, D(u) - V\rangle] \\
&\leq C_n(u^*) + \langle \lambda^*_n, D_n(u^*) - V\rangle + \varepsilon_1(n)k_1 \\
&\quad - [C(u^*) + \langle \lambda^*_n, D(u^*) - V\rangle] \\
&\leq \varepsilon_1(n)k_1 + \varepsilon_1(n)(1 + \langle \lambda^*_n, e\rangle) \\
&\leq 2\varepsilon_1(n)k_1.
\end{aligned}$$

We obtain similarly by the same kind of arguments, for all n large enough,

$$C^{\overline{U}} - C^{\overline{U}}_n \leq 2\varepsilon_1(n)k_1,$$

which concludes the proof. □

There are cases where one knows *a priori* that there exists an optimal policy for **COP** within some $\overline{U}$, but $\mathbf{COP}_n$ has optimal policies only within some larger class of policies, say $\overline{U}'$. This is the case, for example, when **COP** corresponds to the expected average cost problem, for which we showed that under fairly general assumptions, there exist optimal stationary policies; if $\mathbf{COP}_n$ corresponds to the problem with finite horizon (of length n, say), then one has to consider the larger class U_M in order to obtain an optimal policy for $\mathbf{COP}_n$. If we chose for both the finite and infinite horizon $\overline{U} = \overline{U}_M$, then condition (**S**3) would typically not hold. If we chose $\overline{U} = U_S$, then we would only get a statement of the type

$$\lim_{n\to\infty} C_n^{\overline{U}} = C_{ea}(\beta),$$

whereas we wish to obtain

$$\lim_{n\to\infty} C_n = C_{ea}(\beta).$$

To handle these cases, the following will be useful:

Theorem 13.2 (*Convergence of values, extensions*)
Assume (**S**1) − (**S**3) *(restricted to the class of policies* $\overline{U}$*). Assume that for any* $\varepsilon > 0$ *and* $\lambda \geq 0$*, there exist an* ε*-optimal policy* u_ε *within the subclass* $\overline{U} \subset \overline{U}'$*, and some integer* N_0 *(both may depend on* λ *and* ε*) for the problem of minimizing over* $u \in \overline{U}'$ *the Lagrangian*

$$C_n(u) + \langle \lambda, D_n(u) \rangle, \qquad \forall n \geq N_0.$$

Then $\lim_{n\to\infty} C_n^{\overline{U}'} = C^{\overline{U}}$.

Proof. According to Theorem 13.1 we have $\lim_{n\to\infty} C_n^{\overline{U}} = C^{\overline{U}}$. Since $C_n^{\overline{U}} \geq C_n^{\overline{U}'}$, we conclude that that

$$\overline{\lim_{n\to\infty}}\, C_n^{\overline{U}'} \leq C^{\overline{U}}.$$

We shall show that

$$\underline{\lim_{n\to\infty}}\, C_n^{\overline{U}'} \geq C^{\overline{U}},$$

which will establish the convergence of the values. Fix some $0 < \delta_0 < \eta(V)$, and consider n sufficiently large so that $\varepsilon_1(n) < \delta_0$, where $\varepsilon_1(n)$ is defined in (**S**3). It follows that

$$\begin{aligned} C_n^{\overline{U}'} - C^{\overline{U}} &= \inf_{u\in\overline{U}'} \sup_{\lambda\geq 0} [C_n(u) + \langle \lambda, D_n(u) - V \rangle] \\ &\quad - \min_{u\in\overline{U}} [C(u) + \langle \lambda^*, D(u) - V \rangle] \\ &\geq [C_n(u_\varepsilon) + \langle \lambda^*, D_n(u_\varepsilon) - V \rangle] - \varepsilon \end{aligned}$$

$$\begin{aligned} & -[C(u_\varepsilon) + \langle \lambda^*, D(u_\varepsilon) - V\rangle] \\ \geq\ & -\varepsilon_1(n)k_1 - \varepsilon \end{aligned}$$

(where k_1 is defined in Lemma 13.2). □

13.3 Approximation and robustness of the policies

Next, we establish the convergence of optimal policies.

Theorem 13.3 (*Convergence of the policies*)
Assume that the values of $\mathbf{COP}_n(\overline{U})$ *converge to the value of* $\mathbf{COP}(\overline{U})$, *i.e.,* $\lim_{n\to\infty} C_n^{\overline{U}} = C^{\overline{U}}$. *Assume that there is some topology on* $\overline{U}$ *such that*

*(***S***4):* $C(\cdot)$ *and* $D^k(\cdot)$, $k = 1, \ldots, K$ *are lower semi-continuous on* $\overline{U}$.

Consider an increasing sequence of integers $m(n)$, $n = 1, 2, \ldots$ *and a sequence* $\varepsilon_2(n)$ *decreasing to zero. Assume that* $\mathbf{COP}_{m(n)}$ *are feasible, and let* $u_n^* \in \overline{U}$ *be some* $\varepsilon_2(n)$*-optimal policies for* $\mathbf{COP}_{m(n)}$, $n = 1, 2, \ldots$. *Assume that* u_n^* *have some accumulation point* $u^* \in \overline{U}$. *Then* u^* *is optimal for* **COP**.

Proof. From the lower semi-continuity of $D(\cdot)$ and from (**S**3), it follows that

$$\begin{aligned} D^k(u^*) &\leq \lim_{n\to\infty} D^k(u_n^*) \\ &\leq \lim_{n\to\infty} [D^k_{m(n)}(u_n^*) + \varepsilon_1(n)] \\ &\leq \lim_{n\to\infty} [V_k - \varepsilon_1(n)] = V_k. \end{aligned}$$

Hence, u^* is feasible. On the other hand, from the lower semi-continuity of $C(\cdot)$, from (**S**3), and since, by assumption, $\lim_{n\to\infty} C_n^{\overline{U}} = C^{\overline{U}}$, it follows that

$$\begin{aligned} C(u^*) &\leq \lim_{n\to\infty} C(u_n^*) \\ &\leq \lim_{n\to\infty} [C_{m(n)}(u_n^*) + \varepsilon_1(n)] \\ &\leq \lim_{n\to\infty} [C^{\overline{U}}_{m(n)} + \varepsilon_2(n) + \varepsilon_1(n)] = C^{\overline{U}}. \end{aligned}$$

Consequently, $C(u^*) = C^{\overline{U}}$ and u^* is optimal, which establishes the proof. □

Finally, we consider the construction of almost optimal policies. We need the following convexity assumption:

- (**S**5): For any $p, 0 < p < 1$ and any policies $u^1 \in \overline{U}, u^2 \in \overline{U}$, there is a policy $u^p \in \overline{U}$ such that

$$D(u^p) \leq pD(u^1) + (1-p)D(u^2),$$

$$C(u^p) \leq pC(u^1) + (1-p)C(u^2).$$

Theorem 13.4 (*Robustness of the policies*)
Assume (**S1**) − (**S3**) *and* (**S5**).
(*i*) *Let* $u^1 = v$, $u^2 = u^*$, *where* u^* *is optimal for* $\mathbf{COP}(\overline{U})$ (*see* (**S2**)) *and* v *is given in* (**S1**). *Then for any* $\varepsilon_4 > 0$, *there exists some* p *such that the policy* u^p *defined in* (**S5**) *is* ε_4*-optimal for* $\mathbf{COP}_n(\overline{U})$, *for all* n *large enough.*
(*ii*) *Consider some sequence* $\varepsilon_3(n)$, $n = 1, 2, \ldots$ *converging to zero. Let* $u^1 = v$, *and consider the sequence of policies* $u_n^2 \in \overline{U}$ *such that* u_n^2 *is* $\varepsilon_3(n)$*-optimal for* $\mathbf{COP}_n(\overline{U})$. *Then for any* $\varepsilon_4 > 0$, *there exists some* p *such that the policies* $u^p(n)$ *defined in* (**S5**) *when considering the pairs* (u^1, u_n^2) *are* ε_4*-optimal for* $\mathbf{COP}(\overline{U})$, *for all* n *large enough.*

Proof. We first show that for any $p > 0$, u^p is feasible for all n large enough.

$$\begin{aligned} D_n(u^p) &\leq D(u^p) + \varepsilon_1(n)e \\ &\leq pD(v) + (1-p)D(u^*) + \varepsilon_1(n)e \\ &\leq V - p[V - D(v)] + \varepsilon_1(n)e. \end{aligned}$$

So, for all n for which $p[V - D(v)] + \varepsilon_1(n)e \leq 0$, u^p is feasible. Similarly,

$$\begin{aligned} C_n(u^p) &\leq C(u^p) + \varepsilon_1(n) \\ &\leq pC(v) + (1-p)C^{\overline{U}} + \varepsilon_1(n) \\ &\leq 2p\overline{B} + C^{\overline{U}} + \varepsilon_1(n) \\ &\leq C_n^{\overline{U}} + 2p\overline{B} + [C^{\overline{U}} - C_n^{\overline{U}}] + \varepsilon_1(n). \end{aligned}$$

(i) now follows since $C^{\overline{U}} - C_n^{\overline{U}} + \varepsilon_1(n)$ tends to zero (by Theorem 13.1 and by (**S3**)).
(ii) is obtained similarly. For any n,

$$\begin{aligned} D(u^p(n)) &\leq pD(v) + (1-p)D(u_n^2) \\ &\leq pD(v) + (1-p)D_n(u_n^2) + \varepsilon_1(n)e \\ &\leq V - p[V - D(v)] + \varepsilon_1(n)e \end{aligned}$$

and hence for any p, $u^p(n)$ are feasible for all large enough n.

$$\begin{aligned} C(u^p(n)) &\leq pC(v) + (1-p)C(u_n^2) \\ &\leq 2p\overline{B} + C_n(u_n^2) + \varepsilon_1(n) \\ &\leq 2p\overline{B} + C_n^{\overline{U}} + \varepsilon_3(n) + \varepsilon_1(n). \end{aligned}$$

(ii) now follows since $C_n^{\overline{U}} + \varepsilon_3(n) + \varepsilon_1(n)$ tends to $C^{\overline{U}}$ (by Theorem 13.1 and by definition of $\varepsilon_1(n)$ and $\varepsilon_3(n)$). □

Remark 13.3 (*Relaxing some assumptions*)
The results of Theorem 13.4 (i) clearly extend to the setting of Theorem

13.2. This follows from the fact that u^p is ε_4-optimal for $\mathbf{COP}_n(\overline{U})$, for all n large enough, and since the class $\overline{U}$ has an ε-optimal policy for $\mathbf{COP}_n(\overline{U}')$, for all n large enough.

CHAPTER 14

Convergence of discounted constrained MDPs

We apply below the results of Chapter 13, to the convergence of CMDPs in the discount factor.

14.1 Convergence in the discount factor

We first consider the four types of convergence described in Chapter 13, where the limit **COP** is the one with infinite horizon discounted cost, with discount factor $\alpha < 1$, and where $\mathbf{COP}_n$ are with infinite horizon discounted cost with discount factor α_n converging to α. The transition probabilities and immediate costs are the same. The convergence results were already obtained in Altman (1993) using other general convergence theorems (that did not provide the estimation of the error in approximation, as we have here).

We make a weak contracting assumption: the immediate costs are μ-bounded by $\overline{b}$, the transition probabilities are μ-continuous, the initial distribution satisfies $\langle \beta, \mu \rangle < \infty$, and

$$\alpha \sum_{y \notin \mathcal{M}_\alpha} \mathcal{P}_{xay} \mu(y) \leq \xi \mu(x), \xi \in [0, 1). \tag{14.1}$$

It follows then by Theorem 8.4 (iii) and the end of Chapter 10 that one may restrict oneself without loss of optimality to stationary policies, since they are sufficient for both the limiting and the approximating problems. Hence we may consider $\overline{U}$ in the key theorems of Chapter 13 to be the stationary policies.

We assume that the Slater condition holds, i.e., $D_\alpha(\beta, u) < V$ for some policy u, which implies condition (**S**1).

We check all conditions (**S**2)-(**S**5). (**S**2) is established in Corollary 9.2 and Theorem 9.10; Lemma 8.5 (ii) implies (**S**4). (**S**5) follows from Theorem 9.8 (ii). For any discount factor α_1 such that $\alpha_1 < \alpha/\xi$ (where ξ is defined in (14.1),

$$\|C_{\alpha_1}(\bullet, u) - C_\alpha(\bullet, u)\|_\mu$$

$$
\begin{aligned}
&= \left\| \sum_{j=0}^{\infty} \left[(1-\alpha_1)\alpha_1^j - (1-\alpha)\alpha^j \right] P^j(u)c(u) \right\|_\mu \\
&\leq \sum_{j=0}^{\infty} \left| (1-\alpha_1)\alpha_1^j - (1-\alpha)\alpha^j \right| \left(\frac{\xi}{\alpha} \right)^j \overline{b} \\
&\leq \overline{b} \sum_{j=0}^{\infty} \left| \alpha_1^j - \alpha^j \right| \left(\frac{\xi}{\alpha} \right)^j = \overline{b} \left| \frac{1}{1-\xi} - \frac{1}{1-\xi\alpha^{-1}\alpha_1} \right| =: \varepsilon_1(\alpha_1, \alpha)
\end{aligned}
$$

($\overline{b}$ is defined in (7.36), $P^j(u)$ is the j-step transition probability matrix under the stationary policy u, and $c(u)$ is the vector whose components are $c(x,u)$.) This converges to 0 as α_1 converges to α, uniformly in the policies. This establishes (**S**3) (from Chapter 13). Using Theorem 13.1, we have that the difference between $C_{\alpha_1}(\beta)$ and $C_\alpha(\beta)$ is of order $\varepsilon_1(\alpha_1, \alpha)$.

Remark 14.1 (*Alternative conditions*)
If the immediate costs are bounded, or if the following holds:

$$
\sup_{u \in U_S} \sup_{j \in \mathbb{N}} |\beta P^j(u)c(u)| \leq \tilde{b} < \infty, \tag{14.2}
$$

then (**S**3) is even simpler to establish. Indeed, we have

$$
\begin{aligned}
&C_{\alpha_1}(\beta, u) - C_\alpha(\beta, u) \\
&\quad \leq \tilde{b} \sum_{j=0}^{\infty} \left| (1-\alpha)\alpha^j - (1-\alpha_1)\alpha_1^j \right| \\
&\quad \leq \tilde{b} \sum_{j=0}^{\infty} \left| \alpha^j - \alpha_1^j \right| = \tilde{b} \left| \frac{1}{1-\alpha} - \frac{1}{1-\alpha_1} \right|.
\end{aligned}
$$

(Note, however, that (14.2) does not imply our Assumption (14.1).)

14.2 Convergence to the expected average cost

We consider the four types of convergence where the limit **COP** is the one with *infinite horizon expected average cost*, and where $\mathbf{COP}_n$ are with infinite horizon discounted cost with discount factor α_n converging to 1. The transition probabilities and immediate costs are the same.

We consider again the contracting framework; we assume in particular that the MDP is uniform μ-geometric recurrent (Definition 11.5). Finally, we make the unichain assumption (11.2) (from Chapter 11).

It follows then by Theorem 8.4 (iii) and 11.6 that one may restrict oneself without loss of optimality to stationary policies, since they are sufficient for both the limiting and the approximating problems. Hence we may consider $\overline{U}$ in the key theorems of Chapter 13 to be the stationary policies.

(**S**1) holds when assuming the standard Slater condition. (**S**2) is established in Corollary 12.2 and Theorem 12.8; Lemma 11.3 (ii) implies (**S**4). (**S**5) follows from Theorem 12.6 (ii). It remains to establish (**S**3). We prove it for C_{ea}; the same proof holds for D^k_{ea}. Fix an arbitrary stationary policy u, and let $\Pi(u)$ denote the matrix whose rows are all equal to the steady-state probability distribution $\pi(u)$. Recall that our uniformly μ-geometric recurrence assumption implies uniform μ-ergodicity (Theorem 11.11), i.e., there exist constants $\sigma > 0$ and $\tilde{\xi} < 1$ such that for all $u \in U_S$,

$$\begin{cases} \|P^n(u) - \Pi(u)\|_\mu \le \sigma\tilde{\xi}^n, & \forall n \in \mathbb{N}, \\ \|P(u)\|_\mu \le \sigma. \end{cases}$$

Hence

$$\begin{aligned} &\|C_\alpha(\bullet, u) - C_{ea}(\bullet, u)\|_\mu \\ &= \left\| \left[\sum_{j=0}^{\infty} (1-\alpha)\alpha^j P^j(u) - \Pi(u) \right] c(u) \right\|_\mu \\ &= \left\| \sum_{j=0}^{\infty} (1-\alpha)\alpha^j \left[P^j(u) - \Pi(u) \right] c(u) \right\|_\mu \\ &\le \sigma \sum_{j=0}^{\infty} (1-\alpha)\alpha^j \overline{b}\tilde{\xi}^j = \frac{\sigma\overline{b}(1-\alpha)}{1-\alpha\tilde{\xi}} =: \varepsilon_1(\alpha). \end{aligned}$$

Using Theorem 13.1, we have that the difference between $C_\alpha(\beta)$ and $C_{ea}(\beta)$ is of order $\varepsilon_1(\alpha)$.

Remark 14.2 (*The multi-chain case*)
It is possible to obtain similar results for the general multi-chain case under appropriate conditions. This was done for the finite state and actions case in Tidball and Altman (1995). The class of policies $\overline{U}$ they consider is $\mathcal{U}$, which is a dominating class for the multi-chain case. It used the fact that the optimal policies and values of **COP** are obtained by $\mathbf{LP_4}(\beta)$(see Feinberg, 1995) even in the multi-chain case.

14.3 The case of uniform Lyapunov function

Next, we relax the assumption on uniform μ-geometric recurrence, used in the previous section, at the price of losing the explicit estimation on the error in the approximation. We assume that the MDP has uniform Lyapunov function (corresponding to the set $\mathcal{M} = \{0\}$, where 0 is some arbitrary state), and that the immediate costs are ν-bounded.

Since (B2) and (B3) are equivalent to the existence of a uniform Lyapunov function (Section 11.9), then by Theorem 11.6, we may restrict ourselves to stationary policies for the limit problem. The uniform Lyapunov

condition implies that for each discount factor α, the weak contracting condition (10.4) holds for the corresponding discounted cost problem (see Section 10.4). Thus, we may restrict ourselves to stationary policies also for the approximating problems (this is due to Theorem 8.4 (iii) and to the fact that the discounted cost problem is equivalent to a contracting total cost problem).

Condition (**S**1) (from Chapter 13) holds when assuming the standard Slater condition. (**S**2) is established in Corollary 12.2 and Theorem 12.8. Lemma 11.3 implies (**S**4). (**S**5) follows from Theorem 12.6 (ii). It remains to establish (**S**3). We prove it in the next Lemma for C_{ea}; the same proof holds for D^k_{ea}.

Lemma 14.1 (*Uniform convergence of the discounted cost to the expected average cost*)
Consider an MDP with a uniform Lyapunov function. Then $C_\alpha(\beta,u)$ converges to $C_{ea}(\beta,u)$ uniformly over U_S.

Proof. Let $T = T_0$ be the time to hit state 0 (see definition in (6.3)). The discounted cost satisfies the following for $u \in U_S$:

$$C_\alpha(\beta,u) = (1-\alpha)E^u_\beta \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t,A_t) + E^u_\beta \alpha^{T-1} C_\alpha(0,u).$$

Moreover,

$$C_\alpha(0,u) \quad = \quad \frac{E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t,A_t)}{E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1}}$$

(see (12.19)). On the other hand,

$$C_{ea}(\beta,u) = \frac{E^u_0 \sum_{t=1}^{T} c(X_t,A_t)}{E^u_0 T}.$$

Hence

$$|C_{ea}(\beta,u) - C_\alpha(\beta,u)| \le \left|(1-\alpha)E^u_\beta \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t,A_t)\right| \tag{14.3}$$
$$+ \left| E^u_\beta \alpha^{T-1} \frac{E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t,A_t)}{E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1}} - \frac{E^u_0 \sum_{t=1}^{T-1} c(X_t,A_t)}{E^u_0 T - 1} \right|.$$

We shall show that this converges to 0 uniformly in U_S.

We note that

$$\left| E^u_\beta \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t,A_t) \right| \le E^u_\beta \sum_{t=1}^{T-1} \nu(X_t,A_t) \le \langle \beta, \mu \rangle < \infty$$

(the second inequality follows from Lemma 7.5 (ii)). Hence the first term

on the right-hand side of (14.3) converges to 0 as $\alpha \to 1$, uniformly over U_S.

Now,

$$|1 - E^u_\beta \alpha^{T-1}| = (1-\alpha)E^u_\beta \sum_{t=1}^{T-1} \alpha^{t-1} \le (1-\alpha)E^u_\beta T - 1 < (1-\alpha)\langle \beta, u\rangle$$

(this too follows from Lemma 7.5 (ii)). This implies that $E^u_\beta \alpha^{T-1}$ converges to 1 uniformly in u, as $\alpha \to 1$.

Next, we evaluate $E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1} - E^u_0(T-1)$. Fix some integer n.

$$\begin{aligned}
&\left| E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1} - E^u_0(T-1) \right| \\
&\quad \le \quad E^u_0 \sum_{t=1}^{T-1} (1-\alpha^{t-1}) 1\{T \le n\} + 2E^u_0 1\{T > n\} \\
&\quad \le \quad \sum_{t=1}^{n} (1-\alpha^{t-1}) + 2E^u_0 1\{T > n\}
\end{aligned}$$

The term $E^u_0 1\{T > n\}$ can be made arbitrarily small uniformly over $u \in U_S$, by choosing n sufficiently large. This is due to property M1′ (in Section 7.4) and to Lemma 7.5 (i). For that n, $\sum_{t=1}^{n}(1-\alpha^{t-1})$ can be made arbitrarily small in a neighborhood $[\alpha(n), 1]$ by choosing $\alpha(n)$ sufficiently close to 1. This implies the convergence of $E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1}$ to $E^u_0 T - 1$, uniformly in U_S.

By the same type of arguments, we show that $E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t)$ converges to $E^u_0 \sum_{t=1}^{T-1} c(X_t, A_t)$, uniformly in U_S. Indeed,

$$\begin{aligned}
&\left| E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t) - E^u_0 \sum_{t=1}^{T-1} c(X_t, A_t) \right| \\
&\quad \le \quad E^u_0 \sum_{t=1}^{T-1} (1-\alpha^{t-1}) \nu(X_s, A_s) 1\{T \le n\} + 2E^u_0 \hat{M}(X_n) 1\{T > n\} \\
&\quad \le \quad (1-\alpha^{n-1}) \sum_{t=1}^{T-1} \nu(X_t, A_t) + 2E^u_0 \hat{M}(X_n) 1\{T > n\} \\
&\quad \le \quad (1-\alpha^{n-1})\mu(0) + 2E^u_0 \hat{M}(X_n) 1\{T > n\}.
\end{aligned}$$

By choosing n sufficiently large, the second term in the above expression can be made arbitrarily small; this follows again from property M1′ (in Section 7.4) and from Lemma 7.5 (i). For that n, $(1-\alpha^{n-1})\mu(0)$ can be made arbitrarily small in a neighborhood $[\alpha(n), 1]$ by choosing $\alpha(n)$ sufficiently close to 1. We have thus shown that $E^u_0 \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t)$ converges to $E^u_0 \sum_{t=1}^{T-1} c(X_t, A_t)$, uniformly in U_S.

The three last paragraphs now imply the convergence of the second term on the right-hand side of (14.3) to 0, uniformly in $u \in U_S$. This concludes the proof of the lemma. □

CHAPTER 15

Convergence as the horizon tends to infinity

We use below the tools developed in Chapter 13 to study the convergence of finite horizon CMDPs to infinite horizon CMDPs.

15.1 The discounted cost

We consider all four types of convergence described in Chapter 13 where the limit **COP** is the one with infinite horizon discounted cost, with discount factor $\alpha < 1$, and where $\mathbf{COP}_n$ are with horizon of length n, and discounted with the same discount factor α. The transition probabilities and immediate costs are the same. We assume that the weak contracting framework (see Section 10.4) holds, and in particular (10.4): there exists some scalar $\xi \in [0,1)$, a vector $\mu : \mathbf{X}_\alpha \to [1,\infty)$, and a finite set $\mathcal{M}_\alpha$, such that for all $x \in \mathbf{X}, a \in \mathsf{A}$, $\alpha \sum_{y \notin \mathcal{M}_\alpha} \mathcal{P}^\alpha_{xay} \mu(y) \leq \xi \mu(x)$, and the immediate costs are μ-bounded by a constant $\overline{b}$.

One may restrict oneself to Markov policies, since they are sufficient for both the limiting and the approximating problems (see Theorem 6.1). In order to apply below the key theorems, we shall thus consider $\overline{U}$ to be the set of Markov policies.

Remark 15.1 (*Almost optimal stationary policies*)
Note that **COP** has optimal stationary policies (Theorem 8.4 (iii)). One can then show by using Theorem 13.4 that for any $\varepsilon > 0$, there exists some *stationary* u^p (which depends only on ε, not on n) that is ε-optimal for $\mathbf{COP}_n$ for all n large enough.

Conditions (**S**1),(**S**2),(**S**4) and (**S**5) from Chapter 13 were established in Chapter 14. (**S**3) follows from Lemma 8.5 (i). We compute a bound on the approximation error. We have for any Markov policy u,

$$\begin{aligned} &\left\| C^T_\alpha(\beta,u) - C_\alpha(\beta,u) \right\|_\mu \\ &\quad \leq \quad (1-\alpha) \left\| \sum_{t=T}^{\infty} \alpha^t P(u_1)P(u_2)\cdots P(u_t) \right\|_\mu \overline{b} \\ &\quad \leq \quad \frac{(1-\alpha)\xi^{T+1}\overline{b}}{1-\xi} =: \varepsilon_1(T). \end{aligned}$$

Using Theorem 13.1, we have that the difference between $C_\alpha^T(\beta)$ and $C_\alpha(\beta)$ is of order $\varepsilon_1(T)$.

15.2 The expected average cost: stationary policies

This problem is more involved than the previous ones; if we considered the class of Markov policies as candidates for $\overline{U}$, then property (**S**3) is not satisfied. On the other hand, a smaller class of policies might not be dominating for the finite horizon problem. We therefore use the approach of Theorem 13.2 and Remark 13.3.

We consider the four types of convergence where the limit **COP** is the one with infinite horizon expected average cost, and where $\mathbf{COP}_n$ are with finite horizon expected average cost. The transition probabilities and immediate costs are the same.

We consider the case of uniform Lyapunov function (where $\mathcal{M} = \{0\}$, and 0 is an arbitrary state).

It follows by Theorem 11.6 that one may restrict oneself to stationary policies for the limiting case **COP**, and thus we take $\overline{U} = U_S$. For the finite horizon case $\mathbf{COP}_n$, we may choose $\overline{U} = U_M$.

We show first that the four types of convergence, given in Theorems 13.1, 13.3 and 13.4, hold when restricting ourselves to U_S. In other words, we show that the optimal values $C_{ea}^{n\ U_S}$ converge to $C_{ea}(\beta)$, that is, the values of the finite horizon problems restricted to the (non-dominating class of) stationary policies converge to the optimal value of the infinite horizon problem. (This does not mean *a priori* that the values converge without the above restriction.) In particular, we can obtain an optimal policy for the expected average cost as the appropriate limit of stationary policies that are almost optimal for the (restricted) finite horizon case.

Conditions (**S**1),(**S**2),(**S**4) and (**S**5) of Chapter 13 were established in Section 14.3. It remains to establish (**S**3). We shall prove it for C_{ea}; the same proof holds for D_{ea}^k.

Before presenting the proof, we illustrate that (**S**3) holds under the stronger contracting framework: the cost is assumed to be μ-bounded (7.36), the transition probabilities are μ-continuous (Assumption (7.35)), and the initial distribution satisfies $\langle \beta, \mu \rangle < \infty$; the MDP is assumed to be uniformly μ-geometric ergodic (see Definition 11.5). Finally, we make the unichain assumption (11.2) (from Chapter 11).

Fix an arbitrary stationary policy w, and let $\Pi(w)$ denote the matrix whose rows are all equal to $\pi(w)$.

$$C_{ea}^n(x,w) \quad = \quad \sum_{y \in \mathbf{X}} \frac{1}{n} \sum_{t=1}^{n} [P^t(w)]_{xy} c(y,w),$$

$$C_{ea}(x,w) \quad = \quad \langle \pi(w), c(w) \rangle = \sum_{y \in \mathbf{X}} n^{-1} \sum_{t=1}^{n} \pi_y(w) c(y,w)$$

so that

$$\begin{aligned} \|C^n_{ea}(\cdot,w) - C_{ea}(\cdot,w)\|_\mu &\leq n^{-1} \sum_{t=1}^{n} \left\| P^t(w) - \Pi(w) \right\|_\mu \|c(u)\|_\mu \\ &\leq \frac{\sigma \overline{b} \sum_{t=1}^{n} \tilde{\xi}^t}{n} \leq \frac{\sigma \overline{b}}{n(1-\tilde{\xi})}. \end{aligned}$$

(σ and $\tilde{\xi} \in [0,1)$ are given in the Definition 11.5.) This establishes (**S**3) for the more restrictive contracting framework. (**S**3) also holds for the more general case of uniform Lyapunov function. We omit the proof, which can be found in Theorem 4.2 in Cavazos-Cadena (1992).

We thus established the convergence of the finite horizon CMDP restricted to stationary policies, to the infinite horizon one. Moreover, it follows from Theorem 13.1 that the rate of convergence of the values is of the order of n^{-1}.

15.3 The expected average cost: general policies

Next, we consider the problem of the convergence of $\mathbf{COP}_n$ to **COP**, as in the previous section, but without the restriction to stationary policies. The proof of the theorem below is based on an extension of Lemma 1 in Altman and Gaitsgory (1995).

Theorem 15.1 (*Convergence of the finite horizon problem to the infinite horizon*)
Consider an MDP with a uniform Lyapunov function. Assume that the Slater condition holds, i.e., for some stationary policy, $D_{ea}(\beta,u) < V$. Then
(*i*) *The value of the finite horizon problem converges to the value of the infinite horizon one.*
(*ii*) *There exists a stationary policy u_ε which is ε-optimal for the finite horizon constrained MDP, for all horizons n sufficiently large.*

Proof. We shall use Theorem 13.2. We need to show that for any non-negative λ and ε, there exists an ε-optimal stationary policy w (that may depend on λ and ε) for the Lagrangian

$$J^{\lambda,n}_{ea}(\beta,u) := C^n_{ea}(\beta,u) + \langle \lambda, D^n_{ea}(\beta,u) \rangle = \frac{1}{n} \sum_{t=1}^{n} E^u_\beta j^\lambda(X_t, A_t),$$

for all n sufficiently large, where

$$j^\lambda(\cdot,\cdot) := c(\cdot,\cdot) + \langle \lambda, d(\cdot,\cdot) \rangle.$$

(We thus set $\overline{U} = U_S$ and $\overline{U}' = U_M$ in Theorem 13.2. The fact that U_M is a dominating class for the finite horizon problem follows from Theorem 6.1.) Denote the value of the above minimization by $J_{ea}^{\lambda,n}(\beta)$, and let J_{ea}^{λ} be the value corresponding to the expected average cost (with infinite horizon, which thus does not depend on β). For simplicity, we shall omit λ from the notation below.

Let $j_0 \in \mathsf{F}^\mu$ denote some terminal cost, and consider the problem of minimizing the total expected cost during a horizon of n step:

$$J^n(\beta, u, j_0) = \sum_{t=1}^{n} E_\beta^u j(X_t, A_t) + E_\beta^u j_0(X_{n+1}). \tag{15.1}$$

Denote the value of this problem by $J^n(\beta, j_0)$. We shall use the following (see e.g., Puterman, 1994):

Lemma 15.1 (*Computing the optimal value and policy for a finite horizon problem*)
(*i*) $J^n(\beta, j_0)$ *is given by the recursive solution of*

$$\begin{aligned} J^0(x, j_0) &= j_0(x), \\ J^{t+1}(x, j_0) &= \min_{a \in \mathsf{A}(x)} \left\{ j(x,a) + \sum_{y \in \mathbf{Y}} \mathcal{P}_{xay} J^t(x, j_0) \right\} \end{aligned} \tag{15.2}$$

for all $x \in \mathbf{X}$ *(*$J^t(\beta, j_0)$ *is then given by* $\langle \beta, J^t(\cdot, j_0) \rangle$*).*
(*ii*) *Consider the Markov policy* $g = (g^n, g^{n-1}, \ldots, g^1)$*, where* g_i *attains the minimum in (15.2) for* $t = i - 1$*. Then* g *is optimal.*

(Note that any finite horizon problem can be transformed into an infinite horizon problem by incorporating the time into the state space, see e.g., Tidball and Altman, 1995. One can then use Theorem 9.1 to show that the recursive equations above indeed yield the optimal value.)

Let (J, h) be solutions of ACOE (12.6) where c is replaced by j, and such that J and $h \in \mathsf{F}^\mu$ are obtained as the limits (12.9)-(12.10) (the fact that these limits are indeed solutions of ACOE was established in Theorem 12.2). Let g^* be a stationary optimal policy achieving the min in ACOE. Define

$$j_0(x) := h(x).$$

It follows from (15.1) that

$$|J^n(x, 0) - J^n(x, j_0)| \leq |j_0(x)|. \tag{15.3}$$

We now compute $J^n(\beta, j_0)$. By (15.2), we have

$$\begin{aligned} J^0(x, j_0) &= j_0(x), \\ J^1(x, j_0) &= \min_{a \in \mathsf{A}(x)} \left\{ j(x,a) + \sum_{y \in \mathbf{Y}} \mathcal{P}_{xay} J^0(x, j_0) \right\} \end{aligned}$$

$$= \min_{a \in \mathsf{A}(x)} \left\{ j(x,a) + \sum_{y \in \mathbf{Y}} \mathcal{P}_{xay} h(x) \right\}$$
$$= h(x) + J_{ea},$$

where the last equality follows from (12.6). Moreover, the Markov policy $g_1 = g^*$ is optimal. We may now continue recursively and obtain

$$J^n(x, j_0) = h(x) + nJ_{ea};$$

moreover, the Markov policy $g = (g^*, \ldots, g^*)$ is optimal, i.e., for all n,

$$J^n(x, j_0) = J^n(x, g^*, j_0) = J^n(x, g^*, 0) + j_0(x). \tag{15.4}$$

Combining (15.3) with (15.4) we get

$$\frac{|J^n_{ea}(\beta) - J^n_{ea}(\beta, g^*)|}{n} \leq \frac{\langle \beta, h \rangle}{n}.$$

(The latter indeed converges to 0 as n goes to infinity, since it follows from (12.20) that h is μ bounded; thus $\langle \beta, h \rangle$ is finite.) Hence, the stationary policy g^* is ε optimal for the problem of minimizing $J^n_{ea}(\beta, u)$ for all n larger than $\varepsilon |\langle \beta, h \rangle^{-1}|$. This establishes the conditions of Theorem 13.2 from which statement (i) follows. Statement (ii) follows by combining (i) with the first part of the section. □

Theorem 13.2 can also be used in order to establish the convergence of the horizon for the general multi-chain case, under suitable conditions. In particular, for the case of finite states and actions, one may consider in Theorem 13.2 $\overline{U} = \mathcal{U}$. Indeed, it is known that in this class (and in particular, within U_D) there exist ε-optimal stationary policies for all horizon n sufficiently large. Moreover, it can be shown that the approximation error is of the order of n^{-1}. (This follows from Federgruen, 1979.) The fact that (**S**3) holds follows since there is only a finite number of elements in U_D. (**S**2) follows since, when restricting ourselves to $u \in \mathcal{U}$, the performance measures are linear in u, and obtained as a finite linear program (see Tidball and Altman, 1996).

CHAPTER 16

State truncation and approximation

In this chapter we consider several schemes for replacing a problem involving an infinite state space with problems with finitely many states (Schemes I and II), or with a problem in which decisions are taken only in finitely many states (Scheme III). We are then interested in the convergence of the optimal values and policies of the truncated problems to those of the original one, as well as the robustness of optimal policies (or, as we already know, of some modifications of optimal policies). The results of this chapter are useful in two situations.

In the first, we might want to solve a constrained MDP with a countable set of states. The way to do this is via an LP with an infinite set of decision variables. The truncation techniques in this chapter will allow us to use a finite state approximation of the original problem, which can be solved by an LP with finitely many decision variables.

As a second application, consider constrained MDPs with a very large state space, for which an LP solution may be too costly. In some special cases one may extend in a natural way the finite problem to an infinite problem; the latter may possess some special structure, which enables us to solve it with some simple techniques other than those involving infinite LPs. The solution for the extended problem can then serve to approximate the original finite one. Examples of this type are presented in Altman (1993, 1994).

We shall use throughout the chapter the contracting framework (see Definition 7.9 for the total cost, and Definition 11.4 for the expected average cost). (We have presented a different approach and results for the non-contracting framework, for non-negative immediate cost, see Remark 9.4 for the non-constrained case, and Section 9.6 for the constrained case.)

The theory of state truncation (as well as other state approximation schemes, such as discretization) in MDPs is a very active area of research, even in the non-constrained case. Some of the important references in this area are Whitt (1978), White (1980, 1982), Hernández-Lerma (1986, 1989), Cavazos-Cadena (1986), Thomas and Stengos (1985) and Sennott (1997). The case of more than one controller was investigated by Nowak (1985), Whitt (1980), Tidball and Altman (1996a) and Tidball *et al.* (1997). Altman (1993, 1994) presented state-truncation techniques for the constrained MDPs, and the schemes presented in this chapter are extensions of those. Our approach is based on the sensitivity analysis tools developed in Chap-

ter 13, which allows us to obtain not only convergence results but also an estimation of the approximation errors.

In the first two approximating schemes which we present, we modify the 'limit' CMDP in the following way. We consider an increasing set of states $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \ldots$ converging to $\mathbf{X}$ or to a strict subset of $\mathbf{X}$. The nth CMDP ($\mathbf{COP}_n$) is restricted to the set $\mathbf{X}_n$ (states not in $\mathbf{X}_n$ will not be reachable from states within $\mathbf{X}_n$, and will thus be of no of interest). In $\mathbf{COP}_n$, we modify the transition probabilities so as to eliminate all transitions outside the set $\mathbf{X}_n$. The two schemes will differ by the way that such transitions will be replaced.

16.1 The approximating sets of states

The sets $\mathbf{X}_n$ may or may not be given *a priori*. In some problems, the following 'finite neighbors' property may hold: from any $x \in \mathbf{X}$, only finitely many states are reachable in one step. In other words,

$$\text{from any } x \in \mathbf{X},\ \{y : \mathcal{P}_{xay} > 0 \text{ for some } a\} \text{ is finite.} \tag{16.1}$$

(This property holds in particular in many queueing applications.) When it holds, we may construct the sets $\mathbf{X}_n$ as follows.

Definition 16.1 (*n-step reachable sets*)
Let $\mathcal{X}$ be a finite given set (in which we would like to approximate the values and policies), and set $\mathbf{Y}(x) = \{y : \mathcal{P}_{xay} > 0 \text{ for some } a\}$. Then we define $\mathbf{X}_n$ as follows:

$$\mathbf{X}_0 = \mathcal{X}, \qquad \mathbf{X}_{n+1} = \bigcup_{x \in \mathbf{X}_n} \mathbf{Y}(x) \bigcup \mathbf{X}_n. \tag{16.2}$$

$\mathbf{X}_n$ is the set of states reachable in n-steps from $\mathcal{X}$.

The above construction may be useful especially when the set of neighbors of a 'typical' state is not too large. When it is large, then the sets $\mathbf{X}_n$ grow very rapidly, which suggests that obtaining good estimates of optimal value and policies might require an unacceptably high complexity of computations. We thus present an alternative, more general way of constructing finite sets $\mathbf{X}_n$ (even when (16.1) does not hold).

We define a parameterized family $\{\mathbf{X}_n(\varepsilon)\}$, where ε is a positive real number, as follows.

Definition 16.2 (ε-*neighboring sets*)
Define $\mathbf{X}_0(\varepsilon) = \mathcal{X}$ where, again, $\mathcal{X}$ is a given set (in which we would like to approximate the values and policies). $\{\mathbf{X}_n(\varepsilon)\}$ are then chosen to be an arbitrary increasing sequence of finite sets of states that satisfies the following. If for some $l > 0$, say $l = \hat{l}$,

$$\sup_{x \in \mathbf{X}_l(\varepsilon)} \sup_{a \in \mathsf{A}(x)} \sum_{y \notin \mathbf{X}_l(\varepsilon)} \mathcal{P}_{xay} \leq \varepsilon, \tag{16.3}$$

then $\mathbf{X}_n(\varepsilon) = \mathbf{X}_{\hat{l}}$ for all $n > \hat{l}$. Otherwise, $\mathbf{X}_{l+1}(\varepsilon)$ is chosen such that

$$\sup_{x \in \mathbf{X}_l(\varepsilon)} \sup_{a \in \mathsf{A}(x)} \sum_{y \notin \mathbf{X}_{l+1}(\varepsilon)} \mathcal{P}_{xay} \leq \varepsilon. \tag{16.4}$$

In other words, we replace neighboring sets in the previous scheme (16.2) by some 'ε-neighboring sets'; in (16.2), the probability under any policy to go from a state in $\mathbf{X}_n$ to a state which is not in $\mathbf{X}_{n+1}$ is zero. In (16.3) and (16.4), it is less than ε instead. One could also consider weighted versions of (16.3) and (16.4), where $\mathcal{P}_{xay}$ are replaced by $\mathcal{P}_{xay}\mu(y)$.

Next, we consider the case where the sets $\mathbf{X}_n$ are given *a priori*. In that case, we shall be interested in identifying an increasing sequence $m^n(\varepsilon)$, such that the nth step of the approximation will yield an error of the order of $\max(\varepsilon, \xi^n)$, provided that we solve the MDP on the truncated set $\mathbf{X}_{m^n(\varepsilon)}$ (ξ is the contraction factor defined in Definition 7.9).

To that end we begin by defining

$$\delta(r, n) := \sup_{\substack{x \in \mathbf{X}_r \\ a \in \mathsf{A}(x)}} \sum_{y \notin \mathbf{X}_n} \mathcal{P}_{xay}\mu(y).$$

$\delta(r, n)$ is a measure of the error that truncation of $\mathbf{X}$ to $\mathbf{X}_n$ induces in states in $\mathbf{X}_r$. We call it the *induced error index*.

Claim: Due to the contracting assumption (7.34), the following holds

$$\lim_{n \to \infty} \delta(r, n) = 0, \qquad \forall r \in \mathbb{N} \tag{16.5}$$

if $\mathbf{X}_n$ are finite sets for all n.

We leave the proof of the claim to the end of the section.

We use an idea introduced by Cavazos-Cadena (1986) and further developed by Tidball and Altman (1996a). Fix ε arbitrarily small, and define the sequence g_k in the following way. $g_0 = \min\{m : \mathcal{X} \subset \mathbf{X}_m\}$ and recursively,

$$g_k = g(\varepsilon, g_{k-1}), \qquad \text{where} \qquad g(\varepsilon, r) = \min\{m : \delta(r, m) \leq \varepsilon\}. \tag{16.6}$$

Due to Assumption (16.5), this sequence is well defined, and for all k, g_k is finite. $g(\varepsilon, r)$ can be interpreted as follows. If we truncate $\mathbf{X}$ to $\mathbf{X}_m$, where $m \geq g(\varepsilon, r)$, then the induced error index $\delta(r, n)$ for the set $\mathbf{X}_r$ is less than ε. Thus the impact on the total approximation error is of the order of ε. Still, we need to approximate the value of x inside $\mathbf{X}_{g_1}$. This leads us to consider the new set $\mathbf{X}_{g_2}$, etc.

Note that g_1 *may be smaller than* g_0; moreover, we could even have $\mathbf{X}_{g_1} \subset \mathcal{X}$. For example, let $\mathbf{X} = \mathbb{N}$, the set of natural numbers, $\mathbf{X}_n = \{1, 2, \ldots, n\}$ and $\mathcal{X} = \{10\}$. Assume that $\mathcal{P}_{x,a,1} = 1, \forall a$. Then $g_0 = 10$ and $g_1 = 1$. This phenomenon motivates the definition

$$m^k(\varepsilon) = \max\{g_m, m = 0, 1, \ldots, k\}.$$

In the special case that $\mathbf{X}_n$ are given by the μ-weighted versions of (16.3) and (16.4), we get $g^0 = 0$, $g^k = k$ and $m_k(\varepsilon) = g^k = k$ for $k \leq \hat{l}$.

In the next sections, we shall include explicitly the set $\mathcal{X}$ in the notation of the μ-norm:

$$\|q\|_\mu^{\mathcal{X}} = \sup_{x \in \mathcal{X}} \frac{q(x)}{\mu(x)}.$$

Our aim in the approximation schemes below is to obtain convergence of the values and policies. Moreover, let $\mathcal{X}$ be a given finite subset of $\mathbf{X}$. We wish to obtain an estimate of the approximation errors for initial distributions having their support in $\mathcal{X}$.

We conclude the section with the proof of the Claim made earlier in this section.

Proof of Claim: Assume that (16.5) does not hold. Then there exists some $b > 0$ such that for some x,

$$\overline{\lim_{n \to \infty}} \max_a \left(\sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \mu(y) 1\{y \notin \mathbf{X}_n\} \right) = b. \tag{16.7}$$

Let a_n be some actions achieving the max (the fact that the max is achieved follows from the compactness of the action space and continuity assumption (7.35)). Choose a subsequence $n(\ell), \ell = 1, 2, \ldots$ along which the limsup is obtained and along which a_n converges to some action a^*. Then $\mathcal{P}_{xa_{n(\ell)},\bullet}$ converges (pointwise) to the probability $\mathcal{P}_{xa^*,\bullet}$ as $\ell \to \infty$. But then it follows from a dominant convergence theorem (Royden, 1988, Chapter 11 Section 4) and from the fact that $\mathbf{X}_n$ increase to $\mathbf{X}$, that

$$\lim_{\ell \to \infty} \sum_{y \in \mathbf{X}} \mathcal{P}_{xa_{n(\ell)}y} \mu(y) 1\{y \notin \mathbf{X}_{n(\ell)}\} = \sum_{y \in \mathbf{X}} \mathcal{P}_{xa^*y} \mu(y) \cdot 0 = 0,$$

which contradicts (16.7). Hence (16.5) indeed holds. □

16.2 Scheme I: the total cost

The set $\mathcal{X}$ may or may not contain $\mathcal{M}$ (recall that $\mathcal{M}$ is part of the definition of the total expected cost criterion). We assume, however, that $\mathbf{X}_n$ includes some state, called 0, which belongs to $\mathcal{M}$.

In $\mathbf{COP}_n$, we modify the transition probabilities so as to eliminate all transitions outside the set $\mathbf{X}_n$; we replace transitions outside of $\mathbf{X}_n$ by transitions to $0 \in \mathcal{M}$. Hence, $\mathcal{P}^n_{xay}$ is defined by:

$$\mathcal{P}^n_{xay} = \begin{cases} \mathcal{P}_{xa0} + \sum_{z \notin \mathbf{X}_n} \mathcal{P}_{xaz} & y = 0 \\ \mathcal{P}_{xay} & y \neq 0, y \in \mathbf{X}_n \\ 0 & y \notin \mathbf{X}_n \end{cases} \tag{16.8}$$

Both $\mathbf{COP}$ and $\mathbf{COP}_n$ have optimal stationary policies according to

Theorem 8.4. We can therefore consider **COP** and $\mathbf{COP}_n$ restricted to U_S. When applying the results and checking the assumptions of Sections 13.2 and 13.3, we shall use $\overline{U} = U_S$ to conclude that the optimal values and policies converge.

Let $C_{tc}^n(\beta, w)$, $D_{tc}^{k,n}(\beta, w)$, $k = 1, \ldots, K$, be the costs under a policy w corresponding to the nth approximation (i.e., to the transition probabilities $\mathcal{P}^n$). Let $C_{tc}^n(\beta)$ denote the corresponding optimal value.

Fix an arbitrary stationary policy w. From Remark 9.2 it follows that $C_{tc}(\cdot, w)$ and $C_{tc}^n(\cdot, w)$ are the unique solutions in F^μ of the fixed point equations

$$\begin{aligned} \phi(x,w) &= c(x,w) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xwy}\phi(y,w), \qquad x \in \mathbf{X}, \qquad (16.9) \\ \phi^n(x,w) &= c(x,w) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xwy}^n \phi^n(y,w), \qquad x \in \mathbf{X}_n. \end{aligned}$$

Theorem 16.1 (*Convergence of values and policies*)
Consider the contracting framework. Assume that there exists some policy v satisfying the Slater condition

$$D_{tc}(\beta, v) < V. \qquad (16.10)$$

Under Scheme I,
(i) The values $C_{tc}^n(\beta)$ of the truncated MDP converge to the value $C_{tc}(\beta)$ of the original one;
(ii) For any $\varepsilon > 0$, there exists a stationary policy w (characterized in Theorem 13.4(i)) that is ε-optimal for $\mathbf{COP}_n$ for all n sufficiently large;
(iii) Any policy w which is a limit of optimal stationary policies for $\mathbf{COP}_n$ (as n tends to ∞) is optimal for **COP**.

Proof. The proof is obtained by applying Theorems 13.1, 13.3 and 13.4. We show that the assumptions there indeed hold.

(**S**1) holds by Assumption (16.10); (**S**2) is established in Corollary 9.2 and Theorem 9.10; Lemma 8.5 (ii) implies (**S**4). (**S**5) follows from Theorem 9.8 (ii). It remains to establish (**S**3). We prove it for C_{tc}; the same proof holds for D_{tc}^k.

Fix a stationary policy w. We estimate $\|C_{tc}^n(\cdot, w) - C_{tc}(\cdot, w)\|_\mu^{\mathcal{X}}$. We first present a simple proof for the special case where the finite neighbor Assumption (16.1) holds, and when $\mathbf{X}_n$ are defined in (16.2). In that case, for $x \in \mathcal{X}$, we have

$$\begin{aligned} &\frac{1}{\mu(x)} \left| C_{tc}^n(x,w) - C_{tc}(x,w) \right| \\ &\quad \leq \frac{1}{\mu(x)} \sum_{y \in \mathbf{Y}(x)} \mathcal{P}_{xwy}\mu(y) \frac{|C_{tc}^n(y,w) - C_{tc}(y,w)|}{\mu(y)} \end{aligned}$$

$$\leq \quad \xi \, \|C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w)\|^{\mathbf{X}_1}_{\mu} .$$

Continuing this way, we obtain

$$\|C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w)\|^{\mathcal{X}}_{\mu} \leq \xi^n \, \|C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w)\|^{\mathbf{X}_n}_{\mu} .$$

Since

$$\|C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w)\|^{\mathbf{X}_n}_{\mu} \leq \frac{2\overline{b}}{1-\xi},$$

we finally get

$$\|C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w)\|^{\mathcal{X}}_{\mu} \leq \frac{2\overline{b}\xi^n}{1-\xi}.$$

This establishes (**S3**) under the conditions (16.1) and when $\mathbf{X}_n$ are defined in (16.2).

Next we consider the general case. Let $n \geq m^{\ell}(\varepsilon)$, where ℓ is some given integer. Clearly,

$$\|C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w)\|^{\mathcal{X}}_{\mu} \leq \|C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w)\|^{\mathbf{X}_{g_0}}_{\mu}$$

since $\mathcal{X} \subset \mathbf{X}_{g_0}$. For $x \in \mathbf{X}_{g_0}$,

$$\begin{aligned}
&\frac{1}{\mu(x)} \left| C^n_{tc}(x, w) - C_{tc}(x, w) \right| \\
&= \quad \frac{1}{\mu(x)} \left| \sum_{y \notin \mathcal{M}} \mathcal{P}^n_{xwy} C^n_{tc}(y, w) - \mathcal{P}_{xwy} C_{tc}(y, w) \right| \\
&\leq \quad \frac{1}{\mu(x)} \sum_{y \in \mathbf{X}_{g_1} \setminus \mathcal{M}} \mathcal{P}^n_{xwy} \left| C^n_{tc}(y, w) - \mathcal{P}_{xwy} C_{tc}(y, w) \right| \\
&\qquad + \frac{1}{\mu(x)} \sum_{y \notin \mathbf{X}_{g_1}} \left| \mathcal{P}^n_{xwy} C^n_{tc}(y, w) \right| + \left| \mathcal{P}_{xwy} C_{tc}(y, w) \right| \\
&\leq \quad \sum_{y \in \mathbf{X}_{g_1} \setminus \mathcal{M}} \frac{\mathcal{P}_{xwy} \mu(y)}{\mu(x)} \frac{\left| \mathcal{P}^n_{xwy} C^n_{tc}(y, w) - \mathcal{P}_{xwy} C_{tc}(y, w) \right|}{\mu(y)} \\
&\qquad + \frac{1}{\mu(x)} \sum_{y \notin \mathbf{X}_{g_1}} \mathcal{P}_{xwy} \mu(y) \left(\left| \frac{C^n_{tc}(y, w)}{\mu(y)} \right| + \left| \frac{C_{tc}(y, w)}{\mu(y)} \right| \right).
\end{aligned}$$

In the last inequality, the first term is bounded by

$$\xi \parallel C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w) \parallel^{\mathbf{X}_{g_1}}_{\mu}$$

(because of Assumption (7.34)) and the second by $2\overline{b}\varepsilon/(1-\xi)$ (due to the definition of the sequence $\mathbf{X}_{g_k}$ and by Theorem 8.3 (iii)). We obtain

$$\parallel C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w) \parallel^{\mathbf{X}_{g_0}}_{\mu} \; \leq \; \xi \parallel C^n_{tc}(\cdot, w) - C_{tc}(\cdot, w) \parallel^{\mathbf{X}_{g_1}}_{\mu} + 2\frac{\overline{b}\varepsilon}{1-\xi}.$$

In the same way we get for $g_k \le m^\ell(\varepsilon) \le n$

$$\| C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \|^{\mathbf{X}_{g_k}}_\mu \;\le\; \xi \;\| C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \|^{\mathbf{X}_{g_{k+1}}}_\mu + 2\frac{\overline{b}\varepsilon}{1-\xi}.$$

Since

$$\| C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \|^{\mathbf{X}_{g_k}}_\mu \;\le\; \frac{2\overline{b}}{1-\xi},$$

we get for any integer ℓ with $n \ge m^\ell(\varepsilon)$,

$$\| C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \|^{\mathcal{X}}_\mu \le \xi^\ell \frac{2\overline{b}}{1-\xi} + \frac{2\overline{b}\varepsilon}{1-\xi}\left(\frac{1-\xi^\ell}{1-\xi}\right) =: \varepsilon_1(n). \quad (16.11)$$

Since ℓ can be chosen arbitrarily large, and ξ is strictly less than 1, this bound can be as small as needed for n large enough. By applying the same arguments again for for $D^k_{tc}(x,u)$, $k = 1,\dots,K$, we finally establish condition (**S**3). □

Remark 16.1 Other results from Chapter 13 can be used to further characterize the convergence of values and policies. In particular, one may use Theorem 13.1 (ii) to further characterize the rate of convergence of $C^n_{tc}(\beta)$ to $C_{tc}(\beta)$, based on the uniform bound (16.11). Moreover, a construction of almost optimal policies for **COP** based on policies that are optimal for $\mathbf{COP}_n$, or vice versa, can be carried out in a way similar to Theorem 13.4 (ii).

16.3 Scheme II: the total cost

In the previous scheme, we replaced transitions outside of $\mathbf{X}_n$ by transitions to state 0. In some applications this may be undesirable; this is the case when the MDPs with truncated space describe real problems that we wish to approximate by some MDP with an infinite state space. To illustrate this, consider a queue with a finite length L, and assume that the state is the number of customers in the queue. Then typically, if a transition from state L to state $L+1$ were possible in the case of infinite queue, then in the problem with truncated state space, which corresponds to a finite queue, it is replaced by a transition from L to L. In the previous scheme, it would be replaced by a transition to state 0. This would be especially undesirable, since in queueing problems, we usually have the property of transitions to closest neighbors: from each state, only finitely many neighboring states can be reached in one step. So, having a transition from state L to 0 does not describe a realistic model of a finite queue.

Let $\{q^n_{xay}, x, y \in \mathbf{X}, a \in \mathsf{A}(x)\}$ be a sequence of measures such that for all n, $x \in \mathbf{X}_n$, $a \in \mathsf{A}(x)$,

$$q^n_{xay} \ge 0 \text{ for } y \in \mathbf{X}_n, \qquad q^n_{xay} = 0 \text{ for } y \notin \mathbf{X}_n, \qquad \sum_{y \in \mathbf{X}_n} (\mathcal{P}_{xay} + q^n_{xay}) = 1.$$

The transitions for the finite problems are then given by

$$\mathcal{P}^n_{xay} = \begin{cases} \mathcal{P}_{xay} + q^n_{xay} & x, y \in \mathbf{X}_n \\ 0 & \text{otherwise.} \end{cases} \tag{16.12}$$

It follows that $\forall x \in \mathbf{X}_n$,

$$\sum_{y \in \mathbf{X}_n} q^n_{xay} = \sum_{y \notin \mathbf{X}_n} \mathcal{P}_{xay}. \tag{16.13}$$

We make the following assumption on μ and on $\mathbf{X}_n$.

$$\text{For any } n > m \text{ and } x \in \mathbf{X}_n \setminus \mathbf{X}_m, \mu(x) \geq \sup_{y \in \mathbf{X}_m} \mu(x) =: \overline{\mu}_m.$$

Theorem 16.2 (*Convergence of the values and policies*)
Consider the contracting framework. Assume that there exists some policy v satisfying the Slater condition (16.10)*. Then under Scheme II, all the statements of Theorem* 16.1 *hold.*

Proof. The proof is obtained by applying again Theorems 13.1, 13.3 and 13.4. We have to check again assumption (**S**3); the other assumptions (**S**1), (**S**2), (**S**4) and (**S**5) were established already in the beginning of the proof of Theorem 16.1. For any stationary policy w, $C_{tc}(\cdot, w)$ and $C^n_{tc}(\cdot, w)$ are again the unique solutions in F^μ of the fixed-point equations (16.9) (with the new transition probabilities $\mathcal{P}^n$).

We begin by obtaining a bound for $C^n_{tc}(x, w)$, uniformly over n and w.

$$\begin{aligned} C^n_{tc}(x,w) &= c(x,w) + \sum_{y \in \mathbf{X}_n \setminus \mathcal{M}} \mathcal{P}^n_{xwy} C^n_{tc}(y) \\ &= c(x,w) + \sum_{y \in \mathbf{X}_n \setminus \mathcal{M}} \mathcal{P}_{xwy} C^n_{tc}(y) + \sum_{y \in \mathbf{X}_n \setminus \mathcal{M}} q^n_{xwy} C^n_{tc}(y) \\ &\leq c(x,w) + \sum_{y \in \mathbf{X}_n \setminus \mathcal{M}} \mathcal{P}_{xwy} C^n_{tc}(y) + \sum_{y \in \mathbf{X}_n \setminus \mathcal{M}} q^n_{xwy} \overline{\mu}_n \sup_{y' \in \mathbf{X}_n} \frac{C^n_{tc}(y')}{\mu(y')} \\ &\leq c(x,w) + \sum_{y \in \mathbf{X}_n \setminus \mathcal{M}} \mathcal{P}_{xwy} \mu(y) \frac{C^n_{tc}(y)}{\mu(y)} \\ &\quad + \sum_{y \notin \mathbf{X}_n} \mathcal{P}_{xay} \mu(y) \sup_{y' \in \mathbf{X}_n} \frac{C^n_{tc}(y')}{\mu(y')} \end{aligned}$$

We thus conclude that

$$\left\| C^n_{tc}(\cdot, w) \right\|_\mu \leq \overline{b} + \xi \left\| C^n_{tc}(\cdot, w) \right\|_\mu$$

so that

$$\left\| C^n_{tc}(\cdot, w) \right\|_\mu \leq \frac{\overline{b}}{1 - \xi}.$$

Let $n \geq m^\ell(\varepsilon)$, where ℓ is some given integer (hence, in particular, $n \geq g_1$). For $x \in \mathbf{X}_{g_0}$ (and thus, in particular, for $x \in \mathcal{X}$),

$$\begin{aligned}
&\frac{1}{\mu(x)}\,|C^n_{tc}(x,w) - C_{tc}(x,w)| \\
&= \frac{1}{\mu(x)}\left|\sum_{y\notin\mathcal{M}} \mathcal{P}^n_{xwy}C^n_{tc}(y,w) - \mathcal{P}_{xwy}C_{tc}(y,w)\right| \\
&\leq \frac{1}{\mu(x)}\sum_{y\in \mathbf{X}_{g_1}\setminus\mathcal{M}} \mathcal{P}_{xwy}|C^n_{tc}(y,w) - C_{tc}(y,w)| \\
&\quad + \frac{1}{\mu(x)}\sum_{y\in \mathbf{X}_n\setminus\mathcal{M}} q^n_{xwy}\,|C^n_{tc}(y,w)| \\
&\quad + \frac{1}{\mu(x)}\sum_{y\notin \mathbf{X}_{g_1}} \mathcal{P}_{xwy}|C^n_{tc}(y,w)| + \mathcal{P}_{xwy}|C_{tc}(y,w)| \\
&\leq \sum_{y\in \mathbf{X}_{g_1}\setminus\mathcal{M}} \frac{\mathcal{P}_{xwy}\mu(y)}{\mu(x)}\,\frac{\left|\mathcal{P}^n_{xwy}C^n_{tc}(y,w) - \mathcal{P}_{xwy}C_{tc}(y,w)\right|}{\mu(y)} \\
&\quad + \frac{1}{\mu(x)}\sum_{y\in \mathbf{X}_n\setminus\mathcal{M}} q^n_{xwy}\mu(y)\sup_{y'\in\mathbf{X}_n}\frac{|C^n_{tc}(y',w)|}{\mu(y')} \\
&\quad + \frac{1}{\mu(x)}\sum_{y\notin \mathbf{X}_{g_1}} \mathcal{P}_{xwy}\mu(y)\left(\left|\frac{C^n_{tc}(y,w)}{\mu(y)}\right| + \left|\frac{C_{tc}(y,w)}{\mu(y)}\right|\right) \\
&\leq \xi \parallel C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \parallel_\mu^{\mathbf{X}_{g_1}} \\
&\quad + \frac{1}{\mu(x)}\sum_{y\notin \mathbf{X}_{g_1}} \mathcal{P}_{xwy}\mu(y)\sup_{y'\in\mathbf{X}_n}\frac{|C^n_{tc}(y',w)|}{\mu(y')} + \frac{2\overline{b}\varepsilon}{1-\xi} \\
&\leq \xi \parallel C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \parallel_\mu^{\mathbf{X}_{g_1}} + \frac{3\overline{b}\varepsilon}{1-\xi}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\parallel C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \parallel_\mu^{\mathbf{X}_{g_0}} \\
&\quad \leq \xi \parallel C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \parallel_\mu^{\mathbf{X}_{g_1}} + 3\frac{\overline{b}\varepsilon}{1-\xi} =: \varepsilon_1(n).
\end{aligned}$$

As in (16.11), we get for any integer ℓ with $n \geq m^\ell(\varepsilon)$,

$$\parallel C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \parallel_\mu^{\mathcal{X}} \;\leq\; \xi^\ell\frac{2\overline{b}}{1-\xi} + \frac{3\overline{b}\varepsilon}{1-\xi}\left(\frac{1-\xi^\ell}{1-\xi}\right). \tag{16.14}$$

This establishes (**S**3). □

Again, one may use Theorem 13.1 (ii) to further characterize the rate of convergence of $C_{tc}^n(\beta)$ to $C_{tc}(\beta)$; a construction of almost optimal policies for **COP** based on policies that are optimal for $\mathbf{COP}_n$ or vice versa can be carried out as in Theorem 13.4 (ii).

16.4 Scheme III: the total cost

The basic idea of the approximation scheme is to fix some stationary policy $u \in U_S$ and use it in all states except for a subset $\mathbf{X}_n$. The problem is then of determining optimal strategies in the remaining set of states $\mathbf{X}_n$. We are interested in studying the asymptotic behavior of this approach as $\mathbf{X}_n \to \mathbf{X}$. We note that in this approach, the set of policies depends on n (see Remark 13.1):

$$U_n = \{w \in U_S : w_x = u_x,\ \forall x \notin \mathbf{X}_n\}.$$

To avoid this problem, we introduce the projection $\pi : U_S \to U_n$:

$$\pi_x^n(w) = \begin{cases} w(x) & \text{if } x \in \mathbf{X}_n, \\ \\ u(x) & \text{if } x \notin \mathbf{X}_n. \end{cases}$$

We then define for any $w \in U_S$:

$$C_{tc}^n(\beta, w) := C_{tc}(\beta, \pi^n(w)).$$

Using the same techniques as in the previous sections, one can show again that the results of Theorem 16.1 hold also for Scheme III, see Tidball and Altman (1996a, 1996b), Tidball *et al.* (1997).

16.5 The expected average cost

All the results of previous sections hold also for the expected average cost. This is summarized in the following:

Theorem 16.3 (*Convergence of values and policies*)
Consider a uniform geometric recurrent MDP (Definition 11.4) with $n_0 = 1$*, and which is unichain. Assume that there exists some policy* v *satisfying the Slater condition*

$$D_{ea}(\beta, v) < V. \tag{16.15}$$

Under Scheme I, II or III,
(*i*) *The values* $C_{ea}^n(\beta)$ *of the truncated MDP converge to the value* $C_{ea}(\beta)$ *of the original one.*
(*ii*) *For any* $\varepsilon > 0$*, there exists a stationary policy* w (*characterized in Theorem* 13.4(*i*)) *that is* ε*-optimal for* $\mathbf{COP}_n$ *for all* n *sufficiently large.*
(*iii*) *Any policy* w *which is a limit of optimal stationary policies for* $\mathbf{COP}_n$ *(as* n *tends to* ∞*) is optimal for* **COP**.

Proof. The proof is obtained by applying Theorems 13.1, 13.3 and 13.4. We show that the assumptions there indeed hold. (**S**1) holds by Assumption (16.10); (**S**2) is established in Corollary 12.2 and Theorem 12.8; Lemma 11.3 (ii) implies (**S**4). (**S**5) follows from Theorem 12.6 (ii). It remains to establish (**S**3). We prove it for C_{ea}; the same proof holds for D^k_{ea}.

By the definition of uniform geometric ergodicity (Definition 11.4), there exists some finite set $\mathcal{M}$ and

$$\sum_{y \notin \mathcal{M}} [P^{n_0}(u)]_{xy} \mu(y) \leq \xi \mu(x) \tag{16.16}$$

(we assumed $n_0 = 1$). It follows (see Spieksma, 1990) that one may choose some state, say 0, with $0 \in \mathcal{M}$, some μ' and ξ' such that (16.16) holds for $\mathcal{M}' = \{0\}$ and μ' and ξ' (instead of $\mathcal{M}$ and μ and ξ). In other words, we may assume, without loss of generality, that $\mathcal{M}$ contains a single state 0. Define

$$T := \inf_{t>0}\{X_t = 0\}, \qquad M_w(0) := E_0^w T.$$

For any stationary policy, say w, we have

$$C_{ea}(x, w) = \frac{C_{tc}(0, w)}{M_w(0)}$$

(see Chung 1967, p. 91-92), where by $C_{tc}(0, w)$ we mean the standard total costs until we hit the set $\mathcal{M} = \{0\}$. Similarly, we have

$$C^n_{ea}(x, w) = \frac{C^n_{tc}(0, w)}{M_w(0)}, \tag{16.17}$$

where both C^n_{tc} and $M_w(0)$ are the corresponding total expected costs and expected recurrence times corresponding to Scheme I, II, or III. It follows as in the previous sections that $C^n_{tc}(0, w)$ converges to $C_{tc}(0, w)$ uniformly in U_S. Similarly, one can show that $M_w(0)$ converges to $M_w(0)$ uniformly in U_S (this is obtained by identifying $M_w(0)$ as the total expected cost until $\mathcal{M} = \{0\}$ is hit, for the immediate cost of $c'(x, a) = 1$). This, together with (16.17), implies that $C^n_{ea}(x, w)$ converges to $C_{ea}(x, w)$ uniformly in $w \in U_S$, which establishes (**S**3). □

16.6 Infinite MDPs: on the number of randomizations

We know that in CMDPs, optimal stationary policies exist under several sets of conditions, and that they require randomization. A natural question is how many randomizations are needed. For the case of finite-state and action spaces, we have shown in Sections 3.5 and 4.4 that there exists an optimal stationary policy that requires no more than K randomizations, where K is the number of constraints.

More involved techniques have been used by Borkar (1990, 1994) to es-

tablish the result for the case of a countable state space (see also Feinberg and Shwartz, 1995, 1996).

An alternative simple approach to conclude that optimal stationary policies exist for infinite MDPs, which require at most K randomizations, is to use finite state approximations. Indeed, assume that the number of actions available at each state is finite. We established the convergence of stationary policies that are optimal for the truncated-state space to a stationary policy that is optimal for the original problem. Since for each of the truncated problems we know that no more than K randomizations are required, we conclude that there indeed exists an optimal stationary policy for the original problem with that property.

CHAPTER 17

Appendix: Convergence of probability measures

Definition 17.1 *(Vague and weak convergence of measures)*
A sequence of measures f^t over a metric space $\mathcal{K}$ is said to

- Converge vaguely to f if and only if

$$\int g(\omega) f^t(d\omega) \to \int g(\omega) f(d\omega) \tag{17.1}$$

  for all $g \in C_0(\mathcal{K})$, the space of continuous functions that vanish at infinity.
- It converges weakly to f if and only if (17.1) holds for all $g \in C_b(\mathcal{K})$, the space of bounded continuous functions.

Definition 17.2 *(Tightness)*
A set of probability measures $\{f^n\}_{n\in I}$ (I is some set) over $\mathcal{K}$ is called tight if for any $\varepsilon > 0$ there exists some compact set $K_\varepsilon \in \mathbb{K}$ such that

$$f^n(K_\varepsilon) > 1 - \varepsilon, \qquad \forall n \in I.$$

Definition 17.3 *(Uniform integrability of random variables)*
A set $\{R_t\}_{t\in I}$ of random variables over some probability space $(\Omega, \mathcal{F}, P)$ is said to be uniformly integrable if for any $\varepsilon > 0$ there exists some integer n such that

$$E|R_t|1\{|R_t| > n\} < \varepsilon$$

for all t.

Remark 17.1 It clearly follows from the definition that if R_t is uniformly integrable w.r.t. some P, then so is any other sequence R'_t for which $|R'_t| \leq |R_t|$; in fact, the whole set $\{R'_t, R_t, t \in I\}$ is uniformly integrable.

Definition 17.4 *(Uniform integrability of non-negative measures)*
A set of non-negative measures $\{f^t\}_{t\in I}$ over a locally compact set $\mathcal{K}$ is said to be uniformly integrable with respect to a function $c : \mathcal{K} \to \mathbb{R}$ if for any $\delta > 0$, there exists some compact set $\mathcal{K}'$ such that for all $t \in I$,

$$\int 1\{\kappa \notin \mathcal{K}'\} f^t(d\kappa)|c(\kappa)| < \delta.$$

We have (Prohorov's Theorem, see e.g., Billingsley, 1968, Theorems 6.1 and 6.2):

Lemma 17.1 (*Characterization of tightness*)
A set of probability measures F over a separable metric space is tight if and only if every sequence $\{f^t\}_{t\in\mathbb{N}}$ in F has a subsequence $\{f^{t_n}\}_{n\in\mathbb{N}}$ that weakly converges to some probability measure f (which need not be in F).

Lemma 17.2 (*Properties of vague convergence*)
Let $\{f^n\}_{n\in\mathbb{N}}$ be some non-negative measures over some Borel space $\mathcal{B}$, with f^n converging vaguely to f. Then,
(i) For any non-negative continuous function g over $\mathcal{B}$, we have

$$\underline{\lim}_{n\to\infty}\int_{\mathcal{B}} g(\omega)f^n(d\omega) \geq \int_{\mathcal{B}} g(\omega)f(d\omega).$$

(ii) Let $\{f^n\}_{n\in\mathbb{N}}$ be some probability measures over some Borel space $\mathcal{B}$, with f^n converging vaguely to f. The following are equivalent:
a. f^n is tight,
b. f is a probability measure,
c. f^n converges weakly to f.

Proof. (i) Let $\mathcal{B}_n$ be a sequence of compact subsets of $\mathcal{B}$ that increases to $\mathcal{B}$. One can choose an increasing sequence of bounded continuous functions g_n that converges (pointwise) to g such that g_n vanishes outside of $\mathcal{B}_n$.

Since $g \geq 0$, we have for any integer m,

$$\begin{aligned}\underline{\lim}_{n\to\infty}\int_{\mathcal{B}} g(\omega)f^n(d\omega) &\geq \underline{\lim}_{n\to\infty}\int_{\mathcal{B}} g_m(\omega)f^n(d\omega)\\ &= \int_{\mathcal{B}} g_m(\omega)f(d\omega),\end{aligned}$$

where the last equality follows from the definition of vague convergence, and since the function $g_m(\omega) \in C_0(\mathcal{B})$. The lemma is now obtained by applying the monotone convergence theorem (taking $m \to \infty$).

(ii) That (c) implies (b) follows from Portmanteau's Theorem (Billingsley, 1968, p. 11).
The implication (c) to (a) follows from Prohorov's Theorem (Lemma 17.1). Let f^n vaguely converge to f. We show that (a) implies (c). Choose any subsequence n_i of n. Lemma 17.1 implies the existence of a subsequence n_{i_j} of n_i along which f^n weakly converges to some f'. Since weak convergence implies vague convergence, this means that $f' = f$. Since this holds for any sequence n_i, we conclude that f is the weak limit of f^n. Finally, we show that (b) implies (c). Choose any continuous function g bounded by some constant b. Then it follows from part (i) of the lemma applied to the functions $b + g$ and $b - g$, that

$$\begin{aligned}\underline{\lim}_{n\to\infty}\int_{\mathcal{B}} g(\omega)f^n(d\omega) &= -b + \underline{\lim}_{n\to\infty}\int_{\mathcal{B}} [g(\omega)+b]f^n(d\omega)\\ &\geq -b + \int_{\mathcal{B}} [g(\omega)+b]f(d\omega) = \int_{\mathcal{B}} g(\omega)f(d\omega),\end{aligned}$$

as well as

$$\begin{aligned}\overline{\lim_{n\to\infty}} \int_{\mathcal{B}} g(\omega) f^n(d\omega) &= b - \underline{\lim_{n\to\infty}} \int_{\mathcal{B}} [b - g(\omega)] f^n(d\omega) \\ &\leq \quad b - \int_{\mathcal{B}} [b - g(\omega)] f(d\omega) = \int_{\mathcal{B}} g(\omega) f(d\omega),\end{aligned}$$

which establishes the proof. □

The following is proved, e.g., in Theorem 5.3 of Billingsley (1968):

Lemma 17.3 (*Conditions for uniform integrability of RVs*)
Let $\{R_t\}_{t\in\mathbb{N}}$ be a sequence of random variables converging in distribution to a random variable R. Then (i) if R_t is uniformly integrable, then

$$\lim_{t\to\infty} ER_t = ER.$$

(ii) If R_t are integrable and non-negative, then the converse also holds.

Next, we present the counterpart for the uniform integrability of probability measures.

Lemma 17.4 (*Conditions for uniform integrability of non-negative measures*)
Consider a sequence $\{f^t\}_{t\in\mathbb{N}}$ of non-negative measures over some locally compact space X, converging weakly to a measure f. Let $\mu : X \to \mathbb{R}$ be some given continuous function.
(i) If f^t are uniformly integrable w.r.t. to μ, then

$$\lim_{t\to\infty} \langle f^t, \mu\rangle = \langle f, \mu\rangle < \infty. \tag{17.2}$$

(ii) Assume that μ is strictly positive and f^t are integrable with respect to μ. Assume that (17.2) holds. Then f^t are uniformly integrable w.r.t. to μ.

Proof. Assume that f^t are uniformly integrable w.r.t. to μ. Let X_n be an increasing sequence of compact sets converging to X. Choose some ε, and let n be such that $\int 1\{x \notin X_n\}|\mu(x)|f^t(dx) < \varepsilon$ for all t. $|\mu|$ is bounded on X_n by some constant μ_n since it is continuous. Let $\mu^n(x) = \min(|\mu(x)|, \mu_n(x))$. The weak convergence of f^t implies that

$$\lim_{t\to\infty} \int \mu^n(x) f^t(dx) - \int \mu^n(x) f(dx) = 0.$$

Lemma 17.2 (i) implies that

$$\int 1\{x \notin X_n\}|\mu(x)|f(dx) \leq \sup_t \int 1\{x \notin X_n\}|\mu(x)|f^t(dx) \leq \varepsilon.$$

We thus obtain

$$\lim_{t\to\infty} \left| \int \mu(x) f^t(dx) - \int \mu(x) f(dx) \right|$$

$$\begin{aligned} &\leq \lim_{t\to\infty} \left| \int \mu^n(x) f^t(dx) - \int \mu^n(x) f(dx) \right| \\ &+ \sup_t \int 1\{x \notin X_n\} |\mu(x)| f^t(dx) + \int 1\{x \notin X_n\} |\mu(x)| f(dx) \leq 2\varepsilon. \end{aligned}$$

Since this holds for any ε (with the corresponding n), (i) follows.

(ii) Let X_n be, again, an increasing sequence of compact sets converging to X. If f^t are not uniformly integrable with respect to μ, then there exists some $\varepsilon > 0$ and some strictly increasing sequence $t(n)$ such that

$$\int \mu(y) 1\{y \notin X_n\} f^{t(n)}(dy) \geq \varepsilon.$$

(The fact that $t(n)$ can be chosen to be strictly increasing follows from the integrability of f^t for all t.) Since μ is non-negative, this implies that for any n,

$$\underline{\lim}_{t\to\infty} \int \mu(y) 1\{y \notin X_n\} f^t(dy) \geq \varepsilon.$$

f is integrable w.r.t. μ (as $\langle f, \mu\rangle \leq \underline{\lim}_{t\to\infty} \langle f^t, \mu\rangle$). Let N be such that for all $n \geq N$,

$$\int 1\{x \notin X_n\} \mu(x) f(dx) < \varepsilon/2.$$

Define μ_n to be an upper bound on $\mu(x)$ over X_n, and $\mu^n = \min(\mu(x), \mu_n(x))$. Then

$$\begin{aligned} &\langle f, \mu\rangle - \langle f^t, \mu\rangle \\ &\quad \leq \underline{\lim}_{t\to\infty} \left(\int \mu^n(x) f(dx) - \int \mu^n(x) f^t(dx) \right) \\ &\qquad + \varepsilon/2 - \int 1\{x \notin X_n\} \mu(x) f^t(dx) \\ &\quad \leq -\varepsilon/2. \end{aligned}$$

This establishes (ii). □

CHAPTER 18

References

E. Altman (1993), 'Asymptotic properties of constrained Markov decision processes', *ZOR – Methods and Models in Operations Research*, **37**, Issue 2, pp. 151-170.

E. Altman (1994), 'Denumerable constrained Markov decision processes and finite approximations', *Math. of Operations Research*, **19**, pp. 169-191.

E. Altman (1996), 'Constrained Markov decision processes with total cost criteria: occupation measures and primal LP', *ZOR – Mathematical Methods in Operations Research*, **43**, Issue 1, pp. 45-72.

E. Altman (1998), 'Constrained Markov decision processes with total cost criteria: Lagrange approach and dual LP', *ZOR – Mathematical Methods in Operations Research*, **48**, pp. 387-417, 1998.

E. Altman and V. A. Gaitsgory (1993), 'Stability and singular perturbations in constrained Markov decision problems', *IEEE Transactions on Automatic Control*, **38**, pp. 971-975.

E. Altman and V. A. Gaitsgory (1995), 'A hybrid (differential-stochastic) zero-sum game with fast stochastic part', *Annals of the International Society of Dynamic Games*, **3**, pp. 47-59.

E. Altman, A. Hordijk and L. C. M. Kallenberg (1996), 'On the value in constrained control of Markov chains', *ZOR – Methods and Models in Operations Research*, **44**, Issue 3, pp. 387-400.

E. Altman, A. Hordijk and F. M. Spieksma (1997), 'Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards', *MOR*, **22** No. 3, pp. 588-618.

E. Altman and A. Shwartz (1988), 'Markov optimization problems: state-action frequencies revisited', *27th IEEE Conference on Decision and Control*, Austin, Texas, December 1988 (invited paper).

E. Altman and A. Shwartz (1989), 'Optimal priority assignment: a time sharing approach', *IEEE Transactions on Automatic Control* **AC-34**, pp. 1089-1102.

E. Altman and A. Shwartz (1991a), 'Markov decision problems and state-action frequencies', *SIAM J. Control and Optimization,* **29**, pp. 786-809.

E. Altman and A. Shwartz (1991b), 'Adaptive control of constrained Markov chains', *IEEE Transactions on Automatic Control*, **36**, pp. 454-462.

E. Altman and A. Shwartz (1991c), 'Sensitivity of constrained Markov Decision Problems', *Annals of Operations Research*, **32**, pp. 1-22.

E. Altman and A. Shwartz (1991d), 'Adaptive control of constrained Markov chains: criteria and policies', *Annals of Operations Research* **28**, special issue on 'Markov Decision Processes', Eds. O. Hernández-Lerma and J. B. Lasserre, pp. 101-134.

E. Altman and A. Shwartz (1993), 'Time-sharing policies for controlled Markov chains', *Operations Research*, **41**, pp. 1116-1124.

E. Altman and A. Shwartz (1995), 'Constrained Markov games: Nash equilibria', submitted to *Annals of Dynamic Games*.

E. Altman and F. Spieksma (1995), 'The Linear Program approach in Markov decision problems revisited', *ZOR – Methods and Models in Operations Research*, **42**, Issue 2, pp. 169-188.

E. Altman and O. Zeitouni (1994), 'Rate of convergence of empirical measures and costs in controlled Markov chains and transient optimality', *Math. of Operations Research*, **19**, pp. 955-974.

J. Anderson and P. Nash (1987), *Linear Programming in Infinite-Dimensional Spaces*, Wiley, England.

A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh and S. I. Marcus (1993), 'Discrete-time controlled Markov processes with average cost criterion: a survey', *SIAM J. Control and Optimization*, **31**, pp. 282-344.

J. P. Aubin (1993), *Optima and Equilibria, An Introduction to Nonlinear Analysis*, Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.

R. J. Aumann (1964), 'Mixed and behavior strategies in infinite extensive games', Advances in Game Theory, *Ann. Math. Study*, **52**, pp. 627-650.

J. Bather (1973), 'Optimal decision procedures for finite Markov chains. Part II: Communicating systems', *Advances in Applied Probability*, **5**, pp. 521-540.

M. Bayal-Gursoy and K. W. Ross (1992), 'Variability sensitive Markov decision processes', *Math. of Operations Research*, **17**, pp. 558-571.

P. Bernhard (1992), 'Information and strategies in dynamic games', *SIAM J. Cont. and Opt.*, **30**, pp. 212-228.

F. J. Beutler and K. W. Ross (1985), 'Optimal policies for controlled Markov chains with a constraint', *J. Mathematical Analysis and Applications,* **112**, 236-252.

F. J. Beutler and K. W. Ross (1986), 'Time-average optimal constrained semi-Markov decision processes', *Advances of Applied Probability*, **18**, pp. 341-359.

P. Billingsley (1968), *Convergence of Probability Measures*, J. Wiley, New York.

J. R. Birge and R. J. Wets (1986), 'Designing approximating schemes for stochastic optimization problems', *Math. Programm. Study*, **27**, pp. 54-102.

V. S. Borkar (1983), 'On minimum cost per unit time control of Markov chains', *SIAM J. Control Optim.*, **22**, pp. 965-978.

V. S. Borkar (1988), 'A convex analytic approach to Markov decision processes', *Prob. Th. Rel. Fields*, **78**, pp. 583-602.

V. S. Borkar (1990), *Topics in Controlled Markov Chains*, Longman Scientific & Technical.

V. S. Borkar (1993), 'Controlled diffusions with constraints, II', *Journal of Math. Analysis and Appli.*, **176**, No. 2, pp. 310-321.

V. S. Borkar (1994), 'Ergodic control of Markov Chains with constraints — the general case', *SIAM J. Control and Optimization*, **32**, pp. 176-186.

V. S. Borkar and M. M. Ghosh (1990), 'Controlled diffusions with constraints', *Mathematical Analysis and Applications*, **152**, No. 1, pp. 88-108.

A. D. Bovopoulos and A. A. Lazar (1991), 'The effect of delayed feedback information on network performance', *Annals of Operations Res.* **36**, pp. 581-588.

E. B. N. Bui (1989), *Contrôle de l'allocation dynamique de trame dans un multiplexeur integrant voix et données*, TELECOM, Département Réseaux, Paris 89 E 005, June.

R. Cavazos-Cadena (1986), 'Finite-state approximations for denumerable state discounted Markov decision processes', *J. Applied Mathematics and Optimization*, **14**, pp. 27-47.

R. Cavazos-Cadena (1989), 'Weak conditions for the existence of optimal stationary policies in average cost Markov decision chains with unbounded cost', *Kybernetika*, **25**, 145-156.

R. Cavazos-Cadena (1992), 'Existence of optimal stationary policies in average Markov decision processes with a recurrent state', *Appl. Math. Optim.*, **26**, pp. 171-194.

R. Cavazos-Cadena and O. Hernández-Lerma (1992), 'Equivalence of Lyapunov stability criteria in a class of Markov decision processes', *Appl. Math. Optim.*, **26**, pp. 113-137.

R. Cavazos-Cadena and L. I. Sennott (1992), 'Comparing recent assumptions for the existence of average optimal stationary policies', *Operations Research Letters*, **11**, pp. 33-37.

K. L. Chung (1967), *Markov chains with stationary transition probabilities*, 2nd edition, Springer-Verlag, New York.

D. J. Daley and D. Vere-Jones (1988), *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York.

G. B. Dantzig, J. Folkman and N. Shapiro (1967), 'On the continuity of the minimum set of a continuous function', *J. Math. Anal. and Applications,* **17**, pp. 519-548.

G. T. De Ghellinck (1960), 'Les problèmes de décisions séquentielles', *Cahiers du Centre de Recherche Opérationelle*, **2**, pp. 161-179.

R. Dekker and A. Hordijk (1988), 'Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards', *Mathematics of Operations Research*, **13**, pp. 395-421.

R. Dekker, A. Hordijk and F. M. Spieksma (1994), 'On the relation between recurrence and ergodicity properties in denumerable Markov decision chains', *Math. Operat. Res.*, **19**, pp. 539-559.

E. V. Denardo (1970), 'On linear programming in a Markov decision problem', *Management Science*, **16**, pp. 281-288.

E. V. Denardo and B. L. Fox (1968), 'Multichain Markov renewal programs', *SIAM J. of Applied Math.*, **16**, pp. 468-487.

F. D'Epenoux (1960), 'Sur un problème de production et de stockage dans l'aléatoire', *Revue Française de Recherche Opérationelle*, **14**, pp. 3-16.

F. D'Epenoux (1963), 'A probabilistic production and inventory problem', *Management Science*, **10**, 98-108.

C. Derman (1970), *Finite State Markovian Decision Processes,* Academic Press, New York and London.

C. Derman and M. Klein (1965), 'Some remarks on finite horizon Markovian decision models', *Operations Research*, **13**, pp. 272-278.

C. Derman and R. E. Strauch (1966), 'On memoryless rules for controlling sequential control processes', *Ann. Math. Stat.*, **37**, pp. 276-278.

C. Derman and A. F. Veinott, Jr. (1972), 'Constrained Markov decision chains', *Management Science*, **19**, pp. 389-390.

J. L. Doob (1994), *Measure Theory*, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.

N. Dunford and J. T. Schwartz (1988), *Linear operators*, part I, John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore.

E. Dynkin and A. Yushkevich (1979), *Controlled Markov Processes*, Springer-Verlag, Berlin.

A. Federgruen (1979), 'Geometric convergence of value-iteration in multichain Markov decision problems', *Adv. Appl. Prob.,* **11**, pp. 188-217.

E. A. Feinberg (1982), 'Non-randomized Markov and semi-Markov strategies in dynamic programming', *Theor. Probab. and its Applications,* **27**, pp. 116-126.

E. A. Feinberg (1986), 'Sufficient classes of strategies in discrete dynamic programming I: decomposition of randomized strategies and embedded models', *SIAM Theory Probab. Appl.*, **31**, pp. 658-668.

E. A. Feinberg (1986), 'Sufficient classes of strategies in discrete dynamic programming', *Theory Probability Appl.,* **31**, pp. 658-668.

E. A. Feinberg (1991), 'Non-randomized strategies in stochastic decision processes', *Annals of Operations Research*, **29**, pp. 315-332.

E. A. Feinberg (1995), 'Constrained semi-Markov decision processes with average rewards', *ZOR – Methods and Models in Operations Research*, **39**, pp. 257-288.

E. A. Feinberg and D. J. Kim (1996), 'Bicriterion optimization of an M/G/1 queue with a removable server', *Probab. in the Eng. and Inf. Sciences*, **10**, pp. 57-73.

E. A. Feinberg and M. I. Reiman (1994), 'Optimality of randomized trunk reservation', *Probability in the Engineering and Informational Sciences*, **8**, pp. 463-489.

E. A. Feinberg E. and A. Shwartz (1995), 'Constrained Markov decision models with weighted discounted rewards', *Math. of Operations Research*, **20**, pp. 302-320.

E. A. Feinberg E. and A. Shwartz (1996), 'Constrained discounted dynamic programming', *Math. of Operations Research*, **21**, pp. 922-945.

E. A. Feinberg and I. Sonin (1983), 'Stationary and Markov policies in countable state dynamic programming', *Lecture Notes in Mathematics*, **1021**, pp. 111-129.

E. A. Feinberg and I. Sonin (1993), 'The existence of an equivalent stationary strategy in the case of discount factor equal one', unpublished draft.

E. A. Feinberg and I. Sonin (1995), 'Notes on equivalent stationary policies in Markov decision processes with total rewards', *ZOR – Methods and Models in Operations Research*, **44**, pp. 205-221.

A. V. Fiacco (1974), 'Convergence properties of local solutions of convex optimization problems', *J. Optim. Theory Appl.*, **13**, pp. 1-12.

J. A. Filar, L. C. M Kallenberg and H. M. Lee (1989), 'Variance-penalized Markov decision processes', *Math. of Operations Research*, **14**, pp. 147-161.

J. A. Filar and H. M. Lee (1985), 'Gain/variability tradeoffs in undiscounted Markov decision processes', *Proceedings of 24th Conference on Decision and Control IEEE*, pp. 1106-1112.

L. Fisher (1968), 'On recurrent denumerable decision processes', *Ann. Math. Stat.*, **39**, pp. 424-434.

L. Fisher and S. M. Ross (1968), 'An example in denumerable decision processes', *Ann. Math. Stat.,* **39**, pp. 674-675.

V. A. Gaitsgory and A. A. Pervozvanskii (1986), 'Perturbation theory for mathematical programming problems', *JOTA*, pp. 389-410.

K. Golabi, R. B. Kulkarni and G. B. Way (1982), 'A statewide Pavement Management System', *Interfaces*, **12**, pp. 5-21.

M. Haviv (1995), 'On constrained Markov decision processes', *OR Letters*, **19**, Issue 1, pp. 25-28.

W. R. Heilmann (1977), 'Generalized linear programming in Markovian decision problems', *Bonner Math. Schriften*, **98**, pp. 33-39.

W. R. Heilmann (1978), 'Solving stochastic dynamic programming prob-

lems by linear programming — an annotated bibliography', *Z. Oper. Res.*, **22**, pp. 43-53.

O. Hernández-Lerma (1986), 'Finite state approximations for denumerable multidimensional-state discounted Markov decision processes', *J. Mathematical Analysis and Applications*, **113**, pp. 382-389.

O. Hernández-Lerma (1989), *Adaptive Control of Markov Processes*, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo.

O. Hernández-Lerma and D. Hernández-Hernández (1994), 'Discounted cost Markov decision processes on Borel spaces: the linear programming formulation', *J. of Math. Anal. and Appl.*, **183**, pp. 335-351.

O. Hernández-Lerma and J. B. Lasserre (1994), 'Linear programming and average optimality on Borel spaces-unbounded costs', *SIAM J. Control and Optimization*, **32**, pp. 480-500.

O. Hernández-Lerma and J. B. Lasserre (1995), *Discrete-Time Markov Control Processes, Basic Optimality Criteria*, Springer-Verlag, New York, Berlin, Heidelberg.

K. Hinderer (1970), *Foundation of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Vol. 33, Lecture Notes in Operations Research and Mathematical Systems, Springer-Verlag, Berlin.

A. Hordijk (1977), *Dynamic Programming and Markov Potential Theory*, second edition, Mathematical Centre Tracts 51, Mathematisch Centrum, Amsterdam.

A. Hordijk and L. C. M. Kallenberg (1979), 'Linear programming and Markov decision chains', *Management Science*, **25**, pp. 352-362.

A. Hordijk and L. C. M. Kallenberg (1984), 'Constrained undiscounted stochastic dynamic programming', *Mathematics of Operations Research,* **9**, pp. 276-289.

A. Hordijk and J. B. Lasserre (1994), 'Linear programming formulation of MDPs in countable state space: the multichain case', *ZOR – Methods and Models in Operations Research*, **40**, pp. 91-108.

A. Hordijk and F. Spieksma (1989), 'Constrained admission control to a queuing system', *Advances of Applied Probability*, **21**, pp. 409-431.

R. Horn and C. R. Johnson (1985), *Matrix Analysis*, Cambridge Univ. Press.

M. T. Hsiao and A. A. Lazar (1991), 'Optimal decentralized flow control

of Markovian queueing networks with multiple controllers', *Performance Evaluation*, **13**, pp. 181-204.

Y. Huang and L. C. M. Kallenberg (1994), 'On finding optimal policies for Markov decision chains: A unifying framework for mean-variance tradeoffs', *Math. of Operations Research*, **19**, pp. 434-448.

D. Kadelka (1983), 'On randomized policies and mixtures of deterministic policies in dynamic programming', *Methods of Operations Research*, **46**, pp. 67-75.

L. C. M. Kallenberg (1983), *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tracts 148, Amsterdam.

L. C. M. Kallenberg (1994), 'Survey of linear programming for standard and nonstandard Markovian control problems, Part I: Theory', *ZOR – Methods and Models in Operations Research*, **40**, pp. 1-42.

P. Kanniappan and S. M. A. Sastry (1974), 'Uniform convergence of convex optimization problems', *J. Math. Anal. Appli.*, **96**, pp. 1-12.

H. Kawai (1987), 'A variance minimization problem for a Markov decision process', *European Journal of Operations Research*, **31**, pp. 140-145.

J. G. Kemeney, J. L. Snell and A. W. Knapp (1976), *Denumerable Markov Chains*, Springer-Verlag.

L. Kleinrock (1976), *Queuing systems, Volume I.* John Wiley, New York.

P. Kolesar (1970), 'A Markovian model for hospital admission and scheduling', *Management Science*, **16**, pp. 384-396.

G. M. Koole (1988), *Stochastische Dynamische Programmering met Bijvoorwaarden* (translation: Stochastic dynamic programming with additional constraints) Master thesis, Leiden University, The Netherlands.

Y. A. Korilis and A. Lazar (1995a), 'On the existence of equilibria in noncooperative optimal flow control', *J. of the Association for Computing Machinery*, **42**, No. 3, pp. 584-613.

Y. A. Korilis and A. Lazar (1995b), 'Why is flow control hard: optimality, fairness, partial and delayed information', preprint.

D. Krass (1989), *Contributions to the Theory and Applications of Markov Decision Processes*, Ph.D. thesis, Department of Mathematical Sciences, Johns Hopkins Univ., Baltimore, MD.

M. Krein and D. Milman (1940), 'On extreme points of regularly convex sets', *Studia Math.*, **9**, pp. 133-138.

N. Krylov (1985), 'Once more about the connection between elliptic operators and Ito's stochastic equations', *Statistics and Control of Stochastic Processes*, Steklov Seminar 1984 (Krylov N. *et al.*, Eds.), Optimization Software, New York, 69-101.

H. W. Kuhn (1953), 'Extensive games and the problem of information', *Ann. Math. Stud.*, **28**, pp. 193-216.

H. Kushner and J. Kleinman (1971), 'Mathematical programming and the control of Markov chains', *Internat. J. Control*, **13**, pp. 801-820.

J. B. Lasserre (1994), 'Average optimal stationary policies and linear programming in countable state Markov decision processes', *J. Math. Anal. Appl.*, **183**, pp. 233-249.

A. Lazar (1983), 'Optimal flow control of a class of queuing networks in equilibrium', *IEEE Transactions on Automatic Control*, **28**, pp. 1001-1007.

M. B. Lignota and J. Morgan (1992), 'Convergences of marginal functions with dependent constraints', *Optimization*, **23**, pp. 189-213.

R. Lucchetti and R. J. B Wets (1993), 'Convergence of minima of integral functionals, with applications to optimal control and stochastic optimization', *Statistics and Decisions*, **11**, pp. 69-84.

D.-J. Ma and A. M. Makowski (1988), 'A class of steering policies under a recurrence condition', *27th IEEE Conference on Decision and Control*, Austin, TX, December, pp. 1192-1197.

D.-J. Ma and A. M. Makowski (1992), 'A class of two-dimensional stochastic approximations and steering policies for Markov decision processes', *31st IEEE Conference on Decision and Control*, Tucson, Arizona, pp. 3344-3349.

D.-J. Ma, A. M. Makowski and A. Shwartz (1990), 'Stochastic approximations for finite state Markov chains', *Stochastic Processes and Their Applications*, **35**, pp. 27-45.

B. Maglaris and M. Schwartz (1982), 'Optimal fixed frame multiplexing in integrated line- and packet-switched communication networks', *IEEE Transactions on Information Theory*, **IT-28**, pp. 263-273.

A. M. Makowski and A. Shwartz (1987), 'Recurrence properties of a system of competing queues with applications', Research report EE Pub. No. 627, Technion, Haifa, Israel.

A. M. Makowski and A. Shwartz (1992), 'Stochastic approximations

and adaptive control of a discrete-time single server network with random routing', *SIAM J. Control and Optimization*, **30**, pp. 1476-1506.

A. S. Manne (1960), 'Linear programming and sequential decisions', *Management Science*, **6**, pp. 259-267.

S. Meyn and R. Tweedie (1994), *Markov Chains and Stochastic Stability*, Springer-Verlag, New York.

P. Nain and K. W. Ross (1986), 'Optimal priority assignment with hard constraint', *Transactions on Automatic Control,* **31**, pp. 883-888.

A. S. Nowak (1985), 'Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space', *JOTA*, **45**, pp. 592-602.

A. A. Pervozvanskii and V. A. Gaitsgory (1988), *Theory of Suboptimal Decision: Decomposition and Aggregation*, Kluwer Academic Publishers, Dordrecht.

A. B. Piunovskiy (1993), 'Control of random sequences in problems with constraints', *Theory Probab. Appl.*, **38**, No. 4, translated from Russian.

A. B. Piunovskiy (1994), 'Control of jump-like processes in constrained problems', *Avtomatika i Telemekhanika*, **4**, pp. 75-89. Translated into English in *Automation and Remote Control*, **55**, No. 4, 1994.

A. B. Piunovskiy (1995), 'Multicriteria control problems for stochastic jump processes', *Proceedings of 3rd European Control Conference*, Rome, Italy, September, pp. 492-495.

A. B. Piunovskiy (1996), 'A multicriteria model of optimal control of a stochastic linear system', *Automation and Remote Control* **57**, No. 6, Part 1, pp. 831-842.

A. B. Piunovskiy (1997a), *Optimal Control of Random Sequences in Problems with Constraints*, *Mathematics and its Applications*, Kluwer Academic Publishers, Dordrecht, Boston, London.

A. B. Piunovskiy (1997b), 'Optimal control of stochastic sequences in sequences with constraints', *Stochastic Analysis and Applications*, No. 2.

M. Puterman (1994), *Markov Decision Processes*, John Wiley & Sons, New York.

T. E. S. Raghavan and J. A. Filar (1991), 'Algorithms for stochastic pames – a survey', *ZOR – Methods and Models in Operations Research*, **35**, pp. 437-472.

D. Revuz (1975), *Markov Chains*, North-Holland, Amsterdam, The Netherlands.

R. T. Rockafellar (1989), *Conjugate Duality and Optimization*, Society for Industrial and Applied Mathematics, 2nd printing, Philadelphia.

K. W. Ross (1989), 'Randomized and past-dependent policies for Markov decision processes with multiple constraints', *Operations Research*, **37**, pp. 474-477.

K. W. Ross and B. Chen (1988), 'Optimal scheduling of interactive and non-interactive traffic in telecommunication systems', *IEEE Transactions on Automatic Control,* **33**, pp. 261-267.

K. Ross and R. Varadarajan (1989), 'Markov decision processes with sample path constraints: the communicating case', *Operations Research*, **37**, pp. 780-790.

K. Ross and R. Varadarajan (1991), 'Multichain Markov decision processes with a sample path constraint: a decomposition approach', *Math. of Operations Research*, **16**, pp. 195-207.

H. L. Royden (1988), *Real Analysis*, 3rd edition, Macmillan Publishing Company, New York.

M. Schäl (1975), 'Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal', *Z. Wahrscheinlichkeitstheorie und verw. Geb.,* **32**, pp. 179-196.

M. Schäl (1987), 'Estimation and control in discounted dynamic programming', *Stochastics*, **20**, pp. 51-71.

I. E. Schochetman (1990), 'Pointwise versions of the maximum theorem with applications to optimization', *Appl. Math. Lett.*, **3**, pp. 89-92.

I. E. Schochetman and R. L. Smith (1991), 'Convergence of selections with applications in optimization', *J. Math. Anal. Appl.*, **155**, pp. 278-242.

L. I. Sennott (1989), 'Average cost optimal stationary policies in average cost Markov decision processes', *Operations Research*, **37**, pp. 626-633.

L. I. Sennott (1991), 'Constrained discounted Markov decision chains', *Probability in the Engineering and Informational Sciences*, **5**, pp. 463-475.

L. I. Sennott (1993), 'Constrained average cost Markov decision chains', *Probability in the Engineering and Informational Sciences*, **7**, pp. 69-83.

L. I. Sennott (1997), 'On computing average optimal policies with appli-

cation to routing to parallel queues', *Mathematical Methods of Operations Research*, **45**, pp. 45-62.

L. S. Shapley (1953), 'Stochastic games', *Proceedings Nat. Acad. of Science USA*, **39**, pp. 1095-1100.

N. Shimkin (1994), 'Stochastic games with average cost constraints', *Annals of the International Society of Dynamic Games, Vol. 1: Advances in Dynamic Games and Applications*, Eds. T. Basar and A. Haurie, Birkhauser, Boston.

M. J. Sobel (1985), 'Maximal mean/standard deviation ratio in undiscounted MDP', *OR Letters*, **4**, pp. 157-159.

M. J. Sobel (1994), 'Mean-variance tradeoffs in an undiscounted MDP', *Operations Research*, **42**, pp. 175-188.

F. M. Spieksma (1990), *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, Ph.D. thesis, University of Leiden, The Netherlands.

R. Sznadjer and J. A. Filar (1992), 'Some comments on a theorem of Hardy and Littlewood', *J. Optim. Theory Appl.*, **75**, pp. 210-218.

L. C. Thomas and D. Stengos (1985), 'Finite state approximation algorithms for average cost denumerable state Markov decision processes', *OR Spectrum*, **7**, pp. 27-37.

M. Tidball and E. Altman (1996a), 'Approximations in dynamic zero-sum games, I', *SIAM J. Control and Optimization*, **34**, No. 1, pp. 311-328.

M. Tidball, O. Pourtallier and E. Altman (1996b), 'Continuity of optimal values and solutions of convex optimization, and constrained control of Markov chains', submitted to *SIAM J. Control and Optimization*.

M. Tidball, O. Pourtallier and E. Altman (1997), 'Approximations in dynamic zero-sum games, II', *SIAM J. Control and Optimization*, **35**, pp. 2101-2117.

F. Vakil and A. A. Lazar (1987), 'Flow control protocols for integrated networks with partially observed voice traffic', *IEEE Transactions on Automatic Control*, **AC-32**, pp. 2-14.

J. Van Der Wal (1981a), *Stochastic Dynamic Programming*, Mathematical Centre Tract 139, Mathematisch Centrum, Amsterdam.

J. Van Der Wal (1981b), 'On stationary strategies', Eindhoven Univ. of Technology, Dept. of Math., Memorandum-COSOR 81-14, 1981.

J. Wessels (1977), 'Markov Games with unbounded rewards', *Dynamische Optimierung*, M. Schäl (Editor) Bonner Mathematische Schriften, Nr. 98, Bonn.

D. J. White (1980), 'Finite state approximations for denumerable state infinite horizon discounted Markov decision Processes', *J. Mathematical Analysis and Applications*, **74**, pp. 292-295.

D. J. White (1982), 'Finite state approximations for denumerable state infinite horizon discounted Markov decision processes with unbounded rewards', *J. Mathematical Analysis and Applications*, **86**, pp. 292-306.

D. J. White (1987), 'Utility, probabilistic constraints, mean variance of discounted rewards in Markov decision processes', *OR Spectrum*, **9**, pp. 13-22.

D. J. White (1994), 'A mathematical programming approach to a problem in variance penalized Markov decision processes', *OR Spectrum*, **15**, pp. 225-230.

W. Whitt (1978), 'Approximations of dynamic programs, I', *Mathematics of Operations Research*, **3**, No. 3, pp. 231-243.

W. Whitt (1980), 'Representation and approximation of noncooperative sequential games', *SIAM J. Control and Opt.*, **18**, No. 1, pp. 33-43.

C. V. Winden and R. Dekker (1994), 'Markov Decision Models for Building Maintenance: A Feasibility Study', Report 9473/A, ERASMUS University Rotterdam, The Netherlands.

D. Williams (1992), *Probability and Martingales*, Cambridge University Press, Cambridge.

A. A. Yushkevich (1973), 'On a class of strategies in general Markov decision models', *Theory Probab. Appl.* **18**, pp. 777-779.

CHAPTER 19

List of Symbols and Notation

$1\{\text{condition}\}$:= the indicator function which equals one if the condition holds, and is zero otherwise.
$\langle q_1, q_2 \rangle$:= scalar product between two vectors.
$q_1 \leq q_2$:= componentwise ordering between two vectors.
$B_1 \prec B_2$:= an ordering between sets B_1 and B_2, Section 8.1.
$B_1 \propto B_2$:= an ordering between sets B_1 and B_2, Section 11.1.
$\|\cdot\|_\mu$:= norm, defined in Section 7.7.

a.s. – almost sure.

A^c := the complement of a set A.

a, A_t, A := actions, actions at time t, action space, Sections 2.1, 6.1.

B, $\mathcal{B}$:= a Borel set, set of Borel subsets, Section 6.1.

$\underline{b}, \overline{b}, b$:= bounds on the costs, (11.1), (7.36), (16.7).

$\overline{B}$:= upper bound on $C(x, u)$, see (13.1).

B1 – assumption on the ergodic structure, beginning of Chapter 11.

B2, B2(u), B2* – assumptions related to tightness of the occupation measures, end of Section 11.1.

B3, B3(u), B3* – assumptions related to uniform integrability of the occupation measures, Section 11.3.

c := immediate cost, Section 2.1, 6.1.

$\overline{co}$:= the closed convex hull of a set, Section 8.1.

$C^n(\beta, u)$, $C^n_\alpha(\beta, u)$, $C_{tc}(\beta, u)$, $C_\alpha(\beta, u)$, $C_{ea}(\beta, u)$:= finite horizon expected cost, finite horizon discounted expected cost, total expected cost, total discounted expected cost, expected average cost (Sections 2.2, 6.2).

$\mathcal{C}(\rho) := c \cdot \rho$ - linear expressions in the primal LP.

d^k:= immediate costs, Sections 2.2, 6.2.

$D^{k,n}(\beta,u)$, $D^{k,n}_{\alpha}(\beta,u)$, $D(\beta,u)$, $D^k_{\alpha}(\beta,u)$, $D^k_{ea}(\beta,u)$, $D^k_{av}(\beta,u)$:= the other costs (defined together with the corresponding C).

$\mathcal{D}^k(\rho) := d^k \cdot \rho$, the linear expressions in the primal LP.

$\mathbf{DP}_i$ a dual linear programs related to **COP**.

$E^u_{\beta} :=$ expectation related to initial distribution β and policy u.

f_{tc}, f_{α}, $f^t_{ea} :=$ expected occupation measure for total expected cost, total expected discounted cost, expected average costs until time t (Sections 8.1, 10.2 and 11.1).
$F_{ea} :=$ limit set of occupation measures for the expected average cost.
F^{μ}, $\overline{\mathsf{F}^{\mu}}$:= sets of μ-bounded functions (defined in Section 7.7).

g – stationary deterministic policy, Section 2.1.

$G, \mathcal{G}$:= a set together with its σ-algebra, Section 6.1.

h_{α}:= difference between some optimal discounted costs, Section 12.2.
$\underline{h}$:= lower bound on h_{α} (Section 12.2).
h_t, H_t, $\mathbf{H}_t$:= history until t, the space of histories, Sections 2.1, 6.1.

$\mathbf{I}$, I_t:= used as extra randomizing mechanism for policies U_R (Section 6.6).

J^{λ}, j^{λ}:= the Lagrangian, corresponding to the immediate costs and to the cost criterion, respectively (Sections 9.4, 12.6).

k, K:= indices (of constraints), number of constraints.
$\mathcal{K}, \kappa$:= set of state action pairs, a generic element; Sections 2.1, 6.1.
$\mathbb{K}$:= Borel σ-algebra of $\mathcal{K}$.

$\mathbf{L}, \mathcal{L}, \mathbf{L}^{\alpha}$:= set of occupation measures for total cost, see (8.1), expected average cost, see (11.5), discounted cost, see (10.1). In particular, when they have the subscripts M, S, D, they correspond to occupation measures obtained by the Markov, stationary and stationary deterministic policies, respectively.
$\mathbf{LP}_i$ – Primal Linear Programs which are equivalent to **COP**.

$\min B$:= the set of minimal elements in a set B, Section 8.1.
$\overline{m}(\cdot)$:= upper bound on h_{α} (Section 12.2).
$\mathbf{M}$:= set of achievable costs for total cost, see (9.21), expected average cost, see (12.26). In particular, when they have the subscripts M, S, D they

correspond to costs obtained by the Markov, stationary and stationary deterministic policies, respectively.

$\mathcal{M}$:= a set, used in the definition of the total expected cost, Section 6.1.

M^{μ}:= the set of measures q with $E^{q}\mu$ finite (Section 7.7).

$\overline{\mathsf{M}^{\mu}}$:= defined in Section 7.7.

$M_1(G)$, $M(G)$, $\overline{M}(G)$:= set of probability measures over a set G, the set of measures over G. Mixed strategies over $G \subset U$.

$M_u(x)$:= total expected hitting time under policy u from state x, defined below (6.3).

$\hat{M}(\beta, u), \hat{M}(\beta)$:= defined in (7.11) (see also (7.18)) and below (7.18), respectively.

M1,M2, ..., M9 – equivalent properties of uniform Lyapunov functions, Section 7.4.

N1,N2, ..., N6 – equivalent properties of uniform Lyapunov functions, Section 7.5.

$p^{u}_{\beta}(t;\mathcal{X}) := P^{u}_{\beta}(X_t \in \mathcal{X}, T > t)$ and $p^{u}_{\beta}(t;\mathcal{K}) := P^{u}_{\beta}((X_t, A_t) \in \mathcal{K}, T > t)$, Section 6.1.

${}_{\mathcal{M}}P$:= the Taboo matrix, see Section 6.2.

$\mathcal{P}$, P^{u}_{β}:= transition probabilities; probability generated by initial distribution β and policy u (6.6).

$\mathbf{Q}_{tc}$, $\mathbf{Q}_{ea}$, $\mathbf{Q}^{\alpha}$:= feasible sets for the primal LPs, see (8.2), (8.3), (11.5), (10.2).

Q – matrix, Section 7.7, Definition 8.2.

$q, \hat{q}$:= probability distribution; mixed policy with parameter q, Section 6.1.

$\mathcal{S}1$-$\mathcal{S}3$ – assumptions introduced in Section 12.2.

($\mathbf{S}1$)-($\mathbf{S}5$) – conditions defined in Section 13.2 for the convergence of values and policies.

s, t, n:= generic notation for time or horizon length.

$T, \mathcal{T}, T_{\mathcal{M}}, T^{[s]}$:= hitting times, and expected hitting times, see (6.3), (12.11), (11.28), Example 9.1, Section 16.5.

$u, v, U, \mathcal{U}, U_M, U_S, U_D, U_R$:= policies, set of behavioral policies, mixed stationary-deterministic policies, Markov, stationary, stationary-deterministic policies (Section 2.1), and policies with extra randomization (Section 6.6).

$\mathbf{V}_{tc}$, $\mathbf{V}_{ea}$:= sets of achievable costs, see (9.22), (12.27).
$V = (V_1, \ldots, V_K)$:= constants, appear in the constraints, Section 2.1.

w:= stationary (randomized) policy, Section 2.1.
$W_\alpha(u; x)$:= the total cost from x to y, see (12.11).
w.p.1 – with probability one.

x, y, z, X_t, $\mathbf{X}$:= states, state at time t, state space, Section 2.1.

α:= discount factor.

β:= initial distribution.

γ:= used as a probability measure (often used as parameter for mixed policies).
$\hat{\gamma}$:= the mixed policy having parameter γ.

δ:= Dirac measure, Section 2.1.

λ:= Lagrange multiplier.

$\mu, \nu, \xi, \tilde{\xi}, \sigma$:= involved in the definition of MDPs with uniform Lyapunov functions (Definitions 7.4 and 7.5), contracting MDPs, Section 7.7, and Definitions 11.4 and 11.5 of uniform geometric recurrence and ergodicity.

σ:= a constant in the Definition 11.5 of uniform geometric ergodicity.

ρ:= decision variables in the primal LP; ρ corresponds to the occupation measure (8.2) and Section 8.7, (11.5) and Section 11.5.

(ϕ, ψ):= decision variables in the dual LP. ϕ corresponds to the value for the finite horizon case and ψ to the relative cost.

$\pi(\cdot)$:= steady-sate probabilities. $\pi(g)$:= steady-state probabilities corresponding to a stationary policy g.

Δ_f:= a constant used for the decomposition of the occupation measures corresponding to the expected average cost, see Section 11.1.

Θ:= value of the dual program.

Index