INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

CONSTRAINED MARKOV DECISION PROCESSES

Eitan ALTMAN

N° 2574

Mai 1995

PROGRAMME 1

Rapport de recherche

ISSN 0249-6399

CONSTRAINED MARKOV DECISION PROCESSES

Eitan ALTMAN

Programme 1 — Architectures parallèles, bases de données, réseaux et systèmes distribués
Projet MISTRAL

Rapport de recherche n° 2574 — Mai 1995 — 115 pages

Abstract:
This report presents a unified approach for the study of constrained Markov decision processes with a countable state space and unbounded costs. We consider a single controller having several objectives; it is desirable to design a controller that minimize one of cost objective, subject to inequality constraints on other cost objectives. The objectives that we study are both the expected average cost, as well as the expected total cost (of which the discounted cost is a special case). We provide two frameworks: the case were costs are bounded below, as well as the contracting framework. We characterize the set of achievable expected occupation measures as well as performance vectors. This allows us to reduce the original control dynamic problem into an infinite Linear Programming. We present a Lagrangian approach that enables us to obtain sensitivity analysis. In particular, we obtain asymptotical results for the constrained control problem: convergence of both the value and the policies in the time horizon and in the discount factor. Finally, we present and several state truncation algorithms that enable to approximate the solution of the original control problem via finite linear programs.

(Résumé : tsvp)

Unité de recherche INRIA Sophia-Antipolis
2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex (France)
Téléphone : (33) 93 65 77 77 – Télécopie : (33) 93 65 77 65

PROCESSUS DE DECISION MARKOVIENS SOUS CONTRAINTES

Résumé :

Dans ce rapport, nous présentons une approche unifiée pour l'étude des processus de décision Markoviens sous contraintes, à espace d'état dénombrable et avec des coûts non bornés. Nous considérons un seul contrôleur ayant plusieurs objectifs; son but est de concevoir une politique qui minimise un objectif sous des contraintes d'inégalité sur les autres. Comme objectifs, nous étudions le coût moyen, ainsi que le coût total (dont le coût actualisé est un cas particulier). Nous considérons deux cadres possibles: le cas où les coûts sont bornés inférieurement, ainsi que le cas contractif. Nous caractérisons les ensembles de mesures d'occupations atteignables, ainsi que l'ensemble des mesures de performances atteignables. Cela nous permet de réduire le problème original de contrôle dynamique à une programmation linéaire infinie. Nous présentons une méthode de Lagrangien permettant d'obtenir l'analyse de sensibilité. En particulier, nous obtenons des résultats sur le comportement asymptotique du problème de contrôle: la convergence des valeurs et des politiques quand l'horizon converge vers l'infini, ou quand le facteur d'actualisation converge. Finalement, nous présentons plusieurs algorithmes de troncation de l'espace d'état, permettant l'approximation des solutions du problème de contrôle à l'aide d'une programmation linéaire finie.

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 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 bound. Posed in the above way, we may consider our control problem can be viewed as a constrained optimization problem over a given class of policy.

By specifying to control problems rather than optimization problems, we have in mind models of dynamic systems, where decision actions 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 be 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, in turn, determine the different costs, or objectives.

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. We describe especially problems in telecommunications networks, which was one of the richest area of applications of constrained Markov decision processes (CMDPs).

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 objectives are the transmission delay, the throughputs, probabilities of losses of packets (that are due to 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 etc... Different types of traffic defer from each other both 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. Hard 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 sec. When the delay is higher than this limit, it becomes quickly untolerable. For non interactive file transfer, we often wish to minimize delays or to maximize the throughputs.

Controllers of telecommunication systems have often been developed using heuristics and experience. However, there has been a tremendous research effort to solve analytically such problems. 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 is still being 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). Hordijk and Spieksma (1989) considered the problem of Lazar (1983) as well as other admission control problems within the framework of MDPs, and discovered that for some problems, optimal policies are not of a limit type (so called "thinning policies" were shown to be optimal under some conditions).

(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 minized, whereas for some other traffic types, a weighted sum of average delays should be bounded by some given limit. This research stimulated much more research, 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 structure of optimal policies for these types of models is some randomization or 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 into a multi-channel system with finite waiting space. They established the optimality of a randomized trunk reservation policy.

Other problems in telecommunications which have been solved by constrained MDPs are Maglaris and Schwartz (1982), Beutler and Ross (1986), and Bui (1989).

Constrained MDPs had an important impact in many other areas of applications. In Kolesar (1970), a problem of hospital admission scheduling is considered.

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 7,400-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.

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 the all 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 these kind of CMDPs. The first, based on a Linear Program, 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 a 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 one to one correspondence between the optimal solutions of the LP and the optimal policies of the CMDP. This method is the most efficient 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 inattractive for practical problems. The number of calculations is prohibitive" (p. 142). An alternative efficient way for obtaining optimal policies from the LP for the average cost was obtained by Krass (1989) in his thesis. In Chapters 3 and 5 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, 1994,1995,1996).

A second method was introduced by Beutler and Ross (1985,1986) for the case of a single constraint, and is based on a Lagrange approach. It allowed 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 Lagrange techniques for several constraints is quite recent (see e.g. Arapostathis et al., 1993, Piunovskiy (1994), and Altman and Spieksma, 1995), and was not much exploited.

A third method, based on a 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) between 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 the finite state and action spaces, where the mixing is done by some equivalent initial randomization between stationary deterministic policies. These approaches requires in general, a huge number of decision variables. However, there are special applications where this LP can be extremely efficient, and used even for problems with infinite state space (see Altman and Shwartz, 1989), in case where one can eliminate a-priori many sub-optimal stationary deterministic policies.

It turns out that deep connections exist between the three solution methods. Understanding these connections enables us to obtain an elegant complete theory for CMDPs. It also enables us to generalize the second approach to several constraints, and to reduce the complexity of other methods. Finally, it allows to obtain many asymptotic results on convergence of the values and policies of some sequence of CMDPs to a limit one. In particular, convergence in the discount factor, in the horizon, finite state approximations etc... (we present these in Chapters 7 and 8).

1.3 Other types of CMDPs

The type of cost criteria and solution approaches surveyed in the previous Section are those the 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. Such CMDPs are extremely hard to solve, and 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 after some fixed time) and requires no more than $K+1$ randomizations. This extends the results by Ross (1989) and Borkar (1994).

Ross and Varadarajan (1989,1981) 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 (1989,1981) using again a LP approach. An interesting feature of this formulation is that ϵ-optimal policies exist (in the finite state and action case) even under the general multi-chain ergodic structure. This is in contrast with 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 other alternative ways to make the costs more sensitive to deviations from its expectation. One way to achieve this goal is to have some *additional cost related to the variance*. Sobel (1985) proposed to maximize the mean-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 and Kallenberg (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).

Other type of constraints, 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. The case of two controllers ("players") with conflicting objectives was solved by Shimkin, 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 to introduce 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) based on ideas on sensitivity analysis of Linear programs. When there is only a single constraint, techniques based on stochastic approximations can be used to solving the adaptive MDP. This approach was ised by Makowski and Shwartz (1992), Ma et al. (1992), Ma and Makowski (1992).

1.4 The convex analytical approach and occupation measures

We focus in this monograph on two types of cost criteria: the total cost, of which the discounted cost is a special case, and the expected average cost.

For the total cost problem we consider three types of MDPs: the transient MDPs, for which the total expected time spent in each state is finite under any policy, the absorbing MDPs, for which the total expected "life time" of the system is finite under any policy, and contracting MDPs. All three types of MDPs are equivalent for the finite state space, as was shown in Kallenberg (1983); this is however not the case in the countable state space.

Our analysis approach is based on the the properties of the set of occupation measures achievable by different classes of policies. A key property for the analysis is convexity and compactness properties of these sets. We present this analysis in the beginning of Chapter 3, for the total cost, and in the beginning of Chapter 5, 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 actions space). 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 Krylov (1985) who studied controlled diffusion processes, Borkar (1988,1990), Altman and Shwartz (1991a), Altman (1994,1996), Spieksma (1990), and Feinberg and Sonin (1993,1995).

For the different cost criteria, the objectives turn 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 the induced conditional occupation measures. We present these LPs and establish their equivalence to the original control problem in the end of of Chapter 3, (the total cost), and of Chapter 5, (the expected average cost). This approach goes back to Derman (1970) and further developed by Derman and Veinott (1972) and Kallenberg in his thesis (see Kallenberg, 1983, and Hordijk & Kallenberg, 1984). Its extension for the infinite state case is due to Altman and Shwartz (1991a) (the expected average cost) and Altman (1994,1996) (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 be able to restrict to them for the search of optimal policies. Under fairly general conditions, the problem of whether a subclass of policies is dominant can be reduced 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.

An important question is whether the stationary policies, is complete. 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 (1994) for the case of finite state and actions), or, in the infinite case, due to non-tightness (see Borkar, 1990, Ch. 5, Altman and Shwartz (1991a), and Spieksma 1990). However, under some conditions on the ergodic structure, the set of stationary policies turns to be "weakly" complete, i.e. to be proportional to those obtained by any other policy. This property, together with some growth condition on the costs, imply that the stationary policies are dominant. These results, obtained by Borkar (1990), Altman and Shwartz (1991a) are presented in Chapter 5.

For the total cost, for contracting and absorbing MDPs, both the stationary policies as well as the mixed stationary-deterministic policies achieve the same occupation measures. Surprisingly, this result turns not to hold for the more general transient MDPs. Indeed, counter-examples have been presented recently by Feinberg and Sonin (1995). However, the set of stationary policies turns out to have the following property. For any occupation measure achievable by some policy u, there are a stationary policy and some mixed stationary-deterministic policy that achieve an occupation measure that is smaller than or equal to that one achieved by u. These results, obtained in Altman (1996), are presented in Chapter 3.

1.5 The 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 an extensive description and 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). A most 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 the theory of infinite dimension LPs (due to 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) 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 equation. 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* "superharmonic functions": 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 the dual to the one obtained using the convex analytical approach of occupation measures (which we described in the previous Section). This approach is in 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 Lagrange approach, and then applying some minmax Theorem. Indeed, the Lagrange approach allows us to transform a constrained control problem to an equivalent minmax non-constrained control problem. If a saddle point property is shown to hold, then the problem is transformed to 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 finite case.

In Chapters 4 and 6 we describe this approach for obtaining the dual LP (for the total cost and the expected average cost, respectively).

The Lagrangian approach turns to be not only a tool for obtaining a 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 7 Theorems for approximations of the value and policies for CMDPs, which we apply for the study of the convergence in the discount factor (especially, in the neighborhood of 1), and in the horizon (Chapter 7,) as well as the study of state truncation techniques (Chapter 8). In particular, it allows to obtain an estimate of the approximation error. All these results are obtained in Chapter 8 for the contracting framework. An alternative approach for approximations is illustrated in Chapter 4, where state truncation is used for computing the value of the CMDP in the case of immediate costs that are bounded below.

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 3 and 5, so that the restriction is without loss of optimality. The decision variables here are the measures over all stationary deterministic policies. The 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 Tidbal and Altman (1995) to obtain convergence of the values and policies of discounted CMDP to those of expected average MDPs. 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 in the end of Chapters 4 and 6.

CHAPTER 2

Markov decision processes

2.1 The model

Markov decision processes (MDPs), also known as controlled Markov chains, constitute a basic framework for controlling dynamically systems, which evolve in a stochastic way. We focus on discrete time models; we thus observe the system at times $t = 1, 2, ..., T$. T 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 a quantity called the "state" of the system, as well as the control action. Here we mean by behavior both some immediate costs and the evolution, i.e. the probabilities of transitions to different states. MDPs are generalization of (non-controlled) Markov chains, and many useful properties of Markov chains cary on to controlled Markov chains. This makes MDPs a useful tool for stochastic control models. The models that we study in this book are special in the fact 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 countable state space. Generic notation for states will be x, y, z.
- A is a metric set of actions. We denote by $\mathsf{A}(x)$ 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 sets $\mathbb{K}$. A generic notation for an action will be a.
- $\mathcal{P}$ are the transition probabilities; thus $\mathcal{P}_{xay}$ is the probability to move 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 functional which we shall minimize.
- $d : \mathcal{K} \to \mathbb{R}^K$ is a K-dimensional vector of immediate costs, related to some 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 following assumption:

$$c(x, a) \text{ and } d^k(x, a), k = 1, ..., K, \text{ are continuous on } \mathsf{A}(x) \tag{2.1}$$

$$\begin{array}{l}\text{The transition probabilities are continuous, i.e. if } a(n) \to a \\ \text{in } \mathsf{A}(x) \text{ then} \\ \quad \lim_{n\to\infty} \sum_{y\in\mathbf{X}} |\mathcal{P}_{xa(n)y} - \mathcal{P}_{xay}| = 0.\end{array} \tag{2.2}$$

Assumption (2.1) can also be rephrased as: c and d are continuous over $\mathcal{K}$.

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 ones, 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 one may restrict 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 have more information on the system:
(i) Whenever the transition probabilities are not only zero or ones,
(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 modesl 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 control 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: $h_t = (x_1, a_1, ..., x_{t-1}, a_{t-1}, x_t)$. Let $\mathbf{H}_t$ be the set of all possible histories of length t. A policy u is a sequence $u = (u_1, u_2, ...)$ (with T elements) where $u_t : \mathbf{H}_t \to M_1(\mathsf{A})$ is a measurable function that assigns to any history of length t, a probability measure over the set of actions. ($M_1(G)$ stands for the set of probability measures over a set G endowed with the topology of weak convergence of measures). f the history h_t was observed at time t, then the controller chooses an action within $\mathcal{A}$ with probability $u_t(\mathcal{A}|h_t)$, where $\mathcal{A}$ is any subset in $\mathsf{A}(x_t)$. The class of all policies defined as above is denoted by U.

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

- U_M:= Markov policies, for which for any t, u_t is only a function of x_t (and not of the whole history). We may identify a Markov policy with a map $u : \{(x,t) \to M_1(\mathsf{A}(x)), x \in \mathbf{X}, t = 1, 2, ..., T\}$.

- U_S:= stationary policies, which are 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 a map $w : \{x \to M_1(\mathsf{A}(x)), x \in \mathbf{X}\}$. Under any stationary policy w, the state process becomes a Markov chain with transition probabilities $P_{xy}(w) = \int \mathcal{P}_{xay} w_x(da)$. If a stationary probability for the Markov chain exists, it is denoted by $\pi(w)$.

- U_D:= stationary deterministic policies, which are 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 $g : \mathbf{X} \to \mathsf{A}$.

It will often be useful to extend the definition of a policy $u = (u_1, u_2, ...)$ so as to allow u_t to depend not only on h_t, but also on some initial randomizing mechanism. In particular, for any class of policies $G \subset U$, we define $\overline{M}(G)$ to be the class of mixed policies generated by G, 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 sometime 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 suchσ-algebras to Section 2.2.

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). For mixed policies, the construction is done similarly. Moreover, for any mixed policy, the probability space for the state and action processes can be chosen to be 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 (i.e. $\beta = \delta_x$), we shall use the notation P^u_x instead of P^u_β. Denote $p^u_\beta(t; x) = P^u_\beta(X_t = x)$ and $p^u_\beta(t; x, \mathcal{A}) = P^u_\beta(X_t = x, A_t \in \mathcal{A})$, $\mathcal{A} \subset \mathsf{A}(x)$. We have for all $\beta \in M_1(\mathbf{X})$ and policies u,

$$p^u_\beta(t; x) = p^u_\beta(t; x, \mathsf{A}(x)),$$

and for $t > 1$,

$$p^u_\beta(t; x) = \sum_y \int_{\mathsf{A}(y)} p^u_\beta(t-1; y, da) \mathcal{P}_{yax} = \int_{\mathcal{K}} p^u_\beta(t-1; d\kappa) \mathcal{P}_{\kappa x}. \tag{2.3}$$

Next, we define the cost criteria that will appear in the constrained control problem. For any policy u and initial distribution β, the finite horizon cost for a horizon T is defined as

$$C^T(\beta,u)=\sum_{t=1}^{T}E^u_\beta c(X_t,A_t). \tag{2.4}$$

The total cost is defined as

$$C_{tc}(\beta,u)=\sum_{t=1}^{\infty}E^u_\beta c(X_t,A_t). \tag{2.5}$$

For a fixed discount factor α, $0<\alpha<1$, define the discounted cost (finite and infinite horizon) by

$$C^T_\alpha(\beta,u)=(1-\alpha)\sum_{t=1}^{T}\alpha^{t-1}E^u_\beta c(X_t,A_t), \tag{2.6}$$

$$C_\alpha(\beta,u)=(1-\alpha)\sum_{t=1}^{\infty}\alpha^{t-1}E^u_\beta c(X_t,A_t). \tag{2.7}$$

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

$$C^T_{ea}(\beta,u)=\frac{C^T(\beta,u)}{T},\qquad C_{ea}(\beta,u)=\overline{\lim_{T\to\infty}}\,C^T_{ea}(\beta,u). \tag{2.8}$$

We shall also consider the (sample) average cost (which is a random variable):

$$C_{av}(\beta,u)=\overline{\lim_{T\to\infty}}\sum_{t=1}^{T}c(X_t,A_t). \tag{2.9}$$

The costs functions related to the immediate costs d are defined similarly; e.g., the finite horizon cost related to d^k, $k=1,...,K$, is $D^{T,k}(\beta,u)=\sum_{t=1}^T E^u_\beta d^k(X_t,A_t)$. 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 $C(x,u)$.

For a fixed vector $V=(V_1,...,V_K)$ 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. (2.4)-(2.8). 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 componentwize ordering, i.e. $q_1(j)\le q_2(j)$, $j=1,...,K$. The set of policies satisfying the constraints are called feasible. Let $C(\beta)$ be the value of the above problem, with the obvious notation relating to the different costs (2.4)-(2.8) (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.

2.2 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 Subsection 2.1) and "convergence of policies", as will be discussed in Chapter 7.

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 $h_t = (x_1, a_1, ..., x_{t-1}, a_{t-1}, x_t)$, 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. Therefore, Tychonov's Theorem implies that $\prod_{h \in \mathbf{H}} M_1(\mathsf{A}(x(h)))$ is also 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 $\overline{M}(U)$ is now identified with the set of probability measures on the space $(U, \mathcal{G}_U)$. This class of policies is known as the class of *non-behavioral* 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 I as range, and a countable product of compact sets (of measures) $\prod_{i \in I} M_1(\mathsf{A}_I)$ 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 strategies $\overline{M}(U_M)$ 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 $\overline{M}(U_D) = \mathcal{U}$.

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.

2.3 Dominating policies

The class of Markov policies turns out to be very 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). If $u \in U_M$ then $p_\beta^u(t;\cdot)$ can be written in the following vector notation:

$$p_\beta^u(t) = \beta P(u_1) P(u_2) ... P(u_{t-1}),$$

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

Theorem 2.1 *(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,

$$p_\beta^{\hat\gamma}(t;\cdot,\cdot) = p_\beta^v(t;\cdot,\cdot) \tag{2.10}$$

(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, (2.10) holds.

Proof: The proof of (i) is related to Dynkin and Yushkevich (1979) Chapter 3, Sec. 5. We write $p_\beta^{\hat\gamma}$ in an integral form:

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

Define v to be the Markov policy given by

$$v_t(\mathcal{A}|x) := \frac{\int \gamma(du) P_\beta^u(X_t = x, A_t \in \mathcal{A})}{\int \gamma(du) P_\beta^u(X_t = x)} \tag{2.11}$$

for all integers t, states x and $\mathcal{A} \subset \mathsf{A}(x)$, for which the denominator is nonzero. When it is zero, define $v_t(\cdot|x)$ to be an arbitrary probability measure over $\mathsf{A}(x)$. The proof follows by induction. (2.10) clearly holds for $t=1$, since for any policy $u \in U_M$,

$$P_\beta^u(X_1 = x, A_1 \in \mathcal{A}) = \beta(x) u_1(\mathcal{A}|x),$$

and $\int \gamma(du) P(X_1 = x) = \beta(x)$; this implies

$$P_\beta^v(X_1 = x, A_1 \in \mathcal{A}) = \beta(x) v_1(\mathcal{A}|x) = \int \gamma(du) P_\beta^u(X_1 = x, A_1 \in \mathcal{A}).$$

Assume that (2.10) holds for some t, i.e.

$$\begin{aligned} & \int \gamma(du) P_\beta^u(X_t = x, A_t \in \mathcal{A}) && (2.12) \\ = \quad & P_\beta^v(X_t = x, A_t \in \mathcal{A}) = [\beta P(v_1) P(v_2) ... P(v_{t-1})]_x v_t(\mathcal{A}|x). && (2.13) \end{aligned}$$

We show first that

$$\int \gamma(du) P^u_\beta(X_{t+1} = x) = [\beta P(v_1)P(v_2)...P(v_t)]_x. \tag{2.14}$$

Since

$$P^u_\beta(X_{t+1} = x | X_t = y, A_t = a) = \mathcal{P}_{yax}, \qquad P^u_\beta - a.s.$$

for all $u \in U_M$, we obtain by conditioning on X_t, A_t and by (2.12), that the left-hand side of (2.14) equals

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

This implies (2.14). Combining now (2.14) with (2.11), we get

$$\begin{aligned} \int \gamma(du) P^u_\beta(X_{t+1} = x, A_{t+1} \in \mathcal{A}) &= v_t(\mathcal{A}|x) \int \gamma(du) P^u_\beta(X_{t+1} = x) \\ &= [\beta P(v_1)P(v_2)...P(v_t)]_x v_t(\mathcal{A}|x) = P^v_\beta(X_{t+1} = x, A_{t+1} \in \mathcal{A}). \end{aligned}$$

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

Remark 2.1 *(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 some 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 by Feinberg (1982, Thm. 1) and Kadelka.

As a corollary, we may conclude that for any policy $u \in U$, there exists some $v(u) \in U_M$ such that for all t,

$$p^{v(u)}_\beta(t; \cdot, \cdot) = p^u_\beta(t; \cdot, \cdot). \tag{2.15}$$

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

The above property of Markov policies implies that we can always restrict to Markov policies, for any type of cost criterion among those introduced in the beginning of this Chapter, (and more generaly, for any cost criteria that depends on the probability space only through the marginal distributions), since these perform as well as any other class of policies. This motivates the following

Definition 2.1 *(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 above 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*

$$C(\beta, \overline{u}) \le C(\beta, u), \qquad \text{and} \qquad D(\beta, \overline{u}) \le V. \tag{2.16}$$

Here C, D, and **COP** *stand for any one of the cost criteria previously defined. When (2.16) holds, we say that $\overline{u}$ dominates u.*

2.4 Transient and Absorbing MDPs

Definition 2.2 *(Transient and absorbing policies)*
Fix an initial distribution β. A policy u is said to be $\mathbf{X}'$-transient where $\mathbf{X}' \subset \mathbf{X}$, if

$$\sum_{t=1}^{\infty} p_\beta^u(t;x) < \infty \textit{ for any } x \in \mathbf{X}'.$$

It is called $\mathbf{X}'$-absorbing if

$$\sum_{t=1}^{\infty} p_\beta^u(t;\mathbf{X}') < \infty.$$

Definition 2.3 *(Transient and absorbing MDPs)*
An MDP for which all policies are $\mathbf{X}'$-transient ($\mathbf{X}'$-absorbing) is called a $\mathbf{X}'$-transient ($\mathbf{X}'$-absorbing, respectively) MDP. A $\mathbf{X}$-transient MDP is called a transient MDP.

Here are some properties of transient stationary policies.

Lemma 2.1 *(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

$$f = \beta + fP(w), \qquad f \geq 0. \tag{2.17}$$

(where f and β are row vectors on $\mathbf{X}'$, and $P(w)$ is the restriction to $\mathbf{X}'$ of the transition probability matrix of the Markov chain corresponding to the stationary policy w). It is the unique solution to (2.17) among f that satisfy $\lim_{n\to\infty} fP^n(w) = 0$.
(ii) Assume that the MDP is $\mathbf{X}'$-absorbing. Fix some policy u and define $g^*(x) := \sum_{t=1}^{\infty} p_\beta^u(t;x)$. If g^* satisfies (2.17), then $g^*(x) = f^*(x)$ for all $x \in \mathbf{X}'$.

Proof: (i) It follows easily that f^* is indeed a solution of (2.17). Iterating (2.17) we get for all integers n:

$$\begin{aligned} f &= \beta + (\beta + fP(w))P(w) = \beta + \beta P(w) + fP^2(w) \\ &= \sum_{i=1}^{n-1} \beta P^i(w) + fP^n(w) = \sum_{t=1}^{n-1} p_\beta^w(t) + fP^n(w) \end{aligned} \tag{2.18}$$

(i) follows since the above holds for all n and since $f \geq 0$.
(ii) follows 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, $\langle g^*, \mathbf{1} \rangle < \infty$. (Here, and throughout, we use the notation $\langle q_1, q_2 \rangle$ between two vectors to denote the scalar product.) For any integer n and $y \in \mathbf{X}$, the yth column of $P^n(w)$ is bounded by $\mathbf{1}$, so by the the generalized dominance convergence Theorem (Royden (1988) Proposition 11.18),

$$\lim_{n\to\infty} g^* P^n(w) = g^* \left(\lim_{n\to\infty} P^n(w) \right) = 0. \tag{2.19}$$

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 $P^\infty = \overline{\lim}_{n\to\infty} P^n(w)$ is zero. Assume that for some z, $P^\infty_{yz} \neq 0$. There exists some time t for which $p^u_\beta(t;y) > 0$. Consider a policy v that behaves like policy u till time t and then behaves like the stationary policy w. Then

$$\sum_{s=1}^{\infty} p^v_\beta(s;z) \geq p^u_\beta(t,y)\sum_{n=0}^{\infty}[P^n(w)]_{yz} = \infty$$

This contradicts the fact that the MDP is absorbing. This shows that $P^\infty_{yz} = 0$ for all z, from which (2.19) follows. The proof now follows taking the limit as $n \to \infty$ in (2.18). ■

2.5 Contracting MDPs

We begin by introducing the following μ-norm. For any functions $q : \mathbf{X} \to \mathbb{R}$, $Q : \mathbf{X}\times\mathbf{X} \to \mathbb{R}$ and $\mu : \mathbf{X} \to [1,\infty)$, we define

$$\|q\|_\mu = \sup_{x\in\mathbf{X}} \frac{q(x)}{\mu(x)}, \qquad \|Q\|_\mu = \sup_{x\in\mathbf{X}} \frac{\sum_{y\in\mathbf{X}} Q_{xy}\mu(y)}{\mu(x)}. \tag{2.20}$$

It is easily verified that μ 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 $\left\|Q^1Q^2\right\|_\mu \leq \left\|Q^1\right\|_\mu \left\|Q^2\right\|_\mu$. 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 norms, and M^μ to be the set of measures $\mathsf{M}^\mu := \{q \in M(\mathbf{X}) : E^q\mu < \infty\}$ (here, $E^q\mu := \sum_x q(x)\mu(x)$).

With some abuse of notation, we shall say that a function $f : \mathcal{K} \to \mathbb{R}$ is in F^μ if the function defined on $\mathbf{X}$ whose x entry is $\sup_{a\in\mathsf{A}(x)} f(x,a)$, is in F^μ. Similarly, a measure q on $\mathcal{K}$ is said to be in M^μ if the measure $\overline{q}$ in M^μ, where $\overline{q}(x) := q(x, \mathsf{A}(x))$.

Definition 2.4 *(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 some scalar $\xi \in [0,1)$ (called the contracting factor), a vector $\mu : \mathbf{X} \to [1,\infty)$, such that for all $x \in \mathbf{X}, a \in \mathsf{A}(x)$,

$$\sum_{y\notin\mathcal{M}} \mathcal{P}_{xay}\mu(y) \leq \xi\mu(x). \tag{2.21}$$

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 $a(n) \to a$ in $\mathsf{A}(x)$ then*

$$\lim_{n\to\infty} \sum_{y\in\mathbf{X}} \left|\mathcal{P}_{xa(n)y} - \mathcal{P}_{xay}\right| \mu(y) = 0. \tag{2.22}$$

- *$c(x,a)$ and $d^k(x,a), k=1,...,K$ are μ-bounded, i.e. $\exists \overline{b} < \infty$ s.t.*

$$\sup_{u \in U_D} \|c(\cdot,u)\|_\mu < \overline{b} \textit{ and } \sup_{u \in U_D, k} \left\|d^k(\cdot,u)\right\|_\mu < \overline{b}$$

An alternative way to write (2.21) is by introducing the taboo probabilities. We define for any $Q : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$

$$_{\mathcal{M}}Q_{xy} = \begin{cases} Q_{xy} & \text{if } y \notin \mathcal{M}, \\ 0 & \text{otherwize .} \end{cases} \tag{2.23}$$

We further define

$$_{\mathcal{M}}p_x^u(t;x) := P_\beta^u(X_t = x, X_s \notin \mathcal{M}, s = 1,...,t).$$

(2.21) can be rewritten as

$$\sup_{w \in U_D} \|_{\mathcal{M}}P(w)\|_\mu \leq \xi.$$

We now show that the contracting framework implies that the MDP is $\mathbf{X}'$-absorbing where $\mathbf{X}' = \mathbf{X} \setminus \mathcal{M}$, for suitable initial distributions.

The μ-continuity, defined in (2.22) is related to a total variation convergence, weighted by the vector μ. It is related to standard continuity as follows:

Lemma 2.2 *(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) $Q(f)$ and $Q(f)\mu$ 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^*,

$$\lim_{n \to \infty} [Q(u_n)q_n]_x = [Q(u^*)q^*]_x, \qquad \forall x \in \mathbf{X}.$$

Lemma 2.3 *(Rate of convergence)*
Consider the contracting MDP. The μ-norm of $_{\mathcal{M}}p_{(\cdot)}^u(t)$ converges to 0 in a geometric rate, uniformly over all $u \in U$, i.e. for any $x \in \mathbf{X}'$

$$_{\mathcal{M}}p_x^u(t;\mathbf{X}') \leq \sum_{y \in \mathbf{X}'} {}_{\mathcal{M}}p_x^u(t;y)\mu(y) \leq \mu(x)\xi^{t-1}. \tag{2.24}$$

Moreover,

$$_{\mathcal{M}}p_\beta^u(t;\mathbf{X}') \leq \sum_{y \in \mathbf{X}'} {}_{\mathcal{M}}p_\beta^u(t;y)\mu(y) \leq \langle \beta, \mu \rangle \xi^{t-1};$$

and

$$\sum_{t=1}^{\infty} {}_{\mathcal{M}}p_\beta^u(t;\mathbf{X}') \leq \sum_{t=1}^{\infty} \sum_{y \in \mathbf{X}'} {}_{\mathcal{M}}p_\beta^u(t;y)\mu(y) \leq \frac{\langle \beta, \mu \rangle}{1 - \xi},$$

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 (2.15). Viewing ${}_{\mathcal{M}}p^u_{(\cdot)}(t;\cdot)$ as a matrix, we have

$$\|{}_{\mathcal{M}}p^u_{\cdot}(t;\cdot)\|_\mu \le \|{}_{\mathcal{M}}P(v_1)\|_\mu \, \|{}_{\mathcal{M}}P(v_2)\|_\mu \cdots \|{}_{\mathcal{M}}P(v_{t-1})\|_\mu \le \xi^{t-1},$$

which implies

$${}_{\mathcal{M}}p^u_x(t;\mathbf{X}') \le \sum_{y\in\mathbf{X}'} p^u_x(t;y)\mu(y) \le \mu(x)\xi^{t-1}.$$

This concludes the proof. ■

Hence, contracting MDPs are a subclass of absorbing MDPs, which are subclass of transient MDPs. The converse needs 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 nonabsorbing. For the case of finite state space, however, the converse holds, and any transient policy is contracting, see Kallenberg (1983).

Lemma 2.1 can be strengthen:

Lemma 2.4 *(Uniqueness of a bonded 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

$$f(x) = \sum_{t=1}^{\infty} {}_{\mathcal{M}}p^w_\beta(t;x)$$

is the unique μ-bounded solution of

$$f = \beta + fP(w). \tag{2.25}$$

Proof: Let f' be some μ-bounded solution of (2.25). Iterating (2.25), we get:

$$\begin{aligned} f'(y) &= \beta(y) + {}_{\mathcal{M}}p^w_\beta(2;y) + \sum_{x\in\mathbf{X}'} f'(x)[P^2(w)]_{xy} \qquad (2.26)\\ &= \beta(y) + {}_{\mathcal{M}}p^w_\beta(2;y) + {}_{\mathcal{M}}p^w_\beta(3;y) + \sum_{x\in\mathbf{X}'} f'(x)[P^3(w)]_{xy} \\ &= \ldots = \beta(x) + {}_{\mathcal{M}}p^w_\beta(2;x) + \ldots + {}_{\mathcal{M}}p^w_\beta(t;x) + \sum_{x\in\mathbf{X}'} f'(x)[P^t(w)]_{xy} \end{aligned}$$

We have (as in Lemma 2.3)

$$\left|\sum_{x\in\mathbf{X}'} f'(x)[P^t(w)]_{xy}\right| \le \mu(y)\,\|f'\|_\mu \left\|{}_{\mathcal{M}}P^t(w)\right\|_\mu \le \mu(y)\,\|f'\|_\mu\,\xi^t \to 0$$

The proof is established by taking the limit as $t \to \infty$ in (2.26). ■

Remark 2.2 *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 Wessesls (1977), who considers a similar definition but with an empty set $\mathcal{M}$. Allowing nonempty sets $\mathcal{M}$ turns to be especially important in the average cost case.*

CHAPTER 3

The total cost: occupation measures and the primal LP

We study in this chapter the total cost criterion for $\mathbf{X}'$-transient MDPs. Let $\mathcal{M}$ be the complement of $\mathbf{X}'$ in $\mathbf{X}$. We assume throughout this chapter that $c(x,a) = d^k(x,a) = 0, k = 1, ..., K$ for any $x \in \mathcal{M}$, and that the initial distribution β has zero mass on $\mathcal{M}$ ($\mathcal{M}$ may be empty). The total cost criterion has the meaning of the total expected cost till the set $\mathcal{M}$ is reached. In the end of the chapter, we use the theory we developed for the total cost in order to solve the constrained optimal control problems with discounted cost.

3.1 Occupation measure

For any given initial distribution β and policy u, define the occupation measure $f_{tc}(\beta, u; x, \cdot)$ related to the total cost criterion by

$$f_{tc}(\beta, u; x, \mathcal{A}) = \sum_{t=1}^{\infty} {}_{\mathcal{M}}p_\beta^u(t; x, \mathcal{A}), \qquad \mathcal{A} \subset \mathsf{A}(x).$$

With some abuse of notation, we denote $f_{tc}(\beta, u; x) = f_{tc}(\beta, u; x, \mathsf{A}(x))$. Let $f_{tc}(\beta, u) := \{f_{tc}(\beta, u; x, \cdot), x \in \mathbf{X}'\}$. Define $\mathcal{K}' := \{(x,a) : x \in \mathbf{X}', a \in \mathsf{A}(x)\}$, and

$$\mathbf{L}_{\overline{U}}(\beta) = \bigcup_{u \in \overline{U}} \{f_{tc}(\beta, u)\} \text{ for any } \overline{U} \subset U \cup \overline{M}(U_M), \tag{3.1}$$

$$\mathbf{Q}_{tc}(\beta) = \tag{3.2}$$
$$\left\{ \begin{array}{c} \rho \in M(\mathcal{K}') : \displaystyle\sum_{y \in \mathbf{X}'} \int_{\mathcal{A}(y)} \rho(y, da)(\delta_x(y) - \mathcal{P}_{yax}) = \beta(x), x \in \mathbf{X}' \\ \\ \rho(x, \mathsf{A}(x)) = 0 \text{ for } x \in \mathcal{M}, \quad \rho(x, \mathsf{A}(x)) < \infty \text{ for } x \notin \mathcal{M}, \end{array} \right\}$$

where $M(\mathcal{K})$ is the set of nonnegative measures over $\mathcal{K}$. We set $\mathbf{L}(\beta) = \mathbf{L}_U(\beta) \cup \mathbf{L}_{\overline{M(U_M)}}(\beta)$. For the contracting framework we define

$$\mathbf{Q}_{tc}^\mu(\beta) := \mathbf{Q}_{tc}(\beta) \cap \mathsf{M}^\mu.$$

For any sets B, B_1, B_2 in $M(\mathcal{K})$, define

- coB_1 := the convex hull of a set;
- $\min B$:= the set of minimal elements in B, i.e. $\rho \in \min B$ if there does not exist some $\rho' \in B$ such that $\rho'(y,\mathcal{A}) < \rho(y,\mathcal{A})$ for some $y \in \mathbf{X}$ and $\mathcal{A} \subset \mathsf{A}(y)$;
- $B_1 \prec B_2$ if $\forall \rho_2 \in B_2$ there exists $\rho_1 \in B_1$ such that $\rho_1 \leq \rho_2$.

Definition 3.1 *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 3.1 *(Completeness of stationary policies)*
(i) Consider a $\mathbf{X}'$-transient MDP. Then the set of stationary policies is weakly complete.
(ii) If the MDP is $\mathbf{X}'$-absorbing, then the set of stationary policies is complete.

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

$$w_y(\mathcal{A}) = \frac{f_{tc}(\beta,u;y,\mathcal{A})}{f_{tc}(\beta,u;y)}, \quad y \in \mathbf{X}', \mathcal{A} \subset \mathsf{A}(y)$$

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

$$\begin{aligned}
f_{tc}(\beta,u;x) &= \beta(x) + \sum_{t=2}^{\infty} p_\beta^u(t,x) \\
&= \beta(x) + \sum_{t=2}^{\infty} \int_{\mathcal{K}'} p_\beta^u(t-1;d\kappa)\mathcal{P}_{\kappa x} \\
&= \beta(x) + \int_{\mathcal{K}} f_{tc}(\beta,u;d\kappa)\mathcal{P}_{\kappa x} && (3.3) \\
&= \beta(x) + \sum_{y\in\mathbf{X}} f_{tc}(\beta,u;y) \int_{\mathsf{A}(y)} \mathcal{P}_{yax} w_y(da) \\
&= \beta(x) + \sum_{y\in\mathbf{X}} f_{tc}(\beta,u;y) P_{yx}(w) && (3.4)
\end{aligned}$$

Hence, by Lemma 2.1 (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 2.1 (ii) and (3.4). ■

Definition 3.2 *(μ-continuity)*
Consider some $\overline{U} \subset U_M$ and $Q : \overline{U} \times \mathbf{X}$. 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 2.2.

Lemma 3.1 *(Continuity properties of f_{tc})*
(i) For transient MDPs, the map $f_{tc}(\beta,\cdot): U_M \to \mathbf{L}_{U_M}$ is lower semi-continuous. The same holds for the map $f_{tc}(\beta,\cdot): \overline{M}(U_M) \to \mathbf{L}_{U_M}$.
(ii) Consider a contracting MDP. Then the map $f_{tc}(\beta,\cdot): U_M \to \mathbf{L}_{U_M}$ is μ-continuous. The same holds for the map $f_{tc}(\beta,\cdot): \overline{M}(U_M) \to \mathbf{L}_{U_M}$.

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^n). Then $P_{xy}(u_t^n) \to P_{xy}(u)$ 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 $x, y \in \mathbf{X}'$. Moreover, the m step probabilities also converge, i.e. for all integers m:

$$\lim_{n\to\infty} p_\beta^{u^n}(m; y, \mathcal{A}) = \lim_{n\to\infty} \sum_{x\in\mathbf{X}'} \beta(x)[P(u_1^n)P(u_2^n)...P(u_m^n)]_{xy} u_m^n(\mathcal{A}|y)$$
$$= \sum_{x\in\mathbf{X}'} \beta(x)[P(u_1)P(u_2)...P(u_m)]_{xy} u_m(\mathcal{A}|y) = p_\beta^u(m; y, \mathcal{A}), \tag{3.5}$$

for all $y \in \mathbf{X}'$, $\mathcal{A} \subset \mathsf{A}$. (3.5) 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 $P(u_{m+1}^n)$ and $P(u_{m+1})$, 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 (3.5) follows.

Now, we fix some $y \in \mathbf{X}', \mathcal{A} \in \mathsf{A}$, and consider $p_\beta^{u^n}(m; y, \mathcal{A})$ to be a function over the integers m. This function converges pointwize to $p_\beta^u(m; y, \mathcal{A})$. Applying Fatou's Lemma (Royden (1988) Proposition 11.17) with respect to the (infinite) measure over the integers $\mu_n(m) = \mu(m) = 1$, we obtain

$$\varliminf_{n\to\infty} f_{tc}(\beta, u^n; y, \mathcal{A}) \geq f_{tc}(\beta, u; y, \mathcal{A}),$$

which concludes the proof of (i).

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 lower semi-continuity on $\overline{M}(U_M)$ is established by applying again Fatou's Lemma:

$$\varliminf_{n\to\infty} f_{tc}(\beta, \hat{\gamma}^n)$$
$$= \varliminf_{n\to\infty} \langle \gamma^n, f_{tc}(\beta,\cdot)\rangle \geq \langle \varliminf_{n\to\infty} \gamma^n, f_{tc}(\beta,\cdot)\rangle = f_{tc}(\beta, \hat{\gamma}).$$

(ii) Assume again that $u^n \to u$, where $u^n, u \in U_M$. Then by assumption,

$$\lim_{n\to\infty} \sum_{y\in\mathbf{X}'} |P_{xy}(u_1^n) - P_{xy}(u)|\mu(y) = 0$$

for all $x \in \mathbf{X}'$. Since $\sum_{y\in\mathbf{X}'} |P_{xy}(u_1^n) - P_{xy}(u)|\mu(y) \leq \mu(x)$ and $\langle \beta, \mu \rangle < \infty$, it follows from the bounded convergence theorem, that

$$\begin{aligned} & \lim_{n\to\infty} \sum_{x\in\mathbf{X}'} \beta(x) \sum_{y\in\mathbf{X}'} |P_{xy}(u_1^n) - P_{xy}(u)|\mu(y) \\ = & \lim_{n\to\infty} \sum_{y\in\mathbf{X}'} |p_\beta^{u^n}(1;y) - p_\beta^u(1;y)|\mu(y) = 0. \end{aligned} \tag{3.6}$$

Moreover, the m step probabilities also converge, i.e. for all integers m:

$$\lim_{n\to\infty} \sum_{y\in\mathbf{X}'} |p_\beta^{u^n}(m;y) - p_\beta^u(m;y)|\mu(y) = 0. \tag{3.7}$$

This, again is established by induction. It holds for $m = 1$. Assume it holds for arbitrary m.

$$\begin{aligned} & \sum_{y\in\mathbf{X}'} |p_\beta^{u^n}(m+1;y) - p_\beta^u(m+1;y)|\mu(y) \\ = & \sum_{y,z\in\mathbf{X}'} |p_\beta^{u^n}(m;z)P_{zy}(u_m^n) - p_\beta^u(m;z)P_{zy}(u_m)|\mu(y) \\ \leq & \sum_{y,z\in\mathbf{X}'} \big(p_\beta^{u^n}(m;z)|P_{zy}(u_m^n) - P_{zy}(u_m)| \\ & +|p_\beta^{u^n}(m;z) - p_\beta^u(m;z)|P_{zy}(u_m)\big)\mu(y) \\ \leq & \sum_{y,z\in\mathbf{X}'} p_\beta^{u^n}(m;z)|P_{zy}(u_m^n) - P_{zy}(u_m)|\mu(y) \\ & + \sum_{z\in\mathbf{X}'} |p_\beta^{u^n}(m;z) - p_\beta^u(m;z)|\xi\mu(z). \end{aligned}$$

The first summation tends to zero as $n \to \infty$ by the same argument as in (3.6), since

$$\langle p_\beta^{u^n}(m;\cdot), \mu \rangle < \infty$$

by Lemma 2.3. The second summation converges to zero by the induction hypothesis. We conclude that for every m, (3.7) holds, and consequently $f_{tc}^T(\beta, u)$ are μ-continuous on U_M. Next, we observe that for any $u \in U_M$,

$$\sum_{m=T+1}^{\infty} \sum_{y\in\mathbf{X}} p_\beta^u(m;y)\mu(y) \leq \langle \beta, \mu \rangle \sum_{m=T}^{\infty} \xi^m$$

by Lemma 2.3. We conclude that

$$\begin{aligned}
&\sum_{y\in\mathbf{X}'} |f_{tc}(\beta,u^n) - f_{tc}(\beta,u)|\mu(y)\\
&\quad\le \sum_{y\in\mathbf{X}'} |f^T(\beta,u^n;y) - f^T(\beta,u;y)|\mu(y) + 2\langle\beta,\mu\rangle\sum_{m=T}^{\infty}\xi^m.
\end{aligned}$$

Since this holds for every T, and since for every T

$$\lim_{T\to\infty}\sum_{y\in\mathbf{X}'} |f^T(\beta,u^n;y) - f^T(\beta,u;y)|\mu(y) = 0,$$

we conclude that $f_{tc}(\beta,\cdot)$ is μ-continuous on U_M.

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 (Billingseley, 1968) implies:

$$\begin{aligned}
&\lim_{n\to\infty} f_{tc}(\beta,\hat\gamma^n)\\
&\quad= \lim_{n\to\infty}\langle\gamma^n, f_{tc}(\beta,\cdot)\rangle = \langle\lim_{n\to\infty}\gamma^n, f_{tc}(\beta,\cdot)\rangle = C_{tc}(\beta,\hat\gamma).
\end{aligned}$$

■

Lemma 3.2 *(Splitting in a state)*

Choose $w\in U_S$ and a state y. Define $w^a\in U_S$ to be the policy that chooses always 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 $T(y)\stackrel{\text{def}}{=}\inf_{r>1}\{X_r=y\}$, $y\in\mathbf{X}$, with the convention that $\inf\{\emptyset\}=\infty$. Define the total expected number of visits to state y starting from state x as:

$$\mathcal{T}(u;x,y) = E_x^u\left(\sum_{t=2}^{T(y)} 1(X_t,A_t)\right) \tag{3.8}$$

and the probability of ever reaching state y from state x:

$$\overline{p}(u;x,y) := P_x^u(T(y)<\infty).$$

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

$$f_{tc}(x,w;y) = \mathcal{T}(w;x,y) + \overline{p}(u;x,y) f_{tc}(y,w;y)$$

By setting $x = y$ we get

$$f_{tc}(y,w;y) = \frac{\mathcal{T}(w;y,y)}{1-\overline{p}(w;y,y)} = \frac{\int w_y(da)\mathcal{T}(w^a;y,y)}{1-\overline{p}(w;y,y)}$$

$$= \int \frac{w_y(da)(1-\overline{p}(w^a;y,y))}{1-\overline{p}(w;y,y)} f_{tc}(y,w^a;y) = \int_{\mathsf{A}(y)} \gamma(da) f_{tc}(y,w^a;y),$$

which establishes the proof for the case $x = y$. For general x,

$$\begin{aligned} f_{tc}(x,w;y) &= \mathcal{T}(w;x,y) + \overline{p}(u;x,y) f_{tc}(y,w;y) \\ &= \int_{\mathsf{A}(y)} \gamma(da)[\mathcal{T}(w;x,y) + \overline{p}(u;x,y) f_{tc}(y,w^a;y)] \\ &= \int_{\mathsf{A}(y)} \gamma(da) f_{tc}(x,w^a;y). \end{aligned}$$

■

Theorem 3.2 *(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 absorbing MDPs, $\mathbf{L}_{U_S}(\beta)$ *is convex and compact, and satisfies*

$$\mathbf{L}_{\mathcal{U}}(\beta) = \mathbf{L}(\beta) = \mathbf{L}_{U_S}(\beta) = co\mathbf{L}_{U_D}(\beta) = \min \mathbf{Q}_{tc}(\beta).$$

(iii) For contracting MDPs, $\mathbf{L}_{U_S}(\beta)$ *is convex and compact, and satisfies*

$$\mathbf{L}_{\mathcal{U}}(\beta) = \mathbf{L}(\beta) = \mathbf{L}_{U_S}(\beta) = co\mathbf{L}_{U_D}(\beta) = \mathbf{Q}^{\mu}_{tc}(\beta) = \min \mathbf{Q}_{tc}(\beta).$$

Proof: (i) Theorem 2.1 implies that $\mathbf{L}(\beta) = \mathbf{L}_{U_M}(\beta)$ is convex. The weakly completeness of $\mathbf{L}_{U_S}(\beta)$ was established in Theorem 3.1. That $\mathbf{L}(\beta) \subset \mathbf{Q}_{tc}(\beta)$ follows from (3.4). 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 nonzero. We have

$$\begin{aligned} \rho(x,\mathsf{A}(x)) &= \beta(x) + \int_{\mathcal{K}} \rho(d\kappa)\mathcal{P}_{\kappa x} \\ &= \beta(x) + \sum_y \rho(y,\mathsf{A}(y)) \int_{\mathsf{A}(y)} \mathcal{P}_{yax} w_y(da) \\ &= \beta(x) + \sum_y \rho(y,\mathsf{A}(y)) P_{yx}(w). \end{aligned} \tag{3.9}$$

By Lemma 2.1 (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 3.1 (ii), hence $\mathbf{L}_{U_S}(\beta)$ is convex. The compactness of $\mathbf{L}_{U_M}(\beta)$ follows since by Section 2.2 and Lemma 3.1 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 convex hull of $\mathbf{L}_{U_D}(\beta)$ (and thus of $\mathbf{L}_{\mathcal{U}}(\beta)$). Since it is compact, by the Krein-Milman theorem it is the convex hull of its extreme points. Choose some $w \in U_S$. Suppose that w is not deterministic, so that $w_y(\cdot)$ is not concentrated on a single point in $\mathsf{A}(y)$. But then by Lemma 3.2, w is not an extreme point of $\mathbf{L}_{U_S}$. The rest follow from part (i).

(iii) Since contracting MDPs are absorbing (Lemma 2.3), all the statements in (ii) hold. It remains to show that $\mathbf{Q}^{\mu}_{tc}(\beta) \subset \mathbf{L}_{U_S}(\beta)$. By applying Lemma 2.1 to (3.9), this statement will follow if we show that $\lim_{n\to\infty} \sum_y \rho(y, \mathsf{A}(y))[P^n(w)]_{yx}$ is zero. Indeed,

$$\sum_y \rho(y, \mathsf{A}(y))[P^n(w)]_{yx} \leq \sum_y \rho(y, \mathsf{A}(y))\mu(y) \left\| [P^n(w)]_{\cdot,\mathbf{X}'} \right\|_{\mu},$$

which converges to zero by Lemma 2.3. ■

3.2 Relation between cost and occupation measure

We begin by introducing different assumptions on the immediate costs, that will be used when applying either the general transient framework, or the contracting framework. When using the general transient framework we shall assume that the immediate costs are non-negative. We have the following properties of the total costs:

Theorem 3.3 *(Linear representation and boundedness of the cost)*
(i) Assume that the MDP is contracting. Then for any instantaneous cost $c : \mathcal{K} \to \mathbb{R}$ and any β and $u \in U$

$$C_{tc}(\beta, u) = \langle c, f_{tc}(\beta, u)\rangle := \int_{\mathcal{K}} c(\kappa) f_{tc}(\beta, u; d\kappa); \tag{3.10}$$

the finite and infinite horizon total costs are uniformly μ-bounded over all policies:

$$\left\| C^T(\cdot, u) \right\|_{\mu} \leq \frac{\overline{b}}{1-\xi} \qquad\qquad \|C_{tc}(\cdot, u)\|_{\mu} \leq \frac{\overline{b}}{1-\xi}. \tag{3.11}$$

($C_{tc}(\cdot, u)$ is the vector of total cost corresponding to all initial states.)
(ii) Assume that the MDP is transient and the immediate costs are non-negative. Then (3.10) holds for any $u \in U$.

Proof: (i) Fix some policy u. Since $f^T(\beta, u; y)$ converges monotonly to $f_{tc}(\beta, u; y)$ as $T \to \infty$, we have by the monotone convergence theorem

$$\lim_{T\to\infty} \langle f^T(\beta, u; y), \mu\rangle = \langle f_{tc}(\beta, u; y), \mu\rangle.$$

Since c is μ-bounded, we have by the dominated convergence theorem

$$C_{tc}(\beta,u) = \lim_{T\to\infty} C^T(\beta,u) = \lim_{T\to\infty} \langle f^T(\beta,u;y),c\rangle = \langle f_{tc}(\beta,u;y),c\rangle.$$

(3.11) follows from

$$\|C_{tc}(u)\|_\mu = \|\langle c, f_{tc}(\cdot,u)\rangle\|_\mu \le \bar{b}\,\|f_{tc}(\cdot,u)\|_\mu \le \frac{\bar{b}}{1-\xi}, \tag{3.12}$$

and

$$\left\|C^T(u)\right\|_\mu = \left\|\langle c, f^T(\cdot,u)\rangle\right\|_\mu \le \bar{b}\,\|f_{tc}(\cdot,u)\|_\mu \le \frac{\bar{b}}{1-\xi}. \tag{3.13}$$

(ii)

$$\begin{aligned} C_{tc}(\beta,u) &= \sum_{t=1}^{\infty} E_\beta^u c(X_t,A_t) = \sum_{t=1}^{\infty} \int_{\mathcal{K}} p_\beta^u(t;d\kappa)c(\kappa) \\ &= \int_{\mathcal{K}} \sum_{t=1}^{\infty} p_\beta^u(t;d\kappa)c(\kappa) = \langle f_{tc}(\beta,u),c\rangle, \end{aligned}$$

where the change between integration and summation follows since the integrand is non-negative (see Royden, 1988, Corollary 11.14). ∎

Lemma 3.3 *(The transient case: lower semi-continuity of the costs)*
Consider the transient framework, (Definition 2.2 with nonnegative costs). Then $C(\beta,\cdot)$ (and $D^k(\beta,\cdot), k=1,...,K$) are lower semi-continuous on U_S.

Proof: Let $w_n \to w$ be stationary. Then, by Fatou's Lemma (the arguments are as in the proof of Lemma 3.1 (i)) and Theorem 3.3, and by the lower semi-continuity of the occupation measures (Lemma 3.1 (i)), we have

$$\begin{aligned} \varliminf_{n\to\infty} C_{tc}(\beta,w_n) &= \varliminf_{n\to\infty} \langle f_{tc}(\beta,w_n),c\rangle \ge \langle \varliminf_{n\to\infty} f_{tc}(\beta,w_n),c\rangle \\ &\ge \langle f_{tc}(\beta,w),c\rangle = C_{tc}(\beta,w). \end{aligned}$$

(The same argument holds for $D_{tc}(\beta,\cdot)$). ∎

Lemma 3.4 *(Uniform convergence and continuity of the costs)*
Assume that the MDP is contracting. Then
(i) $C^T(\beta,u)$ converges to $C_{tc}(\beta,u)$ uniformly over U as $T\to\infty$.
(ii) $C_{tc}(\beta,u)$ is continuous on U_M.

Proof: (i) For any policy u,

$$|C_{tc}(\beta,u) - C^T(\beta,u)|$$

$$
\begin{aligned}
&\leq \sum_{t=T+1}^{\infty} E_\beta^u |c(X_t, A_t)| = \sum_{t=T+1}^{\infty} \int_{\mathcal{K}} p_\beta^u(t; d\kappa)|c(\kappa)| \\
&\leq \bar{b} \sum_{t=T+1}^{\infty} \sum_{y \in \mathbf{X}'} p_\beta^u(t; y)\mu(y) \leq \frac{\bar{b}\langle\beta, \mu\rangle \xi^T}{1-\xi}
\end{aligned}
$$

which converges to 0 as $T \to \infty$. The last inequality follows from Lemma 2.3.
(ii) Consider any $u, u' \in U_M$. It follows from Theorem (3.3) that

$$
\begin{aligned}
&|C_{tc}(\beta, u) - C_{tc}(\beta, u')| \\
&= \langle c, f_{tc}(\beta, u)\rangle - \langle c, f_{tc}(\beta, u')\rangle \leq \langle c, f_{tc}(\beta, u) - f_{tc}(\beta, u')\rangle \\
&\leq \bar{b} \sum_{y \in \mathbf{X}'} |c, f_{tc}(\beta, u) - f_{tc}(\beta, u')|\mu(y).
\end{aligned}
$$

Since f_{tc} are μ-continuous (Lemma 3.1 (ii). ■

Lemma 3.5 *(Extension to $\mathcal{U}$ and $\overline{M}(U_M)$)*
The results of Lemmas 3.4 and 3.3 as well as Theorem 3.3 holds also for $\overline{M}(U_M)$ (and thus, in particular, for $\mathcal{U}$).

Proof: The extension of Lemma 3.3: 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 lower semi-continuity on $\overline{M}(U_M)$ is established by applying again Fatou's Lemma:

$$
\begin{aligned}
&\underline{\lim}_{n\to\infty} C_{tc}(\beta, \hat{\gamma}^n) \\
&= \underline{\lim}_{n\to\infty} \langle \gamma^n, C_{tc}(\beta, \cdot)\rangle \geq \langle \underline{\lim}_{n\to\infty} \gamma^n, C_{tc}(\beta, \cdot)\rangle = C_{tc}(\beta, \hat{\gamma}).
\end{aligned}
$$

The extension of Lemma 3.4: we thus consider the contracting framework.

$$
\begin{aligned}
&\sup_{\hat{q} \in \overline{M}(U_M)} |C^T(\beta, \hat{q}) - C_{tc}(\beta, \hat{q})| \\
&= \sup_{q \in M_1(U_M)} \left| \int_{U_M} C^T(\beta, u) q(du) - \int_{U_M} C_{tc}(\beta, u) q(du) \right| \\
&\leq \sup_{q \in M_1(U_M)} \int_{U_M} |C^T(\beta, u) - C_{tc}(\beta, u)| q(du) \\
&= \sup_{u \in U_M} \int_{U_M} |C^T(\beta, u) - C_{tc}(\beta, u)|
\end{aligned}
$$

We conclude that Lemma 3.4 (i) holds for $\overline{M}(U_M)$.

Let $q^n, n = 1, 2, \ldots$ and q be probability measures over U_M, and let $\hat{q}^n$ and $\hat{q}$ be the corresponding policies in $\overline{M}(U_M)$. Assume that $\hat{q}^n$ converges to $\hat{q}$ (by which we mean that

q^n converges to q weakly). Then

$$\lim_{n\to\infty} C_{tc}(\beta, \hat{q}^n) = \lim_{n\to\infty} \langle q^n, C_{tc}(\beta,\cdot)\rangle = \langle q, C_{tc}(\beta,\cdot)\rangle = C_{tc}(\beta,\hat{q}).$$

Indeed, this follows (see Billingseley, 1968) since, by Theorem 3.3 (i), $C_{tc}(\beta, u)$, are bounded on U_M, and since, by Lemma 3.4, $C_{tc}(\beta, u)$ are continuous on U_M. This establishes Lemma 3.4 (ii) for $\overline{M}(U_M)$.

The extension of Theorem 3.3 to $\overline{M}(U_M)$ is straightforward. ■

3.3 Dominating classes of policies

Theorem 3.4 *(Dominating policies)*
(i) Consider the transient framework, (Definition 2.2) together with nonnegative costs. Then both U_S and $\mathcal{U}$ are dominating classes of policies.
(ii) Consider the contracting framework, (Definition 2.4). Then any complete class of policies (definition 3.1) is a dominating class of policies.
(iii) Under the assumptions of (i) or of (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 3.3 as well as the weak completeness of the set of stationary policies (Theorem 3.1). We delay the proof for $\mathcal{U}$ to the next Chapter (Corollary 4.1).

(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 $\mathcal{U}$ (Lemma 3.4, 3.3 and 3.5). 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 $\mathcal{U}$. ■

3.4 Equivalent Linear Program

We show below that **COP** is equivalent to a LP with countable number of decision variables and a countable number 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 $\mathcal{C}(\rho) := \langle c, \rho\rangle$ subject to:

$$\mathcal{D}^k(z) := \langle d^k, \rho\rangle \leq V_k, k = 1, ..., K, \qquad \rho \in \mathbf{Q}_{tc}(\beta) \tag{3.14}$$

where $\mathbf{Q}_{tc}(\beta)$ was defined in (3.2).

We similarly define the LP for the contracting case:
$\mathbf{LP}_{\mathbf{1}}^{\mu}(\beta)$: Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle c, \rho \rangle$ subject to:

$$\mathcal{D}^k(z) := \langle d^k, \rho \rangle \le V_k, k = 1, ..., K, \qquad \rho \in \mathbf{Q}_{tc}^{\mu}(\beta). \tag{3.15}$$

We show that there is a one to correspondence between feasible (and optimal) solutions to the LP, and the feasible (and optimal) solutions to **COP**.

Theorem 3.5 *(Equivalence between COP and LP, the transient case)*
Assume that the MDP is transient and the immediate costs are nonnegative. 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 (3.9)) satisfies* $C_{tc}(\beta, w(\rho)) \le \mathcal{C}(\rho)$ *and* $D_{tc}(\beta, w(\rho)) \le \mathcal{D}(\rho)$.
(iii) $\mathbf{LP}_{\mathbf{1}}(\beta)$ *is feasible if and only if* **COP** *is. Assume that* **COP** *is feasible. Then there exists an optimal solution* ρ^* *for* $\mathbf{LP}_{\mathbf{1}}(\beta)$*, and the stationary policy* $w(\rho^*)$ *is optimal for* **COP**.

Proof: We start from (ii). The first claim follows from eq. (3.3). The claims on the costs follow from Theorem 3.3 and Theorem 3.1.
(i) and (iii) now follows from (ii) and Theorem 3.4. ∎

For the contracting case we get similarly:

Theorem 3.6 *(Equivalence between COP and LP, the contracting case)*
Assume that the MDP is contracting. Then
(i) $\mathcal{C}^* = C_{tc}(\beta)$.
(ii) For any $u \in U$, $\rho(u) := f_{tc}(\beta, u) \in \mathbf{Q}_{tc}^{\mu}(\beta)$, $C_{tc}(\beta, u) = \mathcal{C}(\rho(u))$ *and* $D_{tc}(\beta, u) = \mathcal{D}(\rho(u))$*; conversely, for any* $\rho \in \mathbf{Q}_{tc}^{\mu}(\beta)$*, the stationary policy* $w(\rho)$ *(defined above (3.9) satisfies* $C_{tc}(\beta, w(\rho)) = \mathcal{C}(\rho)$ *and* $D_{tc}(\beta, w(\rho)) = \mathcal{D}(\rho)$.
(iii) $\mathbf{LP}_{\mathbf{1}}^{\mu}(\beta)$ *is feasible if and only if* **COP** *is. Assume that* **COP** *is feasible. Then there exists an optimal solution* ρ^* *for* $\mathbf{LP}_{\mathbf{1}}^{\mu}(\beta)$*, and the stationary policy* $w(\rho^*)$ *is optimal for* **COP**.

3.5 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 nonnegative vectors $\lambda \in \mathbb{R}_+^K$.

$$\begin{array}{lll} \mathbf{DP}_{\mathbf{1}}(\beta) : & & \text{Find } \Theta^* := \sup_{\phi,\lambda} \langle \beta, \phi \rangle - \langle \lambda, V \rangle \text{ s.t.} \\ \phi(x) & \le & c(x,a) + \langle \lambda, d(x,a) \rangle + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y), \;\; x \in \mathbf{X}, a \in \mathsf{A}(x). \end{array}$$

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

$$\Theta^* = \mathcal{C}^* = C_{tc}(\beta), \tag{3.16}$$

for both the transient and the contracting framework. For the contracting framework, we shall restrict to $\phi \in \mathsf{F}^\mu$, and for the transient framework, a possible choice is to bounded ϕ.

3.6 The Discounted cost

A simple and natural way to treet the discounted cost is as a simple equivalent total cost problem. Indeed, 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 that serves as a "grave"; whenever the state process reaches it, it stays there forever; moreover, the immediate costs when the process reaches the grave are zero. Finaly, the probability to move from any state in $\mathbf{X}_\alpha$ to x^o is equal to $1-\alpha$, for any action. We summerize 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

$$\mathcal{P}_{xay} = \begin{cases} \alpha \mathcal{P}_{xay} & \text{if } x, y \in \mathbf{X}_\alpha \\ 1-\alpha & \text{if } x \in \mathbf{X}_\alpha, y = x^o \\ 1 & \text{if } x = y = x^o \\ 0 & \text{otherwize .} \end{cases}$$

- There is only one dummy action a^o available at state x^o, i.e. $\mathsf{A}(x^o) = \{a^o\}$, and $c(x^o, a^o) = d^k(x^o, a^o) = 0,\ k = 1, ..., K$. (The immediate cost in other states are unchanged).

Next, we observe that in the equivalent total cost model,

$$\sum_{t=1}^{\infty} p_\beta^u(t, \mathbf{X}_\alpha) = \frac{1}{1-\alpha}$$

for any policy u and initial distribution β on $\mathbf{X}_\alpha$, so that the equivalent MDP is $\mathbf{X}_\alpha$-absorbing.

The occupation measure for the discounted cost are defined as

$$f_\alpha(\beta, u; x, \mathcal{A}) := \sum_{t=1}^{\infty} \alpha^{t-1} p_\beta^u(t; x, \mathcal{A}), \qquad x \in \mathbf{X}_\alpha, \mathcal{A} \in \mathsf{A}(x),$$

and $f_\alpha(\beta, u) := \prod_{x \in \mathbf{X}} f_\alpha(\beta, u; x, \cdot)$. As for the total cost model, define for any class of policies $\overline{U}$

$$\mathbf{L}_{\overline{U}}^{\alpha}(\beta) = \bigcup_{u \in \overline{U}} f_\alpha(\beta, u), \tag{3.17}$$

and define $\mathbf{L}^\alpha := \mathbf{L}^\alpha_U \cup \mathbf{L}^\alpha_{\overline{M(U_M)}}$. A class of policies $\overline{U}$ is complete with respect to the discounted cost problem if $\mathbf{L}^\alpha = \mathbf{L}^\alpha_{\overline{U}}$.

Define the set

$$\mathbf{Q}^\alpha(\beta) = \tag{3.18}$$
$$\left\{ \begin{array}{c} \rho \in M(\mathcal{K}) : \sum_{y \in \mathbf{X}} \int_{\mathcal{A}(y)} \rho(y, da)(\delta_x(y) - \alpha \mathcal{P}_{yax}) = \beta(x), x \in \mathbf{X} \\ \\ \rho(x, \mathsf{A}(x)) = 0 \text{ for } x \in \mathcal{M}, \quad \rho(x, \mathsf{A}(x)) < \infty \text{ for } x \notin \mathcal{M}. \end{array} \right\}$$

Using the above equivalent absorbing total-cost model, we may apply Theorems 3.1 (ii) and Theorem 3.2 to conclude that

Corollary 3.1 *(Properties of occupation measures)*
The set of stationary policies is complete. Moreover, $\mathbf{L}^\alpha_S(\beta)$ is convex and compact, and satisfies

$$\mathbf{L}^\alpha_{\mathcal{U}}(\beta) = \mathbf{L}^\alpha_U(\beta) = \mathbf{L}^\alpha_S(\beta) = co\mathbf{L}^\alpha_D(\beta) = \min \mathbf{Q}^\alpha(\beta).$$

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 nonnegative hold for the discounted cost as well.

We can now apply either the transient or the contracting framework and obtain the corresponding results for the discounted cost.

For the transient case, we thus recover all the results from Lemmas 3.3 and 3.5, as well as Theorems 3.3, 3.4 and 3.5. (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 (3.14)) becomes

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

$$\mathcal{D}^k(z) := \langle d^k, \rho \rangle \leq V_k, k = 1, ..., K, \qquad \rho \in \mathbf{Q}^\alpha(\beta). \tag{3.19}$$

Next, we formulate the conditions for the equivalent total-cost MDP to be contracting in terms of the original discounted MDP. By using (3.18), the following condition on the discounted MDP will imply that the equivalent total cost MDP is $\mathbf{X}_\alpha$ contracting (i.e., eq. (2.21) will hold): 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 $x \in \mathbf{X}, a \in \mathsf{A}$,

$$\alpha \sum_{y \notin \mathcal{M}} \mathcal{P}_{xay} \mu(y) \leq \xi \mu(x). \tag{3.20}$$

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.

It now follows that Lemmas 3.4 and 3.5 as well as Theorems 3.3, 3.4 and 3.6, hold, where the contracting assumption can be replaced by (3.20). The equivalent LP (that corresponds to the one in (3.14)) is again (3.19), where the decision variables ρ are constrained to lie in $\mathbf{Q}^\alpha(\beta) \cap \mathsf{M}^\mu$ instead of just $\mathbf{Q}^\alpha(\beta) \cap \mathsf{M}^\mu$.

Remark 3.1 *(Equivalence of a contracting MDP with a discounted CMDP)*
We established in this Section the theory of discounted cost problem as a special case of the absorbing total cost problem. It turns out that the converse is also true for the special case of contracting MDPs. Indeed, Wan Der Wal (1980) 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.

CHAPTER 4

The total cost: Dynamic and Linear Programming

In the previous Chapter we obtained an LP which was seen to be equivalent to **COP**; it yields the same value, and can be used to compute an optimal stationary policy for **COP**. Historically, this LP was first obtained (for the finite state and action spaces) as a dual to another one using dynamic programming techniques (e.g. Kallenberg, 1983), and we denoted it by $\mathbf{DP_1}(\beta)$. We follow in this Chapter a similar approach to obtain $\mathbf{DP_1}(\beta)$, using dynamic programming arguments and Lagrangian techniques. Then, by using a standard saddle point Theorems we show that there is no duality gap between $\mathbf{DP_1}(\beta)$ and $\mathbf{LP_1}(\beta)$.

In obtaining the Linear Program for the general transient case, we establish, in particular, 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 8 and Altman (1993,1994), we do not need here any Slater type condition.

Some analysis of constrained MDPs was obtained in the past by considering directly the Lagrange formulation, for a single constraint, see Beutler and Ross (1985,1986), Sennott (1991,1993). The use of Lagrange techniques for several constraints is quite recent (see e.g. Arapostathis et al., 1993, Piunovskiy (1994), and Altman and Spieksma, 1995), and was not much exploited. Not only does the Lagrangian approach enable 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 7 and 8).

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 (1986).

This chapter is based on Altman (1995a).

4.1 Non-constrained control: Dynamic and Linear programming

We describe in this section the well known 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).

Introduce the dynamic programming equation:

$$\phi(x) \geq \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} \phi(y) \right] =: T_{tc}\phi(x). \tag{4.1}$$

Theorem 4.1 *(Dynamic programming: the transient case)*
Consider the transient framework, (Definition 2.2 and nonnegative immediate costs). Then
(i) The optimal value $C_{tc}(x), x \in \mathbf{X}'$ is the smallest (componentwize) nonnegative solution of (4.1).
(ii) Any stationary deterministic policy g that chooses at state x an action that achieves the minimum of $[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y)]$ is optimal.
(iii) The optimal value $C_{tc}(x), x \in \mathbf{X}'$ achieves (4.1) with strict equality.

Proof: Consider any nonnegative solution ϕ of (4.1) and let w be the stationary policy that chooses at state x an action that achieves the minimum of $[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay}\phi(y)]$. We iterate (4.1) and obtain:

$$\begin{aligned} \phi(x) &\geq c(x,w) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xwy}\phi(y) = c(x,w) + E_x^w \phi(X_2) \\ &\geq c(x,w) + E_x^w \left[c(X_2, A_2) + E_{X_2}^w \phi(X_3) \right] \\ &= c(x,w) + E_x^w c(X_2, A_2) + E_x^w \phi(X_3) \\ &\geq \ldots \geq \sum_{t=1}^{n} E_x^w c(X_t, A_t) + E_x^w \phi(X_{n+1}) \geq C_{tc}^n(x,w) \end{aligned} \tag{4.2}$$

where the last inequality follows from the fact that ϕ is nonnegative. Since (4.2) holds for all integers n, we conclude that $\phi(x) \geq C_{tc}(x)$. On the other hand,

$$\begin{aligned} C_{tc}(x) &= \min_{u \in U_M} C_{tc}(x,u) = \min_{u \in U_M} \left[c(x,u_1) + E_x^u \sum_{t=2}^{\infty} c(X_t, A_t) \right] \\ &= \min_{u \in U_M} \left[c(x,u_1) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xu_1y} C_{tc}(y) \right] \\ &= \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y) \right]. \end{aligned} \tag{4.3}$$

This establishes (i).
If g is a policy as in (ii), then it follows by applying (4.2) with $\phi = C_{tc}$ and $w = g$ that

$C_{tc}(x) \geq C_{tc}(x,g)$, and hence g is optimal, which establishes (ii).
Combining (ii) with (4.3), we get

$$\begin{aligned} C_{tc}(x) &\geq \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y) \right] \\ &= c(x,g) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xgy} C_{tc}(y,g) = C_{tc}(x,g) = C_{tc}(x). \end{aligned}$$

We conclude that (4.3) holds with equalities everywhere, which establishes (iii). ■

In the following example we show that, indeed, (4.1) may have several solutions larger than the value.

Example 4.1 *(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 $\mathcal{T}$:= the expected time it takes the queue to empty (i.e. to reach state 0). In other words, we wish to compute the total expected cost with respect to the immediate cost $c(x) = c(x,a) = 1$ of reaching the set $\mathcal{M} = \{0\}$. The solution satisfies

$$\begin{aligned} \mathcal{T}(0) &= 0 \\ \mathcal{T}(x) &= 1 + \mu \mathcal{T}(x-1) + (1-\mu-\lambda)\mathcal{T}(x) + \lambda \mathcal{T}(x+1), \end{aligned} \tag{4.4}$$

and, in particular, it satisfies (4.1). (4.4) is a difference equation whose solution is given by

$$const \cdot \left[\left(\frac{\mu}{\lambda} \right)^x - 1 \right] + \frac{x}{\mu - \lambda}. \tag{4.5}$$

The expected time to reach 0 is $\mathcal{T}(x) = x/(\mu - \lambda)$. But any other constant in (4.5) yields another solution of (4.1), so $\mathcal{T}(x)$ is not the unique solution, nor the smallest one. It is, however, the smallest nonnegative solution of (4.5) (and of (4.1)).

■

Theorem 4.2 *(Dynamic programming: the contracting case)*
Consider the contracting framework (Definition 2.4). Then
(i) The optimal value $C_{tc}(x), x \in \mathbf{X}'$ is the unique solution of (4.1) in the class of μ-bounded functions; moreover, it achieves (4.1) with strict equality.
(ii) Any stationary deterministic policy g that chooses at state x an action that achieves the minimum of $[c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C_{tc}(y)]$ is optimal.

Proof: The uniqueness of a solution in F^μ follows by observing that T_{tc} (defined in (4.1) is a contracting operator on F^μ, and hence has a unique fixed point. Indeed, for any $f_1, f_2 \in \mathsf{F}^\mu$,

$$|T_{tc}f_1(x) - T_{tc}f_2(x)| \leq \sup_{a \in \mathsf{A}(x)} \mathcal{P}_{xay}|f_1(y) - f_2(y)|,$$

and hence,

$$\|T_{tc}f_1 - T_{tc}f_2\|_\mu \leq \xi \, \|f_1 - f_2\|_\mu .$$

The rest of the proof of (i) is the same as the one of Theorem 4.1; the inequality before the last one in (4.2) follows from the fact that ϕ is μ bounded, and hence, by eq. (2.24) in Lemma 2.3, for any stationary policy u,

$$\sup_{x,n} \frac{E_x^u \phi(X_n)}{\mu(x)\xi^n} < \infty,$$

so that

$$\lim_{n\to\infty} E_x^u \phi(X_n) = 0.$$

By the same argument, the proof of (ii) also follows from the one in Theorem 4.1. ■

Remark 4.1 *By the same arguments as in the proof of Theorem 4.2 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*

$$\phi(x) \geq c(x,w) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xwy}\phi(y) \tag{4.6}$$

and the above inequality is then achieved as equality.

4.2 Superharmonic functions and Linear Programming

Definition 4.1 *(Superharmonic functions)*
Fix some $\mathbf{X}'$. A function $\phi : \mathbf{X}' \to \mathbb{R}$ is called superharmonic (for the total cost criterion) if it satisfies for all $x \in \mathbf{X}'$ and $a \in \mathsf{A}(x)$:

$$\phi(x) \leq c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay}\phi(y), \tag{4.7}$$

and $\phi(x) = 0, x \notin \mathbf{X}'$.

Theorem 4.3 *(The value and superharmonic functions)*
(i) Consider the transient framework, (Definition 2.2 and nonnegative immediate cost). Let ϕ be a superharmonic function. If for some optimal policy g,

$$\lim_{t\to\infty} E_x^g \phi(X_t) \leq 0 \tag{4.8}$$

then the value $C_{tc} \geq \phi$ (componentwize!).
(ii) Consider the contracting framework (Definition 2.4). Then the value C_{tc} is the largest superharmonic function among the μ-bounded functions.

Proof: (i) From Theorem 4.1 it follows that that C_{tc} is a super-harmonic function. Choose a superharmonic function ϕ and let g be an optimal stationary policy satisfying (4.8). Then

$$\begin{aligned}\phi(x) &\leq c(x,g) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xgy}\phi(y) = c(x,g) + E_x^g\phi(X_2)\\ &\leq c(x,g) + E_x^g\left[c(X_2,A_2) + E_{X_2}^g(\phi(X_3))\right]\\ &= c(x,g) + E_x^g c(X_2,A_2) + E_x^g(\phi(X_3))\\ &\leq \ldots \leq \sum_{t=1}^n E_x^g c(X_t,A_t) + E_x^g\phi(X_{n+1}).\end{aligned}$$

(i) is established by taking the limit as $n\to\infty$.
(ii) follows as the proof of (i), by noting that, as in the proof of Theorem 4.2, for any μ-bounded function ϕ and any stationary policy g, $\lim_{t\to\infty} E_x^g\phi(X_t) = 0$. ■

Motivated by Theorem 4.3, we introduce the the following infinite Linear Program with decision variables $\phi(y), y\in\mathbf{X}'$.

$$\begin{aligned}\mathbf{DP}(\beta): &\qquad \text{Find } \phi^* := \sup_\phi \langle\beta,\phi\rangle \text{ s.t.}\\ \phi(x) &\leq c(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay}\phi(y), \qquad x\in\mathbf{X}', a\in\mathsf{A}(x).\end{aligned}$$

For the transient case (where the immediate cost is assumed to be nonnegative), we may further add nonnegativity 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 contracting framework. Theorem 4.3 (ii) implies the following:

Theorem 4.4 *(The dual linear program, contracting framework)*
Consider the contracting framework. Consider $\mathbf{DP}(\beta)$ *where the decision variables are restricted to the set* $\phi\in\mathsf{F}^\mu$. *Then for any initial distribution* β, $\mathbf{DP}(\beta)$ *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 the transient case, when restricting to functions for which the condition (4.8) from Theorem 4.3 (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 to obtain only a lower bound to the optimal value. Our choice of functions ϕ will turn, however, to be rich enough, to obtain the same value as the one obtained by the richer class of policies satisfying (4.8). We begin by considering the absorbing case.

Theorem 4.5 *(The dual linear program, absorbing case)*
Assume that the MDP is absorbing, and the immediate cost are nonnegative. Consider $\mathbf{DP}(\beta)$ *where the decision variables* ϕ *are all (nonnegative) 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, it follows that for any bounded function ϕ, eq. (4.8) holds for all policies: $\lim_{t\to\infty} E_x^g \phi(X_t) = 0$. Theorem 4.3 (i) implies that

$$C^1(x) \le C_{tc}(x) \tag{4.9}$$

for all x.

Let $\mathbf{X}_n$ be an increasing sequence of 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 intial distribution β, $C_{tc}^n(\beta,u)$ is increasing in n, and hence so is $C_{tc}^n(\beta)$. Denote by $C_{tc}^*(\beta)$ the limit of $C_{tc}^n(\beta)$ as n tends to infinity. By Theorem 4.1, we have

$$C_{tc}^n(x) = \min_{a\in\mathsf{A}(x)} \left[c_n(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y) \right], \qquad x \in \mathbf{X}', \tag{4.10}$$

which implies that for all $x \in \mathbf{X}'$ and $a \in \mathsf{A}(x)$ we have

$$C_{tc}^n(x) \le c_n(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y) \le c(x,a) + \sum_{y\in\mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y). \tag{4.11}$$

Thus C_{tc}^n is a bounded super-harmonic function. Hence

$$C_{tc}^*(x) \le C^1(x) \le C_{tc}(x). \tag{4.12}$$

Let a_n be a minimizing action in (4.10) (it is, of course, a function of x), and let a^* be some limit point obtained by diagonalization. By Fatou's Lemma, applied to (4.10), 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 4.1 (i) that $C_{tc}(x) \le C_{tc}^*(x)$, and hence, by (4.12), we have $C_{tc}(x) = C^1(x)$. ■

Next, we introduce the transient case:

Theorem 4.6 *(The dual linear program, transient case)*
Assume that the MDP is transient, and the immediate cost are nonnegative. Consider $\mathbf{DP}(\beta)$ *where the decision variables* ϕ *are all (nonnegative) 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*

$$c(x,a) = 0 \textit{ for all } x \notin \mathbf{Y}, a \in \mathsf{A}(x). \tag{4.13}$$

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 satisfy (4.13). Since the MDP is transient, it follows that for any function ϕ satisfying (4.13) that eq. (4.8) holds for all policies: $\lim_{t\to\infty} E_x^g \phi(X_t) = 0$. Theorem 4.3 (i) implies that

$$C^1(x) \le C_{tc}(x) \tag{4.14}$$

for all x.

Let there be a sequence of finite sets of states $\mathbf{X}_n$, 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:

$$\mathcal{P}_{xay}^n = \begin{cases} \mathcal{P}_{xay} & \text{if } x, y \in \mathbf{X}_n, \\ 1 & \text{if } x \notin \mathbf{X}_n, y = 0, \\ 0 & \text{otherwize}, \end{cases} \qquad c_n(x,a) = c(x,a) 1\{x \in \mathbf{X}_n\}.$$

Here 0 is some arbitrary (possibly new) state which is not in $\mathbf{X}'$. (The immediate costs outside of $\mathbf{X}'$ are of course 0). 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 intial distribution β, $C_{tc}^n(\beta, u)$ is increasing in n, and so is $C_{tc}^n(\beta)$. Denote by $C_{tc}^*(\beta)$ the limit of $C_{tc}^n(\beta)$ as n tends to infinity. By Theorem 4.1, we have

$$C_{tc}^n(x) = \min_{a \in \mathsf{A}(x)} \left[c_n(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay}^n C_{tc}^n(y) \right], \qquad x \in \mathbf{X}', \tag{4.15}$$

which implies that for all $x \in \mathbf{X}'$ and $a \in \mathsf{A}(x)$ we have

$$C_{tc}^n(x) \le c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay}^n C_{tc}^n(y) \le c(x,a) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xay} C_{tc}^n(y). \tag{4.16}$$

Thus C_{tc}^n is a super-harmonic function that vanishes outside of $\mathbf{X}_n$. Hence

$$C_{tc}^*(x) \le C^1(x) \le C_{tc}(x). \tag{4.17}$$

Let a_n be a minimizing action in (4.15) and let a^* be some limit point obtained by diagonalization. By Fatou's Lemma, applied to (4.15), 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 4.1 (i) that $C_{tc}(x) \le C_{tc}^*(x)$, and hence, by (4.17), we have $C_{tc}(x) = C^1(x)$. ∎

Remark 4.2 *(On the solvability of the dual program)*
Unlike the contracting case, the sup in $\mathbf{DP}(\beta)$ *need not be achieved in the setting of Theorem 4.5 or 4.6 (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 4.3 (State truncation)
We have in fact presented in the proof of Theorem 4.6 a state truncation procedure, that enables 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 4.1 (ii). Since this policy is the limit of policies optimal for the truncated MDPs, we conclude that also the policies converge. A different approach to state truncation will be presented in Chapter 8, which will be used for the contracting framework.

The restriction in the dual LP to bounded functions ϕ in the absorbing case (Theorem 4.5), or to functions ϕ converging to zero for the general transient case (Theorem 4.6) are quite necessary. This can be seen from our Example 4.1. Any function among (4.5) are feasible for the dual LP. The supremum over all these unbounded functions is infinity, and not $\mathcal{T}$.

4.3 Set of achievable costs

Define for any $\overline{U} \subset U \cup \overline{M}(U_M)$ the set of achievable vector performance measures:

$$\mathbf{M}^{tc}_{\overline{U}}(\beta) = \cup_{u\in\overline{U}}\{(C_{tc}(\beta,u), D^k_{tc}(\beta,u), k=1,...,K)\}, \tag{4.18}$$

and set $\mathbf{M}^{tc}(\beta) := \mathbf{M}^{tc}_{U}(\beta) \cup \mathbf{M}^{tc}_{\overline{M}(U_M)}(\beta)$. Define also

$$\mathbf{V}_{tc} := \bigcup_{\rho\in\ \mathbf{Q}_{tc}} \{(\langle c,\rho\rangle, \langle d^1,\rho\rangle, \langle d^2,\rho\rangle, ..., \langle d^K,\rho\rangle)\} \tag{4.19}$$

and

$$\mathbf{V}^{\mu}_{tc} := \bigcup_{\rho\in\ \mathbf{Q}^{\mu}_{tc}} \{(\langle c,\rho\rangle, \langle d^1,\rho\rangle, \langle d^2,\rho\rangle, ..., \langle d^K,\rho\rangle)\}. \tag{4.20}$$

Recall the definition $\min B$ from Section 3.1. The next characterization of achievable costs follows from Theorem 3.2.

Theorem 4.7 *(Characterization of the sets of achievable costs)*
(i) For transient MDPs with nonnegative 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 3.1).
(ii) For absorbing MDPs with nonnegative immediate costs, $\mathbf{M}^{tc}_{U_S}(\beta)$ is convex and compact, and satisfies

$$\mathbf{M}^{tc}_{\mathcal{U}}(\beta) = \mathbf{M}^{tc}(\beta) = \mathbf{M}^{tc}_{U_S}(\beta) = co\mathbf{M}^{tc}_{U_D}(\beta) = \min \mathbf{V}_{tc}(\beta).$$

(iii) For contracting MDPs, $\mathbf{M}^{tc}_{U_S}(\beta)$ is convex and compact, and satisfies

$$\mathbf{M}^{tc}_{\mathcal{U}}(\beta) = \mathbf{M}^{tc}(\beta) = \mathbf{M}^{tc}_{U_S}(\beta) = co\mathbf{M}^{tc}_{U_D}(\beta) = \mathbf{V}^{\mu}_{tc}(\beta) = \min \mathbf{V}_{tc}(\beta).$$

4.4 Constrained control: Lagrange approach

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

Theorem 4.8 *(The Lagrangian)*
Consider either the transient framework, (Definition 2.2 and nonnegative immediate cost) or the contracting framework.
(i) The value function satisfies

$$C_{tc}(\beta) = \inf_{u \in U} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) = \inf_{u \in U_M} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) = \inf_{u \in \overline{M}(U_M)} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) \tag{4.21}$$

where

$$\begin{aligned} J_{tc}^{\lambda}(\beta, u) &:= C_{tc}(\beta, u) + \langle \lambda, D_{tc}(\beta, u) - V \rangle = \sum_{t=1}^{\infty} E_{\beta}^{u} j^{\lambda}(X_t, A_t) - \langle \lambda, V \rangle \\ j^{\lambda}(x, a) &:= c(x, a) + \langle \lambda, d(x, a) \rangle. \end{aligned} \tag{4.22}$$

(ii) The value function satisfies

$$C_{tc}(\beta) = \sup_{\lambda \geq 0} \min_{u \in \overline{M}(U_M)} J_{tc}^{\lambda}(\beta, u) = \sup_{\lambda \geq 0} \min_{u \in U_D} J_{tc}^{\lambda}(\beta, u) \tag{4.23}$$

as well as

$$C_{tc}(\beta) = \inf_{u \in U_S} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) = \inf_{u \in \mathcal{U}} \sup_{\lambda \geq 0} J_{tc}^{\lambda}(\beta, u) \tag{4.24}$$

Remark 4.4 *(Nonconvexity of the Lagrangian)*
We note that the Lagrangian $J_{tc}^{\lambda}(\beta, u)$ is, in general, not convex in the policies. However, the set of achievable costs is convex; this lieds us to use a minimax theorem due to Sion (1958). An alternative approach would be to consider J_{tc}^{λ} over the sets $\overline{M}(U_M)$ or $\mathcal{U}$ of policies; in that case we can use other minimax theorems on convex functions, such as in Rockafelar (1989). We shall use this approach later on.

Before stating Sion's mon-max theorem, we introduce some definitions. Let G_1, G_2 be some convex subsets of some linear topological spaces. We say that a function $\Psi : G_1 \times G_2 \to \mathbb{R}$ is quasi-concave in g_1 if the set $\{g_1 : \Psi(g_1, g_2) > r\}$ is convex for every real number r and every $g_2 \in G_2$. We say that it is quasi-convex in g_2 if the set $\{g_2 : \Psi(g_1, g_2) < r\}$ is convex for every real number r and every $g_1 \in G_1$. If both properties are satisfied then Ψ is called quasi-concave-convex.

Lemma 4.1 *(Sion's minimax Theorem)*
Let G_1, G_2 be some convex subsets of some linear topological spaces. Assume that Ψ is

quasi-concave-convex, that it is upper semi-continuous in g_1 and lower semi-continuous in g_2. Further assume that one of the sets G_1 or G_2 is compact. Then

$$\sup_{G_1}\inf_{G_2}\Psi(g_1,g_2)=\inf_{G_2}\sup_{G_1}\Psi(g_1,g_2)$$

We are now ready to prove Theorem 4.8.

Proof of Theorem 4.8: (i) The first equality in (4.21) is standard: if $u\in U$ is feasible (i.e. it satisfies the constraints $D_{tc}(\beta,u)\le V$) then

$$\sup_{\lambda\ge 0} J_{tc}^{\lambda}(\beta,u)=C_{tc}(\beta,u).$$

Therefore,

$$C_{tc}(\beta)\ge\inf_{u\in U}\sup_{\lambda\ge 0}J_{tc}^{\lambda}(\beta,u).$$

If $u\in U$ is not feasible then it is easily seen that

$$\sup_{\lambda\ge 0}J_{tc}^{\lambda}(\beta,u)=\infty\ge C_{tc}(\beta).$$

We conclude that

$$C_{tc}(\beta)=\inf_{u\in U}\sup_{\lambda\ge 0}J_{tc}^{\lambda}(\beta,u).$$

Similarly, let $C'_{tc}(\beta):=\inf C_{tc}(\beta,u)$ over the set $\{u\in U_M : D_{tc}(\beta,u)\le V\}$. Then we have

$$C'_{tc}(\beta)=\inf_{u\in U_M}\sup_{\lambda\ge 0}J_{tc}^{\lambda}(\beta,u).$$

However, it follows from Section 2.3 that $C'_{tc}(\beta)=C_{tc}(\beta)$. This establishes the second equality. The third equality follows by the same arguments.

(ii) We shall apply Lemma 4.1 where G_1 stands for the convex set $\{\lambda\ge 0\}$, and G_2 for the convex and compact set $\overline{M}(U_M)$ (for a discussion on the compactness, see Section 2.2). It follows from Theorem 4.7 (i) that the function $J^{\lambda}(\beta,u):G_1\times G_2\to\mathbb{R}$ is quasi concave-convex. Finally, $J^{\lambda}(\beta,u)$ is continuous (in fact linear) in λ and lower semi-continuous in u (see Lemma 3.5). Hence by Lemma 4.1, we have

$$\sup_{\lambda\ge 0}\min_{u\in\overline{M}(U_M)}J_{tc}^{\lambda}(\beta,u)=\inf_{u\in\overline{M}(U_M)}\sup_{\lambda\ge 0}J_{tc}^{\lambda}(\beta,u)=C_{tc}(\beta)\qquad(4.25)$$

where the last equality follows from (4.21); this establishes the first equality.

For fixed λ, $J^{\lambda}(\beta,u)$ is minimized by a policy in U_D (by Theorem 4.1 (ii)), i.e.

$$\min_{u\in\overline{U}}J_{tc}^{\lambda}(\beta,u)=\min_{u\in U_D}J_{tc}^{\lambda}(\beta,u).\qquad(4.26)$$

for any class of policies $\overline{U}$ that contains U_D. The proof of (4.23) is established by combining this with eq. (4.25).

By (4.26) we have

$$C_{tc}(\beta)=\sup_{\lambda\ge 0}\min_{u\in U_S}J_{tc}^{\lambda}(\beta,u)=\sup_{\lambda\ge 0}\min_{u\in\mathcal{U}}J_{tc}^{\lambda}(\beta,u).\qquad(4.27)$$

In order to obtain the first equality in (4.24), we apply again Lemma 4.1 with G_1 as the convex set $\{\lambda \geq 0\}$, and G_2 as the convex and compact set $\overline{U}$. where $\overline{U}$ stands for either U_S or $\mathcal{U}$. It follows from Theorem 4.7 (ii) that the function $J^\lambda(\beta, u) : G_1 \times G_2 \to \mathbb{R}$ is quasi concave-convex in the over $\overline{U}$ under the contracting case; it is also concave-convex in the transient case when $\overline{U} = \mathcal{U}$. $J^\lambda(\beta, u)$ is continuous (in fact linear) in λ and lower semi-continuous in $u \in \mathcal{U}$ (see Lemma 3.4 and 2.3).

It remains only to establish the first equality in (4.24) for the transient case. This is done by simply noting that **COP** has optimal minimizers within $u \in U_S$ (according to Theorem 3.4). Therefore, following the same type of ideas as the proof of part (i), we establish (4.24). ■

By the same type of arguments as in the proof of part (i) of Theorem 4.8, we obtain from (4.27) the following Corollary for the transient framework (the contracting case was already established in Chapter 3).

Corollary 4.1 *(Dominance of $\mathcal{U}$)*
Consider the transient framework, (Definition 2.2) with non-negative immediate costs. Then $\mathcal{U}$ is a dominating class of policies.

Corollary 4.2 *(Saddle point)*
Consider either the transient framework (Definition 2.2 and nonnegative immediate cost) or the contracting framework (Definition 2.4). Then for any class of policies $\overline{U}$ that contains either U_S or $\mathcal{U}$, we have

$$C_{tc}(\beta) = \sup_{\lambda \geq 0} \min_{u \in \overline{U}} J^\lambda_{tc}(\beta, u) = \min_{u \in \overline{U}} \sup_{\lambda \geq 0} J^\lambda_{tc}(\beta, u) = \sup_{\lambda \geq 0} J^\lambda_{tc}(\beta, u^*)$$

for some $u^* \in \overline{U}$.

Next, we consider the existence of maximizing Lagrangians.

Theorem 4.9 *(The Lagrangian: Slater condition)*
If there exists some policy u for which $D_{tc}(\beta, u) < V$ then there exist nonnegative Lagrange multipliers $\lambda^* = \{\lambda^*_1, ..., \lambda^*_K\}$ such that

$$C_{tc}(\beta) = \min_{u \in \mathcal{U}} J^{\lambda^*}_{tc}(\beta, u) = \min_{u \in U_D} J^{\lambda^*}_{tc}(\beta, u) \tag{4.28}$$

Moreover, any optimal policy u^* satisfies the Kuhn-Tucker conditions:

$$\lambda^*_k(D^k_{tc}(\beta, u^*) - V_k) = 0, \qquad k = 1, ..., K.$$

Proof: $J^\lambda_{tc}(\beta, u)$ is a convex function over the convex set $\mathcal{U}$, and $C_{tc}(\beta, u)$ and $D^k_{tc}(\beta, u)$ are lower semi-continuous in $\mathcal{U}$ (see Lemma 3.5). By standard minimax theory (see e.g. Rockafelar, 1989 p. 45, and Theorems 17 and 18 in p. 41) it follows that there exist nonnegative Lagrange multipliers $\lambda^* = \{\lambda^*_1, ..., \lambda^*_K\}$

$$\min_{u \in \mathcal{U}} J^{\lambda^*}_{tc}(\beta, u) = \sup_{\lambda \geq 0} \min_{u \in \mathcal{U}} J^\lambda_{tc}(\beta, u) = \min_{u \in \mathcal{U}} \sup_{\lambda \geq 0} J^\lambda_{tc}(\beta, u)$$

which equals $C_{tc}(\beta)$, according to Corollary 4.2. The second equality in (4.28) follows from the fact that for fixed λ, $J^\lambda(\beta,u)$ is minimized by a policy in U_D (by Theorem 4.1 (ii)). The Kuhn-Tucker conditions follow from standard arguments (similar to the proof of part (i) of Theorem 4.8, see Rockafelar 1989, Theorem 15). ■

4.5 The dual LP

Consider the DP with decision variables $\phi(y), y \in \mathbf{X}$ and $\lambda \in \mathbb{R}_+^K$.

$$\mathbf{DP_1}(\beta): \qquad \text{Find } \Theta^*(\beta) := \sup_{\phi,\lambda} \langle \beta, \phi \rangle - \langle \lambda, V \rangle \text{ s.t.}$$
$$\phi(x) \quad \leq \quad c(x,a) + \langle \lambda, d(x,a) \rangle + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay}\phi(y), \quad x \in \mathbf{X}, a \in \mathsf{A}(x).$$

We begin by considering the contracting case. Combining Theorem 4.4 with Theorem 4.8, we show that the value of $\mathbf{DP_1}(\beta)$ equals the value of **COP**. This, together with Theorem 3.6, implies that there is no duality gap between $\mathbf{LP}_\mathbf{1}^\mu(\beta)$ and its dual program $\mathbf{DP_1}(\beta)$.

Theorem 4.10 *(The dual LP, the contracting case)*
Consider the contracting framework. Consider $\mathbf{DP_1}(\beta)$ restricted to $\phi \in \mathsf{F}^\mu$. $\mathbf{DP_1}(\beta)$ 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 transient cases with nonnegative immediate costs. By combining Theorems 4.5, 4.6 and 4.8 we have:

Theorem 4.11 *(The dual LP, the absorbing and transient case)*
Consider either
(i) an absorbing MDP with nonnegative immediate costs, with $\mathbf{DP_1}(\beta)$ restricted to bounded ϕ; or
(ii) a transient MDP with nonnegative immediate costs, with $\mathbf{DP_1}(\beta)$ restricted to ϕ satisfying (4.13).
Then $\mathbf{DP_1}(\beta)$ is feasible ($\phi = 0$ is a feasible solution). Moreover, the value of $\mathbf{DP_1}(\beta)$ equals $C_{tc}(\beta)$.

4.6 State truncation

We consider in this Section transient MDPs with nonnegative immediate costs. We already showed in Remark 4.3 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 4.6). 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.

It should thus not be surprising that a similar result holds for the constrained MDP. Indeed, for any $\lambda \geq 0$ we have by Remark 4.3:

$$\lim_{n\to\infty} J_{tc}^{\lambda,n}(\beta) = \sup_{n\in\mathbb{N}} J_{tc}^{\lambda,n}(\beta) = J_{tc}^{\lambda}(\beta) \tag{4.29}$$

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 defind in (4.22) corresponding to the nth truncated MDP. According to Theorem 4.2, 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 (4.29), 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 8 and Altman (1993,1994), we do not need here any Slater type condition.

4.7 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 4.1 (for the transient framework with nonnegative immediate costs) and Theorem 3.4 (ii) (for the contracting case) that $C_{tc}(\beta)$ is the value of **COP** restricted to $\mathcal{U}$:

$$\min_{u\in\mathcal{U}} C_{tc}(\beta,u) \text{ s.t. } D_{tc}(\beta,u) \leq V.$$

This can be rewritten as a Linear Program:

$$\begin{aligned} \mathbf{LP_2}(\beta): \quad & \min_{\gamma\in M_1(U_D)} \int C_{tc}(\beta,u)\gamma(du) \\ \text{s.t.} \quad & \int D_{tc}^{k}(\beta,u)\gamma(du) \leq V^k, \quad k=1,...,K \end{aligned} \tag{4.30}$$

This yields the following:

Theorem 4.12 *(Relation between* **COP** *and* $\mathbf{LP_2}(\beta)$*)*
Consider either the transient or the contracting framework. Then

(i) $\mathbf{COP}$ *is feasible if and only* $\mathbf{LP_2}(\beta)$ *is feasible (i.e. the set satisfying (4.30) is nonempty). If* $\mathbf{LP_2}(\beta)$ *is feasible then there exists an optimal policy in* $\mathcal{U}$ *for* $\mathbf{COP}$.
(ii) The value of $\mathbf{COP}$ *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* $\mathbf{COP}$.

CHAPTER 5

The expected average cost

We study in this chapter the expected average cost. In a similar way as was done for the total cost, we we shall be especially interested in the following frameworks:
(i) the case for which the costs are bounded below, (known as negative dynamic programming) i.e.

$$\begin{array}{l} c(x,a) \text{ and } d^k(x,a), k=1,...,K, \text{ are bounded below, i.e.} \\ \inf_{\kappa\in\mathcal{K}} c(\kappa) \geq \underline{b} \text{ and } \inf_{k,\kappa\in\mathcal{K}} d^k(\kappa) \geq \underline{b} \text{ for some constant } \underline{b} \end{array} \tag{5.1}$$

for that case, an additional growth condition on the cost will be made.
(ii) the contracting framework, for which the cost is assumed to be μ-bounded (2.4), the transition probabilities are μ-continuous (Assumption (2.22)), and the initial distribution satisfies $\langle\beta,\mu\rangle<\infty$.

Remark 5.1 *(The contracting framework: the expected average cost)*
We introduce in this Chapter the notions of μ-uniform geometric recurrence (Definition 5.3) and μ-uniform geometric ergodicity (Definition 5.4) which are slight modifications of the Definition 2.4 of contracting MDPs. In the context of expected average cost, we shall define the "contracting framework" to be the μ-uniform geometric recurrent MDP, together with the above assumptions on the transition probabilities, immediate costs and initial distribution.

We shall assume throughout this Chapter that

- Under any $w \in U_S$, $\mathbf{X}$ contains a single (aperiodic) ergodic class, and absorption into the positive recurrent class takes place in finite expected time. (5.2)

Note that this assumption may restrict the choice of the initial distribution β. Sufficient and necessary conditions for (5.2) in terms of policies in U_D can be found in Fisher (1968).

5.1 Occupation measure

For any given initial distribution β and policy u, define the finite horizon occupation measure $f_{ea}(\beta,u;x,\cdot)$

$$f_{ea}^t(\beta,u;x,\mathcal{A}) = \frac{1}{t}\sum_{s=1}^{t} p_\beta^u(s;x,\mathcal{A}), \qquad \mathcal{A}\in\mathbf{A}(x). \tag{5.3}$$

We set $f^t_{ea}(\beta,u) := \prod_{x\in\mathbf{X}} f^t_{ea}(\beta,u;x,\cdot)$. 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 are interested in the set of limit points of $f^t_{ea}(\beta;u)$ for different u's. We note that for each β,u and t, $f^t_{ea}(\beta,u)$ is a probability measure over $\mathcal{K}$. However, these probability measures need not be tight, so that their limits may contain elements that are sub-probability measures. We denote by $F_{ea}(\beta,u)$ the compact set obtained as all the limits (in the sense of weak convergence of probability measures) of $\{f^t_{ea}(\beta,u)\}$. In case $F_{ea}(\beta,u) = \{f\}$ is a singleton, we denote $f_{ea}(\beta,u) = f$. Any sub-probability measure on $\mathcal{K}$ can be written as

$$f = \delta_f f' \tag{5.4}$$

where $\delta_f \in [0,1]$, and where f' is a probability measure. Define,

$$\begin{aligned} \mathcal{L}_{\overline{U}}(\beta) &= \bigcup_{u\in\overline{U}} \{F_{ea}(\beta,u)\} \text{ for any } \overline{U} \subset U \cup \overline{M}(U_M), \\ \mathbf{Q}_{ea}(\beta) &= \left\{ \rho \in M_1(\mathcal{K}) : \int_{\mathcal{K}} \rho(d\kappa)(\delta_x(\kappa) - \mathcal{P}_{\kappa x}) = 0, x \in \mathbf{X} \right\} \end{aligned} \tag{5.5}$$

where $M_1(\mathcal{K})$ are the set of probability measures over $\mathcal{K}$. We set $\mathcal{L}(\beta) = \mathcal{L}_U(\beta) \cup \mathcal{L}_{\overline{M(U_M)}}(\beta)$. $\mathcal{L}_{\overline{U}}(\beta)$ are called the expected frequencies achievable by $\overline{U}$.

Definition 5.1 *(Completeness)*
A class of policies $\overline{U}$ is called complete for the expected average criterion (for a given initial distribution β) if

$$\mathcal{L}(\beta) = \mathcal{L}_{\overline{U}}(\beta) \textit{ and } \forall F \in \mathcal{L}_{\overline{U}}(\beta),\ F \textit{ is a singleton }.$$

It is called weakly complete if

$$\mathcal{L}(\beta) \cap M_1(\mathcal{K}) = \mathcal{L}_{\overline{U}}(\beta)$$

$$\textit{and } \forall F \in \mathcal{L}_{\overline{U}}(\beta),\ F \textit{ is a singleton.}$$

Thus, a complete class of policies $\overline{U}$ has the property that the achievable expected frequencies under $\overline{U}$ is the same as under all policies. A weakly complete class of policies achieves all those expected frequencies that have mass 1 over $\mathcal{K}$.

Definition 5.2 *For any sets B_1, B_2 of sub-probability measures over $\mathcal{K}$, 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 $\delta_{f_1} \le \delta_{f_2}$ (where f' and δ_f are defined in (5.4)).*

Theorem 5.1 *(Weakly 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 to some limit $f \in F_{ea}(\beta,u)$. Define γ that maps states y to measures over $\mathsf{A}(y)$:

$$\gamma_y(\mathcal{A}) = \frac{f(y,\mathcal{A})}{f(y,\mathsf{A}(y))}, \quad \mathcal{A} \subset \mathsf{A}(y)$$

whenever the denominator is nonzero. 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 (5.2) 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

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

and hence, $F_{ea}(\beta,u) = \{f^w\}$ is a singleton given by

$$f^w(y,\mathcal{A}) = w_y(\mathcal{A})\pi_w(w). \tag{5.6}$$

We show that $f^w = \delta f$ for some $\delta \in [0,1]$. It follows from (2.3) that for any $x \in \mathbf{X}$,

$$f^t_{ea}(\beta,u;x) - \frac{\beta(x)}{t} = \int_{\mathcal{K}} f^t_{ea}(\beta,u;d\kappa)\mathcal{P}_{\kappa x} - \frac{\int_{\mathcal{K}} p^u_\beta(t;d\kappa)\mathcal{P}_{\kappa x}}{t} \tag{5.7}$$

By applying Fatou's Lemma, we get from (5.7)

$$f(x,\mathsf{A}(x)) = \lim_{n\to\infty} f^{t_n}_{ea}(\beta,u;x) = \lim_{n\to\infty}\int_{\mathcal{K}} f^{t_n}_{ea}(\beta,u;d\kappa)\mathcal{P}_{\kappa x} \geq \int_{\mathcal{K}} f(d\kappa)\mathcal{P}_{\kappa x}. \tag{5.8}$$

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

$$\int_{\mathcal{K}} f(d\kappa)\mathcal{P}_{\kappa x} = \sum_y f(y,\mathsf{A}(y))\int_{\mathsf{A}(y)} \gamma_y(da)\mathcal{P}_{yax} = \sum_y f(y,\mathsf{A}(y))P_{yx}(w), \tag{5.9}$$

which, together with (5.8) leads to

$$f(x,\mathsf{A}(x)) \geq \sum_y f(y,\mathsf{A}(y))P_{yx}(w), \tag{5.10}$$

Measures satisfying (5.10) are called excessive measures; $\pi(w)$ is known to be the unique probability measure over $\mathbf{X}$ satisfying the inequality (5.10), (this is a straight forward extension of Proposition 6.4 in Kemeney, Snell and Knapp (1976), see Altman and Shwartz (1991a)). This, together with the definition of γ, imply that $\{f\} \propto \{f^w\}$, which establishes the proof. ∎

5.2 The contracting framework

It turns out that under the contracting framework, the occupation measures satisfy useful continuity, tightness and uniform integrability properties, which we describe below. They will allow us to show that the stationary policies are in fact complete. We begin by relating the contracting framework to notions of uniform ergodicity and recurrence.

Definition 5.3 *(μ-Uniform geometric recurrence)*
An MDP is μ-uniform geometric recurrent if it satisfies Definition 2.4 (contracting MDP) with the set $\mathcal{M}$ being finite, and where (2.21) is replaced by the weaker requirement that

$$\sum_{y \notin \mathcal{M}} [P^{n_0}(w)]_{xy} \mu(y) \leq \xi \mu(x).$$

for all $w \in U_D$ and for some integer n_0, where $P^n(w)$ is the n step transition probabilities under policy w.

Definition 5.4 *(μ-Uniform geometric ergodicity)*
The MDP is said to be μ-uniform geometrically ergodic if the exists some constants $\sigma > 0$ and $\tilde{\xi} < 1$ such that for all $u \in U_S$,

$$\begin{cases} \|P^n(u) - \Pi(u)\|_\mu \leq \sigma \tilde{\xi}^n, & \forall n \in \mathbb{N}, \\ \|P(u)\|_\mu \leq \sigma \end{cases} \tag{5.11}$$

where $P^n(u)$ is the n step transition probabilities under policy u, and $\Pi(u)$ matrix, the rows of which are equal to the steady state probabilities under u.

We present bellow a part of the remarkable equivalence relation which was established in Spieksma (1990) Chapter 6 and Dekker et al. (1994).

Theorem 5.2 *(μ-uniform geometric ergodicity and recurrence)*
Assume that (5.2) holds. The MDP is μ-uniform geometrically recurrent if and only if it is uniformly geometrically ergodic.

Next we present uniform tightness and integrability properties of contacting MDPs.

Definition 5.5 *(Tightness)*
A set of probability measures $\{f^n\}_{n \in I}$ (I is some set) over $\mathcal{K}$ is called tight if for any $\epsilon > 0$ there exist some compact set $K_\epsilon \in \mathbb{K}$ such that

$$f^n(K_\epsilon) > 1 - \epsilon, \qquad \forall n \in I.$$

Definition 5.6 *(Uniform integrability)*
A set of probability measures $\{f^n\}_{n \in I}$ (I is some set) over $\mathcal{K}$ is said to be uniformly integrable with respect to c if for any $\delta > 0$ there exist some $\epsilon > 0$ such that for any $B \in \mathbb{K}$, t and u such that $f^t_{ea}(\beta, u; B) < \epsilon$,

$$\int_B f^t_{ea}(\beta, u; d\kappa) c(\kappa) < \delta.$$

We have (Billingsley (1968) Theorem 6.1):

Lemma 5.1 *(Characterization of tightness)*
$\{f^t_{ea}(\beta, u)\}_{t \in \mathbb{N}}$ are tight probability measures iver $\mathcal{K}$, if and only if for any converging subsequence $\{f^{t_n}_{ea}(\beta, u)\}_{n \in \mathbb{N}}$, its limit f satisfies $f(\mathcal{K}) = 1$.

Lemma 5.2 *(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 $\mu(y, a) = \mu(y)$.

Proof: The uniform integrability is a direct extension of Lemma 6.8 in Spieksma (1990), who restricted to Markov policies, to fixed initial states, and to uniform integrability with respect to μ. The generalization to any policy follows from Theorem 2.1. The proof in Spieksma (1990) extends in a straight forward way to any initial distribution (satisfying of course $\langle \beta, \mu \rangle < \infty$). The tightness follows from the uniform integrability (see Lemma 6.5 in Spieksma (1990)). ■

If the contracting framework is not assumed, tightness of the occupation measures over the original space $\mathcal{K}$ need not hold (even for a fixed policy u). The Fisher & Ross (1968) Counter Example illustrates such a situation. This example was further analyzed by Spieksma (1990) in Section 11.4, called "The importance of being tight"; she showed for some "badly" behaving policy u that $f_{ea}(x, u; \mathcal{K}) = 0.7$.

Finally, we present a result by Spieksma (1990, Proposition 5.1 p. 97) that establishes the continuity of the occupation measure over the set of stationary policies.

Theorem 5.3 *(Continuity of occupation measures)*
Consider the contracting framework. Then the map $f_{ea}(\beta, \cdot) : U_S \to \mathcal{L}_{U_S}$ is μ-continuous.

Proof: Follows from arguments similar to those in Lemma 3.1 (ii); for any $u \in U_S$ we use $f_{ea}(\beta, u)$ to denote the singleton $F_{ea}(\beta, u) = \{f_{ea}(\beta, u)\}$. We shall show that

$$\lim_{n\to\infty} \sum_{y\in\mathbf{X}} \sup_{\mathcal{A}\subset \mathsf{A}(y)} |f_{ea}(\beta, u^n; y, \mathcal{A}) - f_{ea}(\beta, u; y, \mathcal{A})| \, \mu(y) = 0.$$

By (3.7) in Lemma 3.1 (ii), $p^u_\beta(m; \cdot)$ are μ-continuous in U_S. Moreover, due to Definition 5.11, they converge to $f_{ea}(\beta, u; \cdot)$ uniformly in U_S. This finally implies the μ-continuity of the map $f_{ea}(\beta, \cdot) : U_S \to \mathcal{L}_{U_S}$, see Royden (1988, Problem 8.3.17). ■

5.3 Completeness properties of stationary policies

Theorem 5.4 *(Completeness of stationary policies)*
(i) $\mathcal{L}(\beta)$ is convex and satisfies

$$\mathcal{L}(\beta) = \mathcal{L}_{U_M}(\beta) \propto \mathcal{L}_{U_S}(\beta) = \mathbf{Q}_{ea}(\beta).$$

(ii) In the contracting framework, $\mathcal{L}_{U_S}(\beta)$ is convex and compact, and satisfies

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

Hence the stationary policies are complete.

In order to prove the theorem, we need the following Lemma (that corresponds to Lemma 3.2 in the case of total cost). Its proof is quite technical, and is a straight forward extension of Theorem 5.1 in Altman and Shwartz (1991a), or of Key Lemma in Spieksma (1990) p. 168.

Lemma 5.3 *(Splitting in a state)*
Choose $w \in U_S$ and a state y. Define $w^a \in U_S$ to be the policy that chooses always 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

$$f_{ea}(\beta, w) = \int_{\mathsf{A}(y)} \gamma(da) f_{ea}(\beta, w^a).$$

Proof of Theorem 5.4: (i) Theorem 2.1 implies that $\mathcal{L}(\beta)$ is convex, and that $\mathcal{L}(\beta) = \mathcal{L}_{U_M}(\beta)$. Theorem 5.1 implies that $\mathcal{L}_{U_M}(\beta) \propto \mathcal{L}_{U_S}(\beta)$. Since for each $w \in U_S$, (5.10) is obtained with equality (see paragraph below (5.10)), it follows from (5.9) and (5.10) 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 nonzero. When it is zero, $\gamma_y(\cdot)$ is chosen arbitrarily. Define the stationary policy w as $w_x(\mathcal{A}) = \gamma_y(\mathcal{A})$. 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 $\pi_y(w) = f_{ea}(\beta, u; y)$, $y \in \mathbf{X}$ is the unique solution to $\pi = \pi P_{yx}(w)$, it follows that $\rho(x, A(x)) = f_{ea}(\beta, u; x)$ for all $x \in \mathbf{X}$, and by the definition of γ, $\rho = f_{ea}(\beta, u)$. This establishes $\mathcal{L}_{U_S}(\beta) = \mathbf{Q}_{ea}(\beta)$.
(ii) Choose some policy u and initial distribution β. By Theorem 5.1, 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 5.1 and 5.2 that for any $f \in F_{ea}(\beta, u)$, $f(\mathcal{K}) = 1$. This implies that $\delta = 1$, so that $f^w = f$. Consequently $\mathcal{L}(\beta) = \mathcal{L}_{U_S}(\beta)$.

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 $\mathcal{K}$. Our aim is to find a policy u such that $F_{ea}(\beta, u) = \{f\}$. By Lemmas 5.1 and 5.2, this implies that $f(\mathcal{K}) = 1$.

By Theorem 5.1, there exists a stationary policy g_i that achieves f_i, i.e. $F_{ea}(\beta, g_i) = \{f_i\}$. Let $\epsilon_i := d(f, f_i)$, so that $\lim_{n \to \infty} \epsilon_n = 0$. Consider the nonstationary policy u, that uses g_1 until the time $t_1 := \min\{t : d(f_1, f_{ea}^t(\beta, u)) \le \epsilon_1\}$, and uses g_i between $t_{i-1} + 1$ and t_i, where $t_i := \min\{t > t_{i-1} : d(f_i, f_{ea}^t(\beta, u)) \le \epsilon_i\}$. The fact that $t_n < \infty$ can be proved by induction using the following fact. Suppose the policy u uses g_n from time s onward, and let $\chi_s(z) = p_\beta^u(s, z)$. Then

$$f_{ea}^t(\beta, u; y, \mathcal{A}) = \frac{s}{t} f_{ea}^s(\beta, u; y, \mathcal{A}) + [g_n]_y(\mathcal{A}) \frac{t-s}{t} \sum_{z \in \mathbf{X}} \chi_s(z) \left(\sum_{r=1}^{t-s} [P^r(g_n)]_{zy} \right),$$

(were $P(g_n)$ is the transition probabilities matrix under g_n). It then follows easily that $\lim_{t\to\infty} F^t_{ea}(\beta,u)(y,\mathcal{A}) = \{f_n\}$. Thus

$$d(f, f^{t_n}_{ea}(\beta,u)) \le d(f,f_n) + d(f_n, f^{t_n}_{ea}(\beta,u)) \le 2\epsilon_n$$

and we obtain

$$\lim_{n\to\infty} f^{t_n}_{ea}(\beta,u) = f. \tag{5.12}$$

Since $f(\mathcal{K}) = 1$ (due to Lemmas 5.1 and 5.2), $\mathcal{L}$ is closed and sequentially compact, hence compact.

Next we show that $\mathcal{L}_{U_S}(\beta)$ is equal to the convex hull of $\mathcal{L}_{U_D}(\beta)$. Since it is compact, by the Krein-Milman theorem it is the convex hull of its extreme points. Choose some $w \in U_S$. Suppose that w is not deterministic, so that $w_y(\cdot)$ is not concentrated on a single point in $\mathsf{A}(y)$. But then by Lemma 3.2, w is not an extreme point of $\mathcal{L}_{U_S}$. ■

5.4 Relation between cost and occupation measure

For the contracting framework, assumption (2.4) will suffice to obtain similar linear representation of the cost as was obtained for the total cost case (Section 3.2). When the contracting framework is not assumed, we shall use assumption (5.1) to show that U_S and $\mathcal{U}$ have these properties. For other policies, that representation will not hold in general. To illustrate that, $c(\kappa) = 1$ for all $\kappa \in \mathcal{K}$, consider a policy u for which the occupation measures are not tight (e.g. the Fisher & Ross (1968) Counter Example). Then we may typically have

$$1 = C_{ea}(\beta,u) = C^t_{ea}(\beta,u) > \langle c, f\rangle, \qquad \forall t \in \mathbb{N}.$$

We have the following properties of the expected average costs (see Altman and Shwartz, 1991a, Lemma 2.3):

Theorem 5.5 *(Linear representation of the cost)*
(i) Consider the contracting framework. Then for any β, $u \in U$, and $f \in F_{ea}(\beta,u)$,

$$C_{ea}(\beta,u) \ge \langle c, f\rangle := \int_{\mathcal{K}} c(\kappa) f(d\kappa) \tag{5.13}$$

with strict equality holding for some $f \in F_{ea}(\beta,u)$; the expected average costs are uniformly μ-bounded over all policies:

$$\sup_u \|C_{ea}(\cdot,u)\|_\mu < \infty \tag{5.14}$$

where $C_{ea}(\cdot,u)$ is the vector of expected average cost corresponding to all initial states.
(ii) Fix some β and $u \in U_S$ or $u \in \mathcal{U}$ and assume that the cost is bounded below (i.e. (5.1) holds). 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 (5.13) holds with equality. Moreover, for any policy u and $f \in F_{ea}(\beta,u)$, (5.13) holds.

Proof: (i) Choose any $u \in U$ and let t_n be some sequence along which the limit $f = \lim_{n\to\infty} f_{ea}^{t_n}(\beta, u)$ exists. Then

$$\begin{aligned} C_{ea}(\beta,u) &= \overline{\lim_{t\to\infty}}\, C_{ea}^t(\beta,u) = \overline{\lim_{t\to\infty}} \int f_{ea}^t(\beta,u;d\kappa)c(\kappa) \\ &\geq \overline{\lim_{n\to\infty}} \int f_{ea}^{t_n}(\beta,u;d\kappa)c(\kappa). \end{aligned}$$

Due to the uniform integrability of $f_{ea}^{t_n}(\beta,u;d\kappa)$ w.r.t. the cost c (Lemma 5.2), the integration and limit may be interchanged, see Billingsely (1968, Theorem 5.4). This establishes (5.13). Equality is obtained in (5.13) by choosing t_n so as to achieve the limsup:

$$\overline{\lim_{t\to\infty}} \int f_{ea}^t(\beta,u;d\kappa)c(\kappa) = \lim_{n\to\infty} \int f_{ea}^{t_n}(\beta,u;d\kappa).$$

and so that a limit $f = \lim_{n\to\infty} f_{ea}^{t_n}(\beta,u)$ exists. (That such a choice is possible follows from diagonalization arguments). Finally, (5.14) follows from Lemma 5.2. This establishes (i).

(ii) 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, pp. 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 (5.2)). 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 till state z is first reached. Hence, by the first part of the proof, (5.13) holds. The result is then obtained by the monotone convergence theorem.

Finally, to establish (5.13) for the case of cost bounded below, we choose an arbitrary u, and $f' \in F_{ea}(\beta,u)$, and choose a sequence t_n such that $f' = \lim_{n\to\infty} f_{ea}^{t_n}(\beta,u)$. By applying Fatou's Lemma we obtain

$$\begin{aligned} C_{ea}(\beta,u) &= \overline{\lim_{t\to\infty}} \int f_{ea}^t(\beta,u;d\kappa)c(d\kappa) \geq \overline{\lim_{n\to\infty}} \int f_{ea}^{t_n}(\beta,u;d\kappa)c(d\kappa) \\ &\geq \int \underline{\lim_{n\to\infty}}\, f_{ea}^{t_n}(\beta,u;d\kappa)c(d\kappa) = \langle c, f' \rangle. \end{aligned}$$

■

Next we describe continuity properties of the expected average cost.

Lemma 5.4 *(Continuity of the cost, the contracting framework)*
Consider the contracting framework. Then $C_{ea}(\beta,u)$ is continuous on U_S and on $\mathcal{U}$.

Proof: For U_S, this is an immediate consequence of the μ-continuity of the occupation measures (Theorem 5.3), and from the fact that (5.13) holds with equality for $u \in U_S$. The proof for $\mathcal{U}$ is as in Lemma 3.5. ■

Lemma 5.5 *(Lower semi-continuity of the cost, immediate cost bounded below)*
Assume that the cost is bounded below (i.e. (5.1) holds). $C_{ea}(\beta,\cdot)$ is lower semi-continuous on $\overline{M}(U_M)$, and in particular on $\mathcal{U}$.

Proof: Let $q^n, n = 1, 2, ...$ and q be probability measures over U_M, and let $\hat{q}^n$ and $\hat{q}$ be the corresponding policies in $\overline{M}(U_M)$. Assume that $\hat{q}^n$ converges to $\hat{q}$ (by which we mean that q^n converges to q weakly). Then by Fatou's Lemma,

$$\varliminf_{n\to\infty} C_{ea}(\beta, \hat{q}^n) = \varliminf_{n\to\infty} \langle q^n, C_{ea}(\beta,\cdot)\rangle \geq \langle q, C_{ea}(\beta,\cdot)\rangle = C_{ea}(\beta, \hat{q}).$$

■

5.5 Dominating classes of policies

Theorem 5.6 *(Relation between complete and dominating policies)*
Consider the contracting framework. 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 3.4 (ii). (The basic steps can be found in Altman and Shwartz, 1991a, Theorem 2.8 and Corollary 5.4.) ■

Following an idea by Borkar (1983), which was adapted to constrained MDPs in Altman and Shwartz (1991a), we introduce the following growth condition on the costs: 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

$$\varliminf_{i\to\infty} \{c(\kappa); \kappa \notin \mathcal{K}_i\} = \infty. \tag{5.15}$$

Note that condition (5.15) implies that c are bounded below by some $\underline{b}$. (Note that c and d achieve their minimum over each compact set $\mathcal{K}_i$, since they are continuous in $\mathcal{K}$, bu (2.1)).

A sufficient condition for (5.15), which is frequently used in the literature (e.g. Cavazos-Cadena (1989), Cavazos-Cadena and Sennott (1992)), is the following moment condition:

$$\forall \ell \in \mathbb{R}, \text{ the set } \{x \in \mathbf{X} : \inf_a c(x,a) < \ell \text{ is finite }\} \tag{5.16}$$

Lemma 5.6 *(Tightness of occupation measure for stationary policies)*
The sets $\{f^t_{ea}(\beta,u)\}_{t\in\mathbb{N}}$ are tight for any $u \in U_S$ and any $u \in \mathcal{U}$.

Proof: Assumption (5.2) implies that the steady state probabilities exit under any $u \in U_S$, so that $\sum_{x\in\mathbf{X}} f_{ea}(\beta,u;x) = f_{ea}(\beta,u;\mathcal{K}) = 1$. Lemma 5.1 then implies the required tightness. A similar argument implies the tightness for $\mathcal{U}$. ■

Theorem 5.7 *(Dominance of classes of policies)*
Assume that the growth condition (5.15) holds for the immediate cost c or for d^k, for some $k = 1, ..., K$.
(i) Let $\overline{U}$ be a class of policies that is weakly complete and for which (5.13) holds with equality for both immediate costs c as well as d^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.

Proof: (i) The proof follows Altman and Shwartz (1991a), p. 801, Theorem 4.4. It clearly suffices to show that for any u for which $C_{ea}(\beta, u) < \infty$ and $D^k_{ea}(\beta, u) < \infty,\ k = 1, ..., K$, there exists some $u' \in \overline{U}$ such that

$$C_{ea}(\beta, u') \leq C_{ea}(\beta, u) \qquad D_{ea}(\beta, u') \leq D_{ea}(\beta, u).$$

Thus, assume without loss of generality that (5.15) 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 $\epsilon > 0$ and an increasing sequence $\{t_l\}$ such that $f^{t_l}_{ea}(\mathcal{K}^c_l) > \epsilon$. Denote $c_j := \inf\{c(\kappa) : \kappa \notin \mathcal{K}_j\}$. It follows that

$$C^{t_j}_{ea}(\beta, u) \geq c_j\epsilon + \max(\underline{b}, 0), \quad j \in \mathbb{N}.$$

Since by (5.15), $\lim_{j\to\infty} c_j = \infty$, 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. If $\overline{U}$ is a weakly complete class of policies, then there exists some $u' \in \overline{U}$ such that $f_{ea}(\beta, u') = f$ (this follows from Lemma 5.6). The last part of Theorem 5.5 and the assumption that (5.13) holds with equality for u' implies that u' dominates u. The statement for U_S follows from Theorem 5.1 $\overline{U}$ can be chosen as U_S.
(ii) We delay the proof for $\mathcal{U}$ to the next chapter (Corollary 6.1). ∎

Definition 5.7 *(Strongly dominance)*
A class of policies $\overline{U}$ is said to be a strongly dominating class of policies for the expected average cost, for a given initial distribution β, if the following holds: For any policy $u \in U$ there exists a policy $\overline{u} \in \overline{U}$ such that, for any increasing sequence t_n such that the limits

$$\begin{aligned}\overline{C}_{ea}(\beta, u) &= \lim_{n\to\infty} \frac{1}{t_n} \sum_{s=1}^{t_n} E^u_\beta c(X_t, A_t),\\ \overline{D}^k_{ea}(\beta, u) &= \lim_{n\to\infty} \frac{1}{t_n} \sum_{s=1}^{t_n} E^u_\beta d^k(X_t, A_t)\end{aligned}$$

exist (they may depend on the sequence t_n),

$$C_{ea}(\beta, \overline{u}) \leq \overline{C}_{ea}(\beta, u), \qquad \text{and} \qquad D_{ea}(\beta, \overline{u}) \leq \overline{D}^k_{ea}(\beta, u). \tag{5.17}$$

When (2.16) holds, we say that $\overline{u}$ strongly dominates u.

Note that strongly dominance implies dominance.

By observing the proofs of Theorems 5.7 we may conclude that one may replace dominance by strongly dominance of $\overline{U}$. This allows us to further strengthen Theorem 5.7, and establish the existence of an optimal policy among $\overline{U}$. (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.

Theorem 5.8 *(Strongly-dominance and existence of optimal policies)*
Under the conditions of Theorem 5.7,
(i) $\overline{U}$ is a strongly dominating class.
*(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: (i) is a straightforward extension of the proof of Theorem 5.7.
(ii) Assume that **COP** is feasible, and let g_i be a sequence of stationary policies which are ϵ_i optimal, where $\lim_{i\to\infty}\epsilon_i = 0$. 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 5.4 above (5.12); 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

$$\lim_{n\to\infty} f_{ea}^{t_n}(\beta, u) = f$$

(f is not necessarily a probability measure) and moreover,

$$\overline{C}_{ea}(\beta, u) := \lim_{n\to\infty} C_{ea}^{t_n}(\beta, u) = \lim_{n\to\infty} C_{ea}(\beta, g_n) = C_{ea}(\beta),$$

$$\overline{D}_{ea}^{k}(\beta, u) := \lim_{n\to\infty} D_{ea}^{k,t_n}(\beta, u) = \lim_{n\to\infty} D_{ea}^{k}(\beta, g_n) \le V_k \ k = 1, ..., K.$$

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 5.7 (ii). ∎

Theorem 5.7 holds when the growth condition (5.15) is replaced by a weaker condition, due to Borkar (1983), which was applied to constrained MDPs in Altman and Shwartz (1991a) Section 4.

Definition 5.8 *(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

$$\underline{\lim}_{i\to\infty} \{c(\kappa); \kappa \notin \mathcal{K}_i\} \ge V. \tag{5.18}$$

Theorem 5.9 *(Weakly-completeness and dominance)*
Assume that $\overline{U}$ is weakly complete and that (5.13) 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. $D^k_{ea}(\beta,u') \le V_k$, *and define* $V_0 := C_{ea}(\beta,u')$. *If c is V_0-weakly monotone and d^k are V_k weakly monotone for all* $k = 1,...,K$, *then*
(i) $\overline{U}$ *is a strongly 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 5.7 and 5.8. 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 $\mathcal{U}$. ■

5.6 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 a LP with countable number of decision variables and a countable number of constraints. Consider the following LP:
$\mathbf{LP_3}(\beta)$: Find the infimum $\mathcal{C}^*$ of $\mathcal{C}(\rho) := \langle c, \rho \rangle$ subject to:

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

where $\mathbf{Q}_{ea}(\beta)$ was defined in (5.5). 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 nonzero. We show that there is a one to correspondence between feasible (and optimal) solutions of the LP, and the feasible (and optimal) solutions of **COP**.

Theorem 5.10 *(Equivalence between COP and* $\mathbf{LP_3}(\beta)$*)*
Assume either
(A1): the immediate cost is bounded below (5.1) and satisfies the growth condition (5.15) or (5.18); moreover, for any stationary policy u, the total expected cost to reach some recurrent state z is either finite, or $C_{ea}(z,u) = \infty$. *Or*
(A2): the contracting framework holds (and, in particular, the immediate costs are μ-bounded). 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* $D_{ea}(\beta,w(\rho)) = \mathcal{D}(\rho)$, *with equalities for the contracting case.*
(iii) $\mathbf{LP_3}(\beta)$ *is feasible if and only if* **COP** *is. Assume that* **COP** *is feasible. Then there exists an optimal solution* ρ^* *for* $\mathbf{LP_3}(\beta)$, *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 5.4), by combining

Theorem 5.4 with Theorems 5.5 (ii), 5.6, 5.7 and 5.9. The claims on the costs follow from Theorem 5.5. 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 Theorem 5.4)), and by applying again Theorem 5.5. This establishes (ii), and thus implies (i). (iii) now follows from (ii) and Theorems 5.6, 5.7, 5.8 and 5.9. ■

5.7 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 nonnegative vectors $\lambda \in \mathbb{R}_+^K$.

$$\mathbf{DP_3}(\beta): \qquad \text{Find } \Theta^*(\beta) := \sup_{\psi,\phi,\lambda} \psi - \langle \lambda, V \rangle \text{ s.t.}$$

$$\phi(x) + \psi \quad \leq \quad \left[c(x,a) + \langle \lambda, d(x,a) \rangle + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y) \right], \quad x \in \mathbf{X}, a \in \mathsf{A}(x).$$

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

$$\Theta^* = \mathcal{C}^* = C_{tc}(\beta), \tag{5.19}$$

for both the contracting framework, and the case of costs bounded below. For the contracting framework, we shall restrict to $\phi \in \mathsf{F}^\mu$, and for the case of costs bounded below, a possible choice is to restrict to bounded ϕ.

CHAPTER 6

Expected average cost: Dynamic and Linear Programming

We present in this Chapter dynamic programming, similar to those in Chapter 4, for the unconstrained control problem, and then, using Lagrangian methods and duality methods, obtain a the linear program $\mathbf{DP_3}(\beta)$, which is the dual of the one obtained in the previous Chapter. We show again that there is no duality gap for both the contracting case as well as the case of costs bounded below. As in Chapter 4, we conclude by presenting a different LP approach for computing the optimal values and optimal mixed strategies. This chapter is based on Altman (1995b).

6.1 The non-constrained case: optimality equation

Introduce the (expected) Average Cost Optimality Inequality:

$$\mathbf{ACOI}: \qquad \phi(x) + \psi \geq \min_{a \in \mathsf{A}(x)} \left[c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y) \right], \tag{6.1}$$

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 getting into the 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.

Lemma 6.1 *(Upper bound on the value)*
Let (ψ, ϕ) be a solution of (6.1) and let w be a stationary policy that chooses at state x an action that achieves the minimum of

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

Assume that ϕ satisfies

$$\underline{\lim}_{n \to \infty} \frac{E_x^w \phi(X_n)}{n} \geq 0. \tag{6.2}$$

Then $\psi \geq C_{ea}(x,w)$, and hence $\psi \geq C_{ea}(x)$.

Proof: We iterate (6.1) and obtain:

$$\begin{aligned}
\phi(x) &\geq -\psi + c(x,w) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xwy}\phi(y) = -\psi + c(x,w) + E_x^w \phi(X_2) \\
&\geq -2\psi + c(x,w) + E_x^w \left[c(X_2, A_2) + E_{X_2}^w \phi(X_3) \right] \\
&= -2\psi + c(x,w) + E_x^w c(X_2, A_2) + E_x^w \phi(X_3) \\
&\geq \ldots \geq -n\psi + \sum_{t=1}^{n} E_x^w c(X_t, A_t) + E_x^w \phi(X_{n+1}) \qquad (6.3)
\end{aligned}$$

Dividing in (6.3) by n and going to the limit as n tends to infinity, we conclude that $\psi \geq C_{ea}(x,w) \geq C_{ea}(x)$. ■

Remark 6.1 *Clearly, a sufficient condition for (6.2) to hold is that ϕ is bounded from below.*

Definition 6.1 *(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 $a \in \mathsf{A}(x)$:

$$\phi(x) + \psi \leq c(x,a) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xay}\phi(y). \qquad (6.4)$$

Lemma 6.2 *(Lower bound on the value)*
Assume that there exists some superharmonic pair (ψ, ϕ) such that

$$\overline{\lim_{n\to\infty}} \frac{E_x^u \phi(X_n)}{n} \leq 0 \qquad (6.5)$$

for some Markov policy u. Then $\psi \leq C_{ea}(x,u)$.

Proof: Iterating (6.5), we get

$$\begin{aligned}
\phi(x) &\leq c(x,u_1) - \psi + \sum_{y\in\mathbf{X}} \mathcal{P}_{xu_1y}\phi(y) = c(x,u_1) - \psi + E_x^u \phi(X_2) \\
&\leq c(x,u_1) - 2\psi + E_x^u \left[c(X_2, A_2) + E_{X_2}^u (\phi(X_3)) \right] \\
&= c(x,u_1) - 2\psi + E_x^u c(X_2, A_2) + E_x^u (\phi(X_3)) \\
&\leq \ldots \leq \sum_{t=1}^{n} E_x^u c(X_t, A_t) - n\psi + E_x^u \phi(X_{n+1}).
\end{aligned}$$

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 (6.1) holds in fact with equality. Consider the (expected) Average Cost Optimality Equation:

$$\textbf{ACOE}: \qquad \phi(x) + \psi = \min_{a\in\mathsf{A}(x)} \left[c(x,a) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xay}\phi(y) \right], \qquad (6.6)$$

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 Lemma 6.1 and 6.2 to get the following optimality results:

Lemma 6.3 *(Characterization of optimal value and policy)*
Assume that there exists a pair (ψ, ϕ) satisfying the ACOE (6.6), and that (6.5) 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

$$\underline{\lim}_{n\to\infty} \frac{E_x^w \phi(X_n)}{n} \geq 0.$$

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 the pair (C_{ea}, ϕ') is super-harmonic and for which (6.5) holds for all Markov policies u.

The following converse can be found in Arapostathis et al. (1993):

Lemma 6.4 *(The converse)*
Assume that there exists a pair (ψ, ϕ) satisfying the ACOE (6.6), 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

$$c(x,g) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xgy}\phi(y) = \min_{a\in\mathsf{A}(x)} \left[c(x,a) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xay}\phi(y) \right]. \tag{6.7}$$

Proof: Let $g \in U_S$ be optimal and assume that the state is irreducible and positive recurrent and that (6.7) does not hold. Then there exists some state x_0 and action $a_0 \in \mathsf{A}(x_0)$ such that

$$c(x_0,g) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xgy}\phi(y) = \min_{a\in\mathsf{A}(x)} \left[c(x,a) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xay}\phi(y) \right] + \delta$$

for some $\delta > 0$. Let $g' \in U_S$ be the policy given by

$$g'(x) = \begin{cases} g(x) & \text{if } x \neq x_0 \\ 0 & \text{if } x = x_0. \end{cases}$$

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 that would serve as the pair (ψ, ψ) 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 contracting 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:

$$g^* := \lim_{n\to\infty} g(\alpha_n) \tag{6.8}$$

$$h(x) = \lim_{n\to\infty} \frac{C_{\alpha_n}(x) - C_{\alpha_n}(0)}{1-\alpha_n}, \qquad \forall x \in \mathbf{X} \tag{6.9}$$

$$\psi^* := \lim_{n\to\infty} C_{\alpha_n}(0), \tag{6.10}$$

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.

6.2 Non-constrained control: cost bounded below

We assume that (5.1) holds, i.e. that the costs are bounded below. Without loss of generality, we shall assume that the costs are nonnegative (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 (6.1). 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 are finite.
- $\mathcal{S}$**2:** There exists a nonnegative 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 nonnegative $\overline{m}(x)$ such that $h_\alpha(x) \le \overline{m}(x)$ for every x and α; 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 nonnegative $\overline{m}(x)$ such that $h_\alpha(x) \le \overline{m}(x)$ for every x and α; for every x and $a \in \mathsf{A}(x)$, $\sum_{y\in\mathbf{X}} \mathcal{P}_{xay}\overline{m}(y) < \infty$.

Remark 6.2 *(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 and for references to other sufficient conditions).

We present explicitly some of the above sufficient conditions for the assumption $\mathcal{S}2$ and $\mathcal{S}3$, which will be needed later in a context of approximations of MDPs. Let

$$T = \inf_{t\geq 1}\{X_t = 0\}, \qquad W_\alpha := \sum_{t=1}^{T-1} \alpha^{t-1} c(X_t, A_t). \tag{6.11}$$

Then $C_\alpha(x)$ can be written as

$$C_\alpha(x) = \min_{u\in U_M} \left[(1-\alpha)E_x^u W_\alpha + E_x^u \alpha^{T-1} C_\alpha(0)\right].$$

and

$$\begin{aligned} h_\alpha(x) &= \min_{u\in U_M} \left[E_x^u W_\alpha - \frac{1-E_x^u\alpha^{T-1}}{1-\alpha} C_\alpha(0)\right] \\ &= \min_{u\in U_M} \left[E_x^u W_\alpha - E_x^u \sum_{s=1}^{T-1} \alpha^t C_\alpha(0)\right] \end{aligned} \tag{6.12}$$

Thus for any α,

$$h_\alpha(x) \leq \min_{u\in U_M} E_x^u W_\alpha \leq \min_{u\in U_M} C_{tc}(x,u) \tag{6.13}$$

(where $C_{tc}(x,u)$ is the total expected cost till the hitting of the set $\mathcal{M} = \{0\}$). Hence, if there exists some policy u for which $C_{tc}(x,u)$ is finite for all x, then 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 (6.12) that if the growth condition (5.15) 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 6.5 *(Tauberian Theorem)*
Let $\{a_n\}$ be a sequence of nonnegative real numbers. Then

$$\underline{\lim}_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} a_t \leq \underline{\lim}_{\alpha\to 1}(1-\alpha)\sum_{t=0}^{\infty}\alpha^t a_t \leq \overline{\lim}_{\alpha\to 1}(1-\alpha)\sum_{t=0}^{\infty}\alpha^t a_t \leq \overline{\lim}_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} a_t.$$

The following is due to Sennott (1989):

Theorem 6.1 *(Existence of optimal values and stationary policies)*
Assume $\mathcal{S}1$-$\mathcal{S}3$, and consider nonnegative 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 value discount

$$C_{ea} = \lim_{\alpha\to 1} C_\alpha(x)$$

(this limit is independent on the sequence α_n in (6.8)).
(ii) Any stationary policy g^* that is obtained as limit of α-discount optimal policies $g(\alpha_n)$ (as in (6.8)) is optimal.
(iii) The pair (ψ^*, h) given in (6.9)-(6.10) satisfies the ACOI (6.1). If moreover, $\mathcal{S}3^*$ holds, then they satisfy the ACOE (6.6).

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

$$C_{\alpha_n}(0) + h_{\alpha_n}(x) = c(x, g(\alpha_n)) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xg(\alpha_n)y} h_{\alpha_n}(y). \tag{6.14}$$

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

$$\psi^* + h(x) \ge \min_{a\in \mathsf{A}(x)} \left[c(x,a) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xay} h(y) \right].$$

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 $a \in \mathsf{A}(x)$,

$$C_{\alpha_n}(0) + h_{\alpha_n}(x) \le c(x,a) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xay} h_{\alpha_n}(y). \tag{6.15}$$

Thus, we get by using $\mathcal{S}3$ and applying the dominated convergence Theorem

$$C_{\alpha_n}(0) + h(x) \le c(x,a) + \sum_{y\in\mathbf{X}} \mathcal{P}_{xay} h(y). \tag{6.16}$$

We conclude that g^* minimizes the right hand side of ACOI (6.1), so by Lemma 6.1, g^* satisfies $C_{ea}(x, g^*) \le \psi^*$. On the other hand, it follows from Lemma 6.5 that for any policy u,

$$C_{ea}(x,u) \ge \overline{\lim_{\alpha_n\to 1}}\, C_{\alpha_n}(x,u) \ge \overline{\lim_{\alpha_n\to 1}}\, C_{\alpha_n}(x) = \psi^*. \tag{6.17}$$

We thus conclude that (i) and (ii) hold. ∎

6.3 Dynamic programming approach: the contracting framework

We consider in this section the contracting framework (as defined in Chapter 5, i.e. uniform μ-geometric recurrent MDPs, μ-bounded cost, μ-continuous transition probabilities, and initial distribution satisfying $\langle \beta, \mu \rangle < \infty$). We obtain similar characterization of the optimal value and policy as in the previous section. is finite.

Theorem 6.2 *(Existence of optimal values and stationary policies)*
Consider the contracting framework. 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 discount

$$C_{ea}(\beta) = \lim_{\alpha \to 1} C_\alpha(\beta)$$

(this limit is independent on the sequence α_n in (6.8)).
(ii) Any stationary policy g^* that is obtained as limit of α-discount optimal policies $g(\alpha_n)$ (as in (6.8)) is optimal.
(iii) The pair (ψ^*, h) given in (6.9)-(6.10) satisfies the ACOE (6.1).

Proof: For each α_n, the following holds for any fixed $x \in \mathbf{X}$:

$$C_{\alpha_n}(0) + h_{\alpha_n}(x) = c(x, g(\alpha_n)) + \sum_{y \in \mathbf{X}} P_{xy}(g(\alpha_n)) h_{\alpha_n}(y) \tag{6.18}$$

(see 6.14). h_α are μ-bounded, uniformly in α (this follows from Proposition 5.1 in Spieksma, 1990, p. 97, which presents a Laurent expansion of $C_\alpha(u, \cdot)$) thus, there exists some constant $\overline{m}$ such that

$$|h_\alpha(x)| \le \overline{m} \mu(x),$$

for all α in the neighborhood of 1. Since $P(g(\alpha_n))$ has a finite μ norm (and uniformly bounded in n), we may take the liminf in both sides of (6.18) 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).

h is μ-bounded, and $\|P^n(u)\|_\mu$ is bounded, uniformly in all policies and all integers n (see Proposition 5.1.ix in Spieksma, 1990, p. 97); it then follows that for any policy u,

$$\lim_{n \to \infty} \frac{E^u_\beta h(X_n)}{n} = 0 \tag{6.19}$$

for all policies u. Lemma 6.3 (i) now implies that ψ^* is the optimal value, from which statement (i) follows, and also implies statement (ii). ■

6.4 Super-harmonic functions and linear programming

Theorem 6.3 *(The value and superharmonic functions, the contracting framework)*
Consider the contracting framework. Then
(i) The pair (C_{ea}, h) is superharmonic, and h is μ bounded.
(ii) For any other superharmonic pair (ψ, ϕ) for which $\phi : \mathbf{X} \to \mathbb{R}$ is μ-bounded, we have $C_{ea} \geq \psi$.
(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} is the largest constant for which there exist a μ-bounded function $\overline{\phi} : \mathbf{X} \to \mathbb{R}$ such that $(C_{ea}, \overline{\phi})$ is within the above class.)

Proof: (i) follows from Theorem 6.2 (i) and (iii).
(ii) follows from Lemma 6.3 (ii), since for any μ-bounded ϕ,

$$\sup_{n,u} \|\phi(X_n)\|_\mu < \infty$$

by the same arguments as in the proof of Theorem 6.2, and hence (6.5) hold. ∎

Motivated by Theorem 6.3, we introduce the the following infinite Linear Program with decision variables $\psi \in \mathbb{R}$ and $\phi(y), y \in \mathbf{X}$, used to compute the optimal expected average value of **COP**.

$$\begin{aligned} \mathbf{DP}(\beta): \quad & \text{Find } \Theta^* := \sup_{\psi,\phi} \psi \text{ s.t.} \\ \phi(x) + \psi \quad \leq \quad & c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xay} \phi(y), \qquad x \in \mathbf{X}, a \in \mathsf{A}(x). \end{aligned}$$

Theorem 6.3 implies the following:

Theorem 6.4 *(The dual linear program, the contracting framework)*
Consider the contracting framework. Consider the dual program $\mathbf{DP}(\beta)$, where ϕ is restricted to the linear space F^μ. $\mathbf{DP}(\beta)$ is feasible; its value equals $C_{ea}(\beta)$. $\phi(x) = h(x), x \in \mathbf{X}$ is an optimal solution.

We now obtain a similar statement for the case of nonnegative costs, when restricting to bounded functions (for which (6.5) clearly holds under any policy). As for the Linear program obtained for the case of total cost for transient MDPs, the fact that we restrict to a subclass of functions satisfying the conditions of Lemma 6.3 (ii) might lead to only a lower bound on the value. 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 (6.5).

Theorem 6.5 *(The dual linear program, nonnegative immediate costs)*
Assume that the immediate costs are nonnegative, and the standard moment condition (5.16) holds. Assume further that conditions $\mathcal{S}1$ and $\mathcal{S}3$ hold, and that there exists some 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. (6.5) holds for all policies, we have by Lemma 6.2

$$C^1(\beta) \leq C_{ea}(\beta). \tag{6.20}$$

Consider a set of approximating **COP** with an immediate cost $c_n(x,a) = \min\{n, c(x,a)\}$; denote by $C_\alpha^n(x,u)$ the corresponding infinite horizon expected discounted cost. Denote by $C_{tc}^n(\beta,u)$ the corresponding total expected cost till state 0 is reached. Denote

$$h_\alpha^n(x) := C_\alpha^n(x) - C_\alpha^n(0).$$

The pair $(C_\alpha^n(0), h_\alpha^n)$ is super-harmonic, since, by the same arguments as those that yield (6.14),

$$\begin{aligned} C_\alpha^n(0) + h_\alpha^n(x) &\leq c_n(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xa(\alpha_n)y} h_\alpha^n(y) \\ &\leq c(x,a) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xa(\alpha_n)y} h_\alpha^n(y). \end{aligned} \tag{6.21}$$

Consider an arbitrary sequence α_n converging to 1, along which the following limits exist:

$$C^* = \lim_{n\to\infty} C_{\alpha_n}^n(0), \qquad h^*(x) = \lim_{n\to\infty} h_{\alpha_n}^n(x), \quad \forall x, \qquad g^* = \lim_{n\to\infty} g^*(n),$$

where $g^*(n)$ is an optimal stationary policy for the α_n-discounted MDP. Since c_n are bounded by n, we have $C_\alpha^n(x) \leq n/(1-\alpha)$. Hence $h_\alpha^n(x)$ are bounded (in x) by $n/(1-\alpha)$. Thus, for any fixed $\alpha \in (0,1)$ and n, we have

$$C^* \leq C^1(\beta). \tag{6.22}$$

For any α and n we have (as follows from (6.13))

$$h_\alpha^n(x) \leq \inf_u C_{tc}^n(x,u) \leq \inf_u C_{tc}(x,u),$$

and thus in particular, $h^*(x) \leq \inf_u C_{tc}(x,u)$ is finite. For each $x \in \mathbf{X}$ and n, we have

$$C_{\alpha_n}^n(0) + h_{\alpha_n}^n(x) = c_n(x, g(n)) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xg(n)y} h_{\alpha_n}^n(y)$$

(as in (6.14). One may verify from the growth condition (5.16) and from (6.12) that $h_\alpha^n(x)$ are uniformly bounded from below by some constant that does not depend on α nor on n (which implies in particular condition $\mathcal{S}2$). Taking the limit as n tends to infinity, we get by Fatou's Lemma

$$C^* + h^*(x) \geq c(x, g^*) + \sum_{y \in \mathbf{X}} \mathcal{P}_{xg^*y} h^*(y)$$

We conclude that (C^*, h^*) satisfy ACOI. Since h^* is bounded below, it satisfies (6.2), so by Lemma 6.1, $C_{ea}(\beta) \leq C^*$. This, together with (6.20) and (6.22) establishes the proof. ■

Remark 6.3 *(The nonnegativity of the immediate cost)*
It is in (6.21) that we made use of the nonnegativity of the immediate cost. One can relax the nonnegativity 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 be nonnegative.

Remark 6.4 *(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 6.5 *(On the methodology of approximation)*
The method used to establish the convergence of the approximation scheme in the proof of Theorem 6.5 is similar in spirit to the method used by Sennott (1995) to obtain finite state approximation.

Remark 6.6 *(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.

6.5 Set of achievable costs

Define for any $\overline{U} \subset U \cup \overline{M}(U_M)$ the set of achievable vector performance measures:

$$\mathbf{M}^{ea}_{\overline{U}}(\beta) = \cup_{u\in\overline{U}}\{(C_{ea}(\beta,u), D^k_{ea}(\beta,u), k=1,...,K)\}, \tag{6.23}$$

and set $\mathbf{M}^{ea}(\beta) := \mathbf{M}^{ea}_U(\beta) \cup \mathbf{M}^{ea}_{\overline{M}(U_M)}(\beta)$. Define also

$$\mathbf{V}_{ea} := \bigcup_{\rho\in\ \mathbf{Q}_{ea}} \{(\langle c,\rho\rangle, \langle d^1,\rho\rangle, \langle d^2,\rho\rangle, ..., \langle d^K,\rho\rangle)\}. \tag{6.24}$$

The next characterization of achievable costs follows from by combining Theorems 5.4, 5.5, 5.8 and 5.9.

Theorem 6.6 *(Characterization of the sets of achievable costs)*
(i) Assume that the immediate cost is bounded below (5.1) and satisfies the growth condition (5.15) or (5.18). Then
$\mathbf{M}^{ea}(\beta)$ *is 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 3.1).
(ii) In the contracting framework, $\mathbf{M}^{ea}_{U_S}(\beta)$ *is convex and compact, and satisfies*

$$\mathbf{M}^{ea}_{\mathcal{U}}(\beta) = \mathbf{M}^{ea}(\beta) = \mathbf{M}^{ea}_{U_S}(\beta) = co\mathbf{M}^{ea}_{U_D}(\beta) = \mathbf{V}_{ea}(\beta).$$

6.6 Constrained control: Lagrange approach

By the same arguments as the ones used to establish Theorem 4.8 and 4.9, we now obtain:

Theorem 6.7 *(The Lagrangian)*
Consider either (1) the immediate cost bounded below (5.1) and satisfying the growth condition (5.15) or (5.18), or (2) the contracting framework.
(i) Let $\overline{U}$ be any class of policies containing U_s. The value function satisfies

$$C_{ea}(\beta) = \inf_{u \in \overline{U}} \sup_{\lambda \geq 0} J^{\lambda}_{ea}(\beta, u) \tag{6.25}$$

where

$$J^{\lambda}_{ea}(\beta, u) \quad := \quad C_{ea}(\beta, u) + \langle \lambda, D_{ea}(\beta, u) - V \rangle.$$

(ii) For any class $\overline{U}$ containing U_D, the value function satisfies

$$C_{ea}(\beta) = \sup_{\lambda \geq 0} \min_{u \in \overline{M}(U_S)} J^{\lambda}_{ea}(\beta, u) = \sup_{\lambda \geq 0} \min_{u \in \overline{U}} J^{\lambda}_{ea}(\beta, u). \tag{6.26}$$

and

$$C_{ea}(\beta) = \inf_{u \in \mathcal{U}} \sup_{\lambda \geq 0} J^{\lambda}_{ea}(\beta, u). \tag{6.27}$$

Proof: (i) (6.25) is standard: if $u \in \overline{U}$ is feasible (i.e. it satisfies the constraints $D_{ea}(\beta, u) \leq V$) then

$$\sup_{\lambda \geq 0} J^{\lambda}_{ea}(\beta, u) = C_{ea}(\beta, u).$$

Since there exists an optimal policy for **COP** within $\overline{U}$,

$$C_{ea}(\beta) \geq \inf_{u \in \overline{U}} \sup_{\lambda \geq 0} J^{\lambda}_{ea}(\beta, u).$$

If $u \in \overline{U}$ is not feasible then it is easily seen that

$$\sup_{\lambda \geq 0} J^{\lambda}_{ea}(\beta, u) = \infty \geq C_{tc}(\beta).$$

We conclude that

$$C_{tc}(\beta) = \inf_{u \in \overline{U}} \sup_{\lambda \geq 0} J^{\lambda}_{ea}(\beta, u).$$

The first part of (6.26) is obtained by applying Sion's minimax Theorem to the sets $G_1 := \{\lambda \geq 0\}$ and $G_2 := \overline{M}(U_S)$, as was done in the proof of part (ii) in Theorem 4.8.

For fixed λ, $J^{\lambda}(\beta, u)$ is minimized by a policy in U_D, i.e.

$$\min_{u \in \overline{U}} J^{\lambda}_{ea}(\beta, u) = \min_{u \in U_D} J^{\lambda}_{ea}(\beta, u). \tag{6.28}$$

for any class of policies $\overline{U}$ that contains U_D. This yields the second equality in (6.26).

We use again Sion's theorem for the set $G_2 = \mathcal{U}$, which yields (6.27). ■

By the same type of arguments as in the proof of part (i) of Theorem 4.8, we obtain from (6.27) the following

Corollary 6.1 *(Dominance of $\mathcal{U}$)*
Let the immediate costs be bounded below (5.1) and satisfy the growth condition (5.15) or (5.18). Then $\mathcal{U}$ is a dominating class of policies.

Corollary 6.2 *(Saddle point)*
Consider either the case of immediate cost bounded below (5.1) and satisfying the growth condition (5.15) or (5.18), or the contracting framework (with μ-bounded immediate cost). 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 summerized in the following Theorem, whose proof follows the same lines as the proof of Theorem 4.9.

Theorem 6.8 *(The Lagrangian: Slater condition)*
If there exists some policy u for which $D_{ea}(\beta,u) < V$ then there exist nonnegative Lagrange multipliers $\lambda^* = \{\lambda^*_1, ..., \lambda^*_K\}$ such that

$$C_{ea}(\beta) = \min_{u\in\mathcal{U}} J^{\lambda^*}_{ea}(\beta,u) = \min_{u\in U_D} J^{\lambda^*}_{ea}(\beta,u)$$

Moreover, any optimal policy u^* satisfies the Kuhn-Tucker conditions:

$$\lambda^*_k(D^k_{ea}(\beta,u^*) - V_k) = 0, \qquad k = 1, ..., K.$$

6.7 The dual LP

For any $u \in U_s$,

$$J^{\lambda}_{ea}(\beta,u) = \lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} E^u_\beta j^\lambda(X_s,A_s)$$

where

$$j^\lambda(x,a) := c(x,a) + \langle\lambda, d(x,a)\rangle.$$

This, together with the results of Section 6.4, 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 $\lambda \in \mathbb{R}^K_+$.

$$\begin{aligned} &\mathbf{DP_3}(\beta): &&\text{Find } \Theta^*(\beta) := \sup\nolimits_{\psi,\phi,\lambda} \psi - \langle\lambda, V\rangle \text{ s.t.} \\ &\phi(x) + \psi \quad\leq &&c(x,a) + \langle\lambda, d(x,a)\rangle + \sum_{y\in\mathbf{X}} \mathcal{P}_{xay}\phi(y), \quad x\in\mathbf{X}, a\in \mathsf{A}(x). \end{aligned}$$

Combining Theorem 6.7 with the results of Section 6.4, we get:

Theorem 6.9 *(The dual LP, contracting case)*
Consider the contracting framework. Fix an initial distribution β, such that $\langle\beta,\mu\rangle < \infty$. $\mathbf{DP_3}(\beta)$ 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 (6.9)) is an optimal solution.*

Theorem 6.10 *(The dual LP, cost bounded below)*
Assume that for each one of the immediate cost functions $c(\cdot,\cdot)$, and $d^k(\cdot,\cdot)$, $k=1,...,K$ the following hold:

- *the immediate cost are nonnegative,*
- *the standard moment condition (5.16) holds,*
- *conditions $\mathcal{S}1$ and $\mathcal{S}3$ hold for the (nonconstrained) 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 Theorem 5.10 to 6.9 and 6.10 we see that there is no duality gap between $\mathbf{LP_3}(\beta)$ and $\mathbf{DP_3}(\beta)$.

6.8 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 (5.15) or (5.18), or if the contracting framework holds, then we know by Corollary 6.1 ior by Theorem 5.7 (i) that $C_{ea}(\beta)$ is equal to the value of **COP** restricted to $\mathcal{U}$:

$$\min_{u\in\mathcal{U}} C_{ea}(\beta,u) \text{ s.t. } D_{ea}(\beta,u)\le V.$$

This can be rewritten as a Linear Program:

$$\begin{aligned} \mathbf{LP_4}(\beta): \quad & \min_{\gamma\in M_1(U_D)} \int C_{ea}(\beta,u)\gamma(du) \\ \text{s.t.} \quad & \int D^k_{ea}(\beta,u)\gamma(du)\le V^k, \quad k=1,...,K \end{aligned} \tag{6.29}$$

This yields the following:

Theorem 6.11 *(Relation between* **COP** *and* $\mathbf{LP_4}(\beta)$*)*
Consider either the case of immediate cost bounded below and satisfying the growth condition (5.15) or (5.18), or the contracting framework. Then
(i) **COP** *is feasible if and only* $\mathbf{LP_4}(\beta)$ *is feasible (i.e. the set satisfying (6.29) is nonempty). If* $\mathbf{LP_4}(\beta)$ *is feasible then there exists an optimal policy in $\mathcal{U}$ for* **COP**.
(ii) The value 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**.

CHAPTER 7

Sensitivity analysis

7.1 Introduction

We consider in this Chapter a sequence $\mathbf{COP}_n$, $n = 1, 2, ...$ of CMDPs and a "limit" CMDP, denoted by $\mathbf{COP}_\infty$, or simply by **COP**. **COP** is assumed to be feasible, and therefore, under to 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 yes, 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 following. We first introduce a general framework for approximations, that will provide sufficient conditions for having convergence in the sense of (i) and (ii) above, and will provide also the rate of convergence. It turns out that the answers for (iii) and for (iv) is in general negative, unlike the unconstrained case. The reason is that an optimal policy for $\mathbf{COP}_n$ may be unfeasible 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 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 8 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 on approximations, recall 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 there 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 a.s. 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 an CMDP whose transition probabilities are those given by the currently estimated ones. 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 finite dimensional state, e.g. Dantzig et al. (1967), Pervozvanskii and Gaitsgory (1986,1988). Several special issues of scientific journals were focused on such questions questions, as well as other related sensitivity, stability and parametric analysis: *Mathematical Programming* **21**, 1984, *Annals of Operations Research* **27**, 1990. Convergence results for constrained dynamic control problems were 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 were obtained by Altman and Gaitsgory (1993).

Our approach below to obtain convergence conditions is based on Lagrange techniques, and they are related to the techniques in Rockafelar (1989).

7.2 Key Theorems for approximation

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 $\mathbf{COP}(\overline{U})$:

$$\inf_{u \in \overline{U}} C(u) \quad s.t. \quad D(u) \leq V$$

Denote by $C^{\overline{U}}$ the value of $\mathbf{COP}(\overline{U})$. Assume that

$$|C(u)| < \overline{B} \tag{7.1}$$

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 following. 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:

$$\inf_{u \in \overline{U}} C_n(u) \quad s.t. \quad D_n(u) \leq V.$$

Denote by $C_n^{\overline{U}}$ the value of $\mathbf{COP}_n(\overline{U})$.

Remark 7.1 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 8.4. All the results we present here generalize to this case, using the same types of arguments, see Tidball and Altman (1995). However, for the simplicity of presentation we restrict to the simpler model.

We introduce the following assumptions.

- (**S**1): Slater type condition:

$$\exists v \in \overline{U} \text{ such that } D(v) < V. \tag{7.2}$$

- (**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

$$\begin{aligned} C^{\overline{U}} &= C(u^*) = \min_{u \in \overline{U}} \max_{\lambda \geq 0} \left[C(u) + \langle \lambda, D(u) - V \rangle \right] \\ &= \max_{\lambda \geq 0} \min_{u \in \overline{U}} \left[C(u) + \langle \lambda, D(u) - V \rangle \right] \\ &= \max_{\lambda \geq 0} \left[C(u^*) + \langle \lambda, D(u^*) - V \rangle \right] \\ &= \min_{u \in \overline{U}} \left[C(u) + \langle \lambda^*, D(u) - V \rangle \right]. \end{aligned}$$

  We shall use sometime 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 $\epsilon_1(n) \in \mathbb{R}$, $n = 1, 2, \ldots$ such that for all $u \in \overline{U}$,

$$\lim_{n \to \infty} \epsilon_1(n) = 0,$$

  and

$$|C_n(u) - C(u)| < \epsilon_1(n), \qquad |D_n^k(u) - D^k(u)| < \epsilon_1(n), \; k = 1, \ldots, K.$$

Remark 7.2 *(The unconstrained case)*
Our results will be applicable even for unconstrained MDPs. In that case (**S**1) and (**S**2) hold trivially.

The following theorem establishes the convergence of the value, and the rate of convergence.

Theorem 7.1 *Denote* $\eta(V) := \min_{k=1,\dots,K}[V_k - D^k(v)]$. *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* $\epsilon_1(n)$, *i.e.*

$$\overline{\lim_{n\to\infty}} \frac{|C^{\overline{U}} - C_n^{\overline{U}}|}{\epsilon_1(n)} \le \left(1 + \frac{2\overline{B}}{\eta(V)}\right). \tag{7.3}$$

In order to establish the theorem, we need the following Lemmas.

Lemma 7.1 *Assume (**S**2) and (**S**1). Then*

$$\langle \lambda^*, e \rangle \le \frac{2\overline{B}}{\eta(V)}. \tag{7.4}$$

Proof:

$$\begin{aligned} -\overline{B} &\le C^{\overline{U}} \\ &= \min_{u\in\overline{U}} [C(u) + \langle \lambda^*, D(u) - V \rangle] \\ &\le C(v) + \langle \lambda^*, D(v) - V \rangle \\ &\le \overline{B} + \langle \lambda^*, D(v) - V \rangle. \end{aligned}$$

Hence,

$$\langle \lambda^*, V - D(v) \rangle \le 2\overline{B}. \tag{7.5}$$

We then obtain (7.4) by noting that $\eta(V)\langle \lambda^*, e \rangle \le \langle \lambda^*, V - D(v) \rangle$. ■

The following Lemma shows that a property similar to (**S**2) holds also for $\mathbf{COP}_n(\overline{U})$, for n large enough.

Lemma 7.2 *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*

$$\begin{aligned} C_n^{\overline{U}} &= \inf_{u\in\overline{U}} \sup_{\lambda\ge 0} [C_n(u) + \langle \lambda, D_n(u) - V \rangle] \\ &\le \sup_{\lambda\ge 0} \inf_{u\in\overline{U}} [C_n(u) + \langle \lambda, D_n(u) - V \rangle] + 2\epsilon_1(n) k_1 \end{aligned}$$

Moreover, there exists $u_n^* \in \overline{U}$ *and* $\lambda_n^* \in R^K$ *with* $\lambda_n^* \ge 0$, *and*

$$\langle \lambda_n^*, e \rangle \le \frac{2\overline{B}}{\eta(V) - \delta_0}. \tag{7.6}$$

such that

$$\begin{aligned}&\inf_{u\in\overline{U}}\sup_{\lambda\geq 0}\left[C_n(u)+\langle\lambda, D_n(u)-V\rangle\right]\\ &\quad\leq \inf_{u\in\overline{U}}\left[C_n(u)+\langle\lambda_n^*, D_n(u)-V\rangle\right]+\epsilon_1(n)k_1\end{aligned}$$

and

$$\begin{aligned}&\sup_{\lambda\geq 0}\inf_{u\in\overline{U}}\left[C_n(u)+\langle\lambda, D_n(u)-V\rangle\right]\\ &\quad\geq \sup_{\lambda\geq 0}\left[C_n(u_n^*)+\langle\lambda, D_n(u_n^*)-V\rangle\right]-\epsilon_1(n)k_1\end{aligned}$$

Proof: We shall prove the Lemma by using for u_n^*, λ_n^* the pair
$(u^*_{V-\epsilon_1(n)e}, \lambda^*_{V-\epsilon_1(n)e})$
defined in (**S**2) (corresponding to $\mathbf{COP}(\overline{U})$ with the right hand side constraint V replaced by $V-\epsilon_1(n)e$). Consider n sufficiently large so that $\epsilon_1(n)<\delta_0$.

$$\begin{aligned}&\inf_{u\in\overline{U}}\sup_{\lambda\geq 0}\left[C_n(u)+\langle\lambda, D_n(u)-V\rangle\right]\\
\leq\ & \sup_{\lambda\geq 0}\left[C_n(u^*_{V-\epsilon_1(n)e})+\langle\lambda, D_n(u^*_{V-\epsilon_1(n)e})-V\rangle\right]\\
\leq\ & \sup_{\lambda\geq 0}\left[C(u^*_{V-\epsilon_1(n)e})+\epsilon_1(n)+\langle\lambda, D(u^*_{V-\epsilon_1(n)e})-(V-\epsilon_1(n))\rangle\right]\\
=\ & \inf_{u\in\overline{U}}\left[C(u)+\epsilon_1(n)+\langle\lambda^*_{V-\epsilon_1(n)e}, D(u)+\epsilon_1(n)-(V-\epsilon_1(n))\rangle\right]\\
\leq\ & \inf_{u\in\overline{U}}\left[C(u)+\langle\lambda^*_{V-\epsilon_1(n)e}, D(u)-(V-\epsilon_1(n))\rangle\right]+\epsilon_1(n)k_1\\
\leq\ & \inf_{u\in\overline{U}}\left[C_n(u)+\epsilon_1(n)+\langle\lambda^*_{V-\epsilon_1(n)e}, D_n(u)+\epsilon_1(n)-V\rangle\right]+\epsilon_1(n)k_1\\
\leq\ & \inf_{u\in\overline{U}}\left[C_n(u)+\langle\lambda^*_{V-\epsilon_1(n)e}, D_n(u)-V\rangle\right]+2\epsilon_1(n)k_1\\
\leq\ & \sup_{\lambda\geq 0}\inf_{u\in\overline{U}}\left[C_n(u)+\langle\lambda, D_n(u)-V\rangle\right]+2\epsilon_1(n)k_1.\end{aligned}$$

The feasibility follows from the fact that (7.7) is finite for all n large, as the first term in (7.7) equals to the value of $\mathbf{COP}(\overline{U})$with $V-\epsilon_1(n)e$ replacing V; the latter is bounded by $\overline{B}$ since (**S**1) implies that for all n large enough, $\mathbf{COP}(\overline{U})$with $V-\epsilon_1(n)e$ replacing V is feasible. The other assertions of the Lemma follow from the above equation. The upper bound of λ_n^* follows by applying Lemma 7.1 to $\mathbf{COP}(\overline{U})$with $V-\epsilon_1(n)e$ replacing V. ■

Proof of Theorem 7.1: Choose some small $\delta_0>0$ as in Lemma 7.2. It follows from Lemma 7.2 and especially the bound (7.6) that for all n large enough,

$$C_n^{\overline{U}}-C^{\overline{U}}$$

$$
\begin{aligned}
&= \sup_{\lambda\geq 0}\inf_{u\in\overline{U}}[C_n(u)+\langle\lambda, D_n(u)-V\rangle] \\
&\quad -\max_{\lambda\geq 0}\min_{u\in\overline{U}}[C(u)+\langle\lambda, D(u)-V\rangle] \\
&\leq C_n(u^*)+\langle\lambda_n^*, D_n(u^*)-V\rangle+\epsilon_1(n)k_1 \\
&\quad -[C(u^*)+\langle\lambda_n^*, D(u^*)-V\rangle] \\
&\leq +\epsilon_1(n)k_1+\epsilon_1(n)(1+\langle\lambda_n^*, e\rangle) \\
&\leq 2\epsilon_1(n)k_1
\end{aligned}
$$

We obtain similarly by the same kind of arguments, for all n large enough,

$$C - C_n \leq \epsilon_1(n)k_1$$

which concludes the proof. ∎

There are cases where one knows a-priory 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. However, 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. If we chose for both the finite and infinite horizon $\overline{U} = \overline{U}_M$, then condition (**S**3) will typically not hold. If we chose $\overline{U} = U_S$ then we will shall 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 7.2 *Assume (*$\mathbf{S}$*1)-(*$\mathbf{S}$*3) (restricted to the class of policies $\overline{U}$). Assume that for any $\epsilon > 0$ 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 7.1 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. It follows that

$$\begin{aligned} C_n - C &= \inf_{u\in\overline{U}'} \sup_{\lambda\geq 0}[C_n(u) + \langle\lambda, D_n(u) - V\rangle] \\ &\quad - \min_{u\in\overline{U})}[C(u) + \langle\lambda^*, D(u) - V\rangle] \\ &\geq [C_n(u_\epsilon) + \langle\lambda^*, D_n(u_\epsilon) - V\rangle] + \epsilon \\ &\quad -[C(u_\epsilon) + \langle\lambda^*, D(u_\epsilon) - V\rangle] \\ &\geq \epsilon_1(n)k_1 + \epsilon \end{aligned}$$

(where k_1 is defined in Lemma 7.2 and ϵ_1 is defined in (**S**3)). ■

Next, we establish the convergence of optimal policies.

Theorem 7.3 *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* $D^k(\cdot)$, $k = 1, ..., K$ *are lower semi-continuous on* $\overline{U}$.

Consider an increasing sequence of integers $m(n)$, $n = 1, 2, ...$ *and a sequence* $\epsilon_2(n)$ *decreasing to zero. Assume that* $\mathbf{COP_{m(n)}}$ *are feasible, and let* $u_n^* \in \overline{U}$ *be some* $\epsilon_2(n)$*-optimal policies for* $\mathbf{COP_{m(n)}}$, $n = 1, 2, ...$ *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

$$\begin{aligned} D^k(u^*) &\leq \lim_{n\to\infty} D^k(u_n^*) \\ &\leq \lim_{n\to\infty}[D^k_{m(n)}(u_n^*) + \epsilon_1(n)] \\ &\leq \lim_{n\to\infty}[V_k - \epsilon_1(n)] = V_k. \end{aligned}$$

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

$$\begin{aligned} C(u^*) &\leq \lim_{n\to\infty} C(u_n^*) \\ &\leq \lim_{n\to\infty}[C_{m(n)}(u_n^*) + \epsilon_1(n)] \\ &\leq \lim_{n\to\infty}[C^{\overline{U}}_{m(n)} + \epsilon_2(n) + \epsilon_1(n)] = C^{\overline{U}}. \end{aligned}$$

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 7.4 *Assume (**S**1)-(**S**3) and (**S**5).*
*(i) Let $u^1 = v$, $u^2 = u^*$. Then for any $\epsilon_4 > 0$, there exists some p such that the policy u^p defined in (**S**5) is ϵ_4-optimal for $\mathbf{COP}_n(\overline{U})$, for all n large enough.*
*(ii) Consider some sequence $\epsilon_3(n)$, $n = 1, 2, \ldots$ 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 $\epsilon_3(n)$-optimal for $\mathbf{COP}_n(\overline{U})$. Then for any $\epsilon_4 > 0$, there exists some p such that the policies $u^p(n)$ defined in (**S**5) 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.

$$\begin{aligned} D_n(u^p) &\leq D(u^p) + \epsilon_1(n)e \\ &\leq pD(v) + (1-p)D(u^*) + \epsilon_1(n)e \\ &\leq V - p[V - D(v)] + \epsilon_1(n)e. \end{aligned}$$

So, for all n for which $p[V - D(v)] + \epsilon_1(n)e \leq 0$, u^p is feasible. Similarly,

$$\begin{aligned} C_n(u^p) &\leq C(u^p) + \epsilon_1(n) \\ &\leq pC(v) + (1-p)C^{\overline{U}} + \epsilon_1(n) \\ &\leq 2p\overline{B} + C^{\overline{U}} + \epsilon_1(n)e \\ &\leq C_n^{\overline{U}} + 2p\overline{B} + [C^{\overline{U}} - C_n^{\overline{U}}] + \epsilon_1(n)e \end{aligned}$$

(i) now follows since $C^{\overline{U}} - C_n^{\overline{U}} + \epsilon_1(n)$ tends to zero (by Theorem 7.1 and by (**S**3)).
(ii) is obtained similarly. For any n,

$$\begin{aligned} D(u^p(n)) &\leq pD(v) + (1-p)D(u_n^2) \\ &\leq pD(v) + (1-p)D_n(u_n^2) + \epsilon_1(n)e \\ &\leq V - p[V - D(v)] + \epsilon_1(n)e \end{aligned}$$

and hence for any p, $u^p(n)$ are feasible for all large enough n.

$$\begin{aligned} C(u^p(n)) &\leq pC(v) + (1-p)C(u_n^2) \\ &\leq 2p\overline{B} + C_n(u_n^2) + \epsilon_1(n) \\ &\leq 2p\overline{B} + C_n^{\overline{U}} + \epsilon_3(n) + \epsilon_1(n) \end{aligned}$$

(i) now follows since $C_n^{\overline{U}} + \epsilon_3(n) + \epsilon_1(n)$ tends to $C^{\overline{U}}$ (by Theorem 7.1 and by definition of $\epsilon_1(n)$ and $\epsilon_3(n)$). ■

Remark 7.3 *(Relaxing some assumptions)*
The results of Theorem 7.4 (i) clearly extend to the setting of Theorem 7.2. 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.

7.3 Discounted cost: convergence in the discount factor

We first consider the four types of convergence 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 assume that the contracting framework holds, i.e. that the MDP is uniform μ-geometric recurrent (Definition 5.4), the immediate costs are μ-bounded (2.4), the transition probabilities are μ-continuous, and the initial distribution satisfies $\langle \beta, \mu \rangle < \infty$.

It follows then by Theorem 3.4 (iii) that one may restrict without loss of optimality to stationary policies, since they are sufficient for both the limiting and the approximating problems. Hence we may consider in the Key theorems $\overline{U}$ 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 4.2 and Theorem 4.9; Lemma 3.4 (ii) implies (**S**4). (**S**5) follows from Theorem 4.7 (ii). For any discount factor α_1 such that $\alpha_1 < \alpha/\xi$ (where ξ is defined in Definition 2.4),

$$
\begin{aligned}
&\|C_{\alpha_1}(\bullet, u) - C_\alpha(\bullet, u)\|_\mu \\
&= \left\| \sum_{j=0}^{\infty} \left[(1-\alpha_1)\alpha_1^j - (1-\alpha)\alpha^j \right] P^j(u) c(u) \right\|_\mu \\
&\le \sum_{j=0}^{\infty} \left| (1-\alpha_1)\alpha_1^j - (1-\alpha)\alpha^j \right| \left(\frac{\xi}{\alpha} \right)^j \overline{b} \\
&\le \overline{b} \sum_{j=0}^{\infty} \left| (\alpha_1^j - \alpha^j) \right| \left(\frac{\xi}{\alpha} \right)^j = \overline{b} \left| \frac{1}{1-\xi} - \frac{1}{1-\xi\alpha^{-1}\alpha_1} \right| =: \epsilon_1(\alpha_1, \alpha)
\end{aligned}
$$

($\overline{b}$ is defined in (2.4), $P^j(u)$ 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). Using Theorem 7.1, we have that the difference between $C_{\alpha_1}(\beta)$ and $C_\alpha(\beta)$ is of order $\epsilon_1(\alpha_1, \alpha)$.

7.4 Convergence of the discounted problem to the expected average problem

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. 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 consider again the contracting framework (as in Section 7.3). Finally, we make the unichain assumption (5.2) (from Chapter 5).

It follows then by Theorem 3.4 (iii) and 5.6 that one may restrict without loss of optimality to stationary policies, since they are sufficient for both the limiting and the approximating problems. Hence we may consider in the Key theorems $\overline{U}$ to be the stationary policies.

(**S**1) holds when assuming the standard Slater condition. (**S**2) is established in Corollary 6.2 and Theorem 6.8; Lemma 5.4 (ii) implies (**S**4). (**S**5) follows from Theorem 6.6 (ii). It remains to establish (**S**3). We prove it for C_{ea}; a 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 $\pi(u)$. Then

$$
\begin{aligned}
&\|C_\alpha(\bullet,u) - C_{ea}(\bullet,u)\|_\mu \\
&= \left\| \left[\sum_{j=0}^{\infty} (1-\alpha)\alpha^j P^j(u) - \Pi(u) \right] c(u) \right\|_\mu \\
&= \left\| \sum_{j=0}^{\infty} (1-\alpha)\alpha^j \left[P^j(u) - \Pi(u) \right] c(u) \right\|_\mu \\
&\le \sigma \sum_{j=0}^{\infty} (1-\alpha)\alpha^j \overline{b}\tilde{\xi}^j = \frac{\sigma \overline{b}(1-\alpha)}{1-\alpha\tilde{\xi}} =: \epsilon_1(\alpha)
\end{aligned}
$$

Using Theorem 7.1, we have that the difference between $C_\alpha(\beta)$ and $C_{ea}(\beta)$ is of order $\epsilon_1(\alpha)$.

Remark 7.4 *(The multichain case)*
It is possible to obtain similar results for the general multichain 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 multichain 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.*

7.5 Convergence in the horizon: discounted cost

We consider all four types of convergence 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 that goes to infinity, and discounted with the same discount factor α. The transition probabilities and immediate costs are the same. We assume that the contracting framework holds (Definition 2.4).

One may restrict Markov policies, since they are sufficient for both the limiting and the approximating problems (see Theorem 2.1). In order to apply below the Key theorems, we shall thus consider $\overline{U}$ to be the Markov policies.

Remark 7.5 *(Almost optimal stationary policies)*
Note that **COP** *has optimal stationary policies (Theorem 3.4 (iii)). One can then show by using Theorem 7.4 that for any $\epsilon > 0$, 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) were established in Section 7.3. (**S**3) follows from Lemma 3.4 (i). Since we specialize to discounted cost, for which we relaxed the contracting assumption (2.21) to (3.20), we repeat the calculation for C_α; a same proof holds for D_α^k. We have for any Markov policy u,

$$\begin{aligned} & \left\| C_\alpha^T(\beta, u) - C_\alpha(\beta, u) \right\|_\mu \\ \leq \quad & (1-\alpha) \left\| \sum_{t=T}^{\infty} \alpha^t P(u_1)P(u_2)...P(u_t) \right\|_\mu \overline{b} \\ \leq \quad & \frac{(1-\alpha)\alpha^{T+1}\overline{b}}{1-\xi} =: \epsilon_1(T). \end{aligned}$$

(Using Theorem 7.1, we have that the difference between $C_\alpha^T(\beta)$ and $C_\alpha(\beta)$ is of order $\epsilon_1(T)$.)

7.6 Convergence in the horizon: expected average cost

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 is not dominating for the finite horizon problem. We therefore use the approach of Theorem 7.2 and Remark 7.3.

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 infinite horizon discounted cost with discount factor α_n converging to 1. 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 consider again the contracting framework, i.e. the cost is assumed to be μ-bounded (2.4), the transition probabilities are μ-continuous (Assumption (2.22)), and the initial distribution satisfies $\langle \beta, \mu \rangle < \infty$; the MDP is assumed to be uniformly geometric ergodic (see Definition 5.4). Finally, we make the unichain assumption (5.2) (from Chapter 5).

It follows by 5.6 that one may restrict to optimal policies for the limiting case **COP**, 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 type of convergence, given in Theorem 7.1, 7.3 and 7.4 holds when restricting 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 value 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-priory 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) were established in Section 7.4. It remains to establish (**S**3). We prove it for C_{ea}; a same proof holds for D_{ea}^k. Fix an arbitrary stationary policy w, and let $\Pi(w)$ denote the matrix whose rows are all equal to $\pi(w)$.

$$
\begin{aligned}
C_{ea}(x,w) &= \sum_{y\in\mathbf{X}} \frac{1}{T}\sum_{t=1}^{T}[P^t(w)]_{xy}c(y,w), \\
C_{ea}(x,w) &= \langle \pi(w), c(w)\rangle = \sum_{y\in\mathbf{X}} T^{-1}\sum_{t=1}^{T}\pi_y(w)c(y,w)
\end{aligned}
$$

so that

$$
\begin{aligned}
\left\|C_{ea}^T(\cdot,w) - C_{ea}(\cdot,w)\right\|_\mu &\le T^{-1}\sum_{t=1}^{T}\left\|P^t(w)-\Pi(w)\right\|_\mu \|c(u)\|_\mu \\
&\le \frac{\sigma\overline{b}\sum_{t=1}^{T}\tilde{\xi}^t}{T} \le \frac{\sigma\overline{b}}{T(1-\tilde{\xi})}.
\end{aligned}
$$

This establishes (**S**3), which proves the convergence of the finite horizon CMDP restricted to stationary policies, to the infinite horizon one. Moreover, it follows from Theorem 7.1 that the rate of convergence of the values is of the order of T^{-1}.

Next, we consider the original problem of the convergence of $\mathbf{COP}_n$ to $\mathbf{COP}$, i.e. 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 7.5 *(Convergence of the finite horizon problem to the infinite horizon)*
Consider the contracting framework with $\langle\beta,\mu\rangle$ finite. Assume that the Slater condition holds, i.e. for some stationary policy, $D_{ea}(\beta,u) < V$. (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 T sufficiently large.

Proof: We shall use Theorem 7.2. We need to show that for any nonnegative λ and ϵ, there exists an ϵ-optimal stationary policy w (that may depend on λ and ϵ) for the Lagrangian

$$
J_{ea}^{\lambda,T}(\beta,u) := C_{ea}^T(\beta,u) + \langle\lambda, D_{ea}^T(\beta,u)\rangle = \frac{1}{T}\sum_{t=1}^{T} E_\beta^u j^\lambda(X_t,A_t),
$$

for all T 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 7.2. The fact that U_M is a dominating class for the finite horizon problem follows from Theorem 2.1.) Denote the value of the above minimization by $J_{ea}^{\lambda;T}(\beta)$, 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 below λ from the notation.

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 T step:

$$J^T(\beta, u, j_0) = \sum_{t=1}^{T} E_{\beta}^{u} j(X_t, A_t) + E_{\beta}^{u} j_0(X_{T+1}). \tag{7.7}$$

Denote the value of this problem by $J^T(\beta, j_0)$. We shall use the following (see e.g. Puterman, 1994):

Lemma 7.3 *(Computing the optimal value and policy for a finite horizon problem)*
(i) $J^T(\beta, j_0)$ is given by the recursive solution of

$$\begin{aligned} J^0(x, j_0) &= j_0(x), \\ J^{t+1}(x, j_0) &= \min_{a \in \mathsf{A}(x)} \left\{ j(x,a) + \sum_{y \in \mathbf{Y}} \mathcal{P}_{xay} J^t(x, j_0) \right\} \end{aligned} \tag{7.8}$$

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 $g = (g^T, g^{T-1}, ..., g^1)$, where g_i attains the minimum in (7.8) for $i = t$. Then g is optimal.

(Note that any finite horizon problem can be embedded into an infinite horizon problem by encorporating the time into the state space, see e.g. Tidball and Altman, 1995. Theorem 4.1 then implies that the recursive equations above indeed yield the optimal value).

Let (J, h) be solutions of ACOE (6.6) where c is replaced by j, such that $h \in \mathsf{F}^{\mu}$ are obtained as the limits of the discounted cost optimal values in (6.9)-(6.10) (the fact that these limits are indeed solutions of ACOE Theorem 6.1). Let g^* be a stationary optimal policy achieving the min in ACOE. Define

$$j_0(x) := h(x).$$

It follows from (7.7) that

$$J^T(x, 0) + j_0(x) \geq J^T(x, j_0). \tag{7.9}$$

We now compute $J^T(\beta, j_0)$. By (7.8), we have

$$\begin{aligned} J^0(x, j_0) &= j_0(x), \\ J^1(x, j_0) &= \min_{a \in \mathsf{A}(x)} \left\{ j(x,a) + \sum_{y \in \mathbf{Y}} \mathcal{P}_{xay} J^0(x, j_0) \right\} \end{aligned}$$

$$
\begin{aligned}
&= \min_{a \in \mathsf{A}(x)} \left\{ j(x,a) + \sum_{y \in \mathbf{Y}} \mathcal{P}_{xay} h(x) \right\} \\
&= h(x) + J_{ea},
\end{aligned}
$$

where the last equality follows from (6.6). Moreover the Markov policy $g_1 = g^*$ is optimal. We may now continue recursively and obtain

$$J^T(x, j_0) = h(x) + T J_{ea};$$

moreover, the Markov policy $g = (g^*, ..., g^*)$ is optimal, i.e. for all T,

$$J^T(x, j_0) = J^T(x, g^*, j_0) = J^T(x, g^*, 0) + j_0(x). \tag{7.10}$$

Combining (7.9) with (7.10) we get

$$J_{ea}^T(\beta) \geq J_{ea}^T(\beta, g^*) + \frac{\langle \beta, h \rangle}{T}.$$

Hence, the stationary policy g^* is ϵ optimal for the problem of minimizing $J_{ea}^T(\beta, u)$ for all T larger than $\epsilon \langle \beta, h \rangle^{-1}$. This establishes the conditions of Theorem 7.2 from which statement (i) follows. Statement (ii) follows by combining (i) with the first part of the Section. ∎

Theorem 7.2 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 7.2 $\overline{U} = \mathcal{U}$. Indeed, it is known that in this class (and in particular, within U_D) there exist ϵ-optimal stationary policies for all horizon T sufficiently large. Moreover, it can be shown that the approximation error is of the order of T^{-1}. (This follows from Federgruen, 1979). The fact that (**S**3) holds follows since there are only a finite number of elements in U_D. (**S**2) follows since, when restricting 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 8

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. 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, ot 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 number of states. The only way at this point to do that is via an LP with infinite number 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 nice structure, which enables to solve it with some simple techniques other than the ones 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 2.4 for the total cost, and Remark 5.1) for the expected average cost). (We presented a different approach and resulats for the non-contracting framework, with cost bounded below, see Remark 4.3 for the non-constrained case, and Section 4.6 for the constrained case).

The theory of state truncation (as well as other state approximations 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 that area are Whitt (1978) White (1980,1982), Hernández-Lerma (1986,1989), Cavazos-Cadena (1986), Thomas and Stengos (1985) and Sennott (1995). The case of more that one controller was investigated in Nowak (1985), Whitt (1980), Tidball and Altman (1995) and Tidball et al. (1995). Altman (1993,1994) presented state truncation techniques for the constrained MDPs, and the scheme presented in this Chapter are extensions of those. Our approach is based on the sensitivity analysis tools developed in the beginning of Chapter 7, which allows us not only to obtain convergence results but also estimation on 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 $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, ...$ converging to $\mathbf{X}$, such that $\mathcal{M} \in \mathbf{X}_1$ ($\mathcal{M}$ is defined in the beginning of Chapter 3). The nth CMDP ($\mathbf{COP}_n$) is restricted to the set $\mathbf{X}_n$ (states not in $\mathbf{X}_n$ will be considered to be transient, and not 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.

8.1 The approximating sets of states

The sets $\mathbf{X}_n$ may or may not be given a-priory. In some problems, the following "finite neighbours" property may hold: from any $x \in \mathbf{X}$, only finitely many states are reachable in one step. In other words,

$$\text{From any } x \in \mathbf{X},\ \{y : \mathcal{P}_{xay} > 0 \text{ for some } a\} \text{ is finite.} \tag{8.1}$$

(This property holds in particular in many queueing applications). When it holds, we may construct the sets $\mathbf{X}_n$ in the following way. 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 $\mathbf{X}_n$ in the following way:

$$\mathbf{X}_0 = \mathcal{X}, \qquad \mathbf{X}_{n+1} = \bigcup_{x \in \mathbf{X}_n} \mathbf{Y}(x) \bigcup \mathbf{X}_n \tag{8.2}$$

The above construction may be useful especially when the sets of neighbours of a "typical" state is not too large. When it is large, then the sets $\mathbf{X}_n$ become large 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 (8.1) does not hold). We define a parametrized family $\{\mathbf{X}_n(\epsilon)\}$, where ϵ is a positive real number. Define $\mathbf{X}_0(\epsilon) = \mathcal{X}$, where, again, $\mathcal{X}$ is a given set (in which we would like to approximate the values and policies). $\{\mathbf{X}_n(\epsilon)\}$ are then chosen to be an arbitrary sequence increasing to $\mathbf{X}$ that satisfies the following. If for some $l > 0$, say $l = \hat{l}$,

$$\sup_{x \in \mathbf{X}_l(\epsilon)} \sup_{a \in \mathsf{A}(x)} \sum_{y \notin \mathbf{X}_l(\epsilon)} \mathcal{P}_{xay} \le \epsilon \tag{8.3}$$

then $\mathbf{X}_n(\epsilon) = \mathbf{X}$ for all $n > \hat{l}$. Otherwise, $\mathbf{X}_{l+1}(\epsilon)$ is chosen such that

$$\sup_{x \in \mathbf{X}_l(\epsilon)} \sup_{a \in \mathsf{A}(x)} \sum_{y \notin \mathbf{X}_{l+1}(\epsilon)} \mathcal{P}_{xay} \le \epsilon. \tag{8.4}$$

In other words, we replace neighbouring sets in the previous scheme (8.2) by some "ϵ-neighbouring sets"; in (8.2), 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 (8.3) and (8.4), it is less than ϵ. One could also consider weighted versions of (8.3) and (8.4), 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 $g_n = g_n(\epsilon)$, such that the nth step of the approximation will yield an error of the order of ϵ. To that end, we begin by defining

$$\delta(r,n) := \sup_{\substack{x\in\mathbf{X}_r\\ a\in\mathsf{A}(x)}} \sum_{y\notin\mathbf{X}_n} \mathcal{P}_{xay}\mu(y).$$

Due to the contracting assumption (2.21), the following holds

$$\lim_{n\to\infty}\delta(r,n)=0, \qquad \forall r \tag{8.5}$$

if $\mathbf{X}_n$ are finite sets for all n. This follows from the following argument. Assume that (8.5) does not hold. Then, there exists some $b>0$ such that for some x,

$$\overline{\lim_{n\to\infty}}\max_a\left(\sum_{y\in\mathbf{X}}\mathcal{P}_{xay}\mu(y)1\{y\notin\mathbf{X}_n\}\right)=b. \tag{8.6}$$

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 (2.22)). Choose a subsequence $n(\ell),\ell=1,2,...$ along which the limsup is obtained and along which a_n converges to some action a^*. Then $\mathcal{P}_{xa_{n(\ell)},\bullet}$ converges (pointwize) to the probability $\mathcal{P}_{xa^*,\bullet}$ as $\ell\to\infty$. But then it follows from a dominant convergence Theorem (Royden 1988, Ch. 11 Sec. 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 (8.6). Hence (8.5) indeed holds.

We use an idea introduced by Cavazos-Cadena (1986) and further developed by Tidball and Altman (1995). 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,

$$g_k=g(\epsilon,g_{k-1}), \quad g(\epsilon,r)=\min\{m:\delta(r,m)\le\epsilon\}, \tag{8.7}$$

where δ is defined above (8.5). Due to assumption (8.5) this sequence is well defined, and for all k, g_k is finite. Finally, we define

$$m^k(\epsilon)=\max\{g_m,m=0,1...,k\}.$$

In the approximation schemes that we introduce in the next sections that involve the use of the truncated state space $\mathbf{X}_n$ instead of the original one, we show that for $n\ge m^k(\epsilon)$, the approximation error of the value for states in $\mathcal{X}$ will be of order $\max(\epsilon,\xi^n)$ (ξ is the conraction factor defined in Definition 2.4).

In the special case that $\mathbf{X}_n$ are given by the μ-weighted versions of (8.3) and (8.4), we get $\epsilon(\mathcal{X})=0$, $g^0=0$, and hence $g^k=k$ and $m_k(\epsilon,\epsilon(\mathcal{X}))=g^k=k$ for $k\le\hat{l}$.

We shall use in the next Sections the following version of the μ-norm, adapted to the truncated state space. For any $\mathcal{X}\subset\mathbf{X}$ and any functions $q:\mathbf{X}\to\mathbb{R}$, we define

$$\|q\|_\mu^{\mathcal{X}}=\sup_{x\in\mathcal{X}}\frac{q(x)}{\mu(x)}.$$

Our aim in the approximation schemes below are 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 $\mathcal{X}$.

8.2 Scheme I: the total cost

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 some state $0 \in \mathcal{M}$. Hence, $\mathcal{P}^n_{xay}$ is defined by:

$$\mathcal{P}^n_{xay} = \begin{cases} \mathcal{P}_{xa0} + \sum_{z \notin \mathbf{X}_n} \mathcal{P}_{xaz} & y = 0 \\ \mathcal{P}_{xay} & y \neq 0, y \in \mathbf{X}_n \\ 0 & y \notin \mathbf{X}_n \end{cases} \tag{8.8}$$

Both $\mathbf{COP}$ and $\mathbf{COP}_n$ have optimal stationary policies according to Theorem 3.4. We can therefore consider $\mathbf{COP}$ and $\mathbf{COP}_n$ restricted to U_S. When applying the results and checking the assumptions of Section 7.2, we shall use $\overline{U} = U_S$ to conclude that the optimal values and policies converge.

Let $C^n_{tc}(\beta, w)$, $D^{k,n}_{tc}(\beta, w)$, $k = 1, ..., K$, the costs under a policy w corresponding to the nth approximation (i.e. to the transition probabilities $\mathcal{P}^n$). Let $C^n_{tc}(\beta)$ denote the corresponding optimal value. For any policy u, the value of $C^n_{tc}(x, u)$ and $D^{k,n}_{tc}(x, u)$ outside of $\mathbf{X}_n$ are taken to be zero. (All values are zero also in $\mathcal{M}$.)

Fix an arbitrary stationary policy w. From Remark 4.1 it follows that $C^n_{tc}(\cdot, w)$ and $C^n_{tc}(\cdot, w)$ are the unique solutions in F^μ of the fixed point equations

$$\phi(x, w) = c(x, w) + \sum_{y \in \mathbf{X}'} \mathcal{P}_{xwy} \phi(y, w) \tag{8.9}$$

$$\phi^n(x, w) = \begin{cases} c(x, w) + \sum_{y \in \mathbf{X}'} \mathcal{P}^n_{xwy} \phi^n(y, w), & \text{for } x \in \mathbf{X}_n \\ 0 & \text{for } x \notin \mathbf{X}_n \end{cases}$$

Theorem 8.1 *(Convergence of values and policies)*
Consider the contracting framework. Assume that there exists some policy v satisfying the Slater condition

$$D_{tc}(\beta, v) < V. \tag{8.10}$$

Under Scheme I,
(i) The values $C^n_{tc}(\beta)$ of the truncated MDP converge to the value $C_{tc}(\beta)$ of the original one;
(ii) For any $\epsilon > 0$, there exists a stationary policy w (characterized in Theorem 7.4 (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 $\mathbf{COP}$.

Proof: The Proof is obtained by applying Theorems 7.1, 7.3 and 7.4. We show that the assumptions there indeed hold.

(**S**1) holds by assumption (8.10); (**S**2) is established in Corollary 4.2 and Theorem 4.9; Lemma 3.4 (ii) implies (**S**4). (**S**5) follows from Theorem 4.7 (ii). It remains to establish (**S**3). We prove it for C_{tc}; a 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 neighbour assumptions (8.1) holds, and when $\mathbf{X}_n$ are defined in (8.2). In that case, for $x \in \mathcal{X}$, we have

$$\begin{aligned} & \frac{1}{\mu(x)} \left| C_{tc}^n(x,w) - C_{tc}(x,w) \right| \\ & \leq \sum_{y \in \mathbf{Y}(x)} \mathcal{P}_{xwy} \left| C_{tc}^n(y,w) - C_{tc}(y,w) \right| \\ & \leq \xi \left\| C_{tc}^n(\cdot,w) - C_{tc}(\cdot,w) \right\|_\mu^{\mathbf{X}_1}. \end{aligned}$$

Continuing that way, we obtain for

$$\|C_{tc}^n(\cdot,w) - C_{tc}(\cdot,w)\|_\mu^{\mathcal{X}} \leq \xi^n \|C_{tc}^n(\cdot,w) - C_{tc}(\cdot,w)\|_\mu^{\mathbf{X}_n},$$

and since

$$\|C_{tc}^n(\cdot,w) - C_{tc}(\cdot,w)\|_\mu^{\mathbf{X}_n} \leq \frac{2\overline{b}}{1-\xi},$$

we finally get

$$\|C_{tc}^n(\cdot,w) - C_{tc}(\cdot,w)\|_\mu^{\mathcal{X}} \leq \frac{2\overline{b}\xi^n}{1-\xi}.$$

This establishes (**S**3) under the conditions (8.1) and when $\mathbf{X}_n$ are defined in (8.2).

Next we consider the general case. Let $n \geq m^\nu(\epsilon)$, where ν is some given integer. Clearly,

$$\|C_{tc}^n(\cdot,w) - C_{tc}(\cdot,w)\|_\mu^{\mathcal{X}} \leq \|C_{tc}^n(\cdot,w) - C_{tc}(\cdot,w)\|_\mu^{\mathbf{X}_{g_0}}$$

since $\mathcal{X} \subset \mathbf{X}_{g_0}$. For $x \in \mathbf{X}_{g_0}$,

$$\begin{aligned} & \frac{1}{\mu(x)} \left| C_{tc}^n(x,w) - C_{tc}(x,w) \right| \\ = \quad & \frac{1}{\mu(x)} \left| \sum_{y \notin \mathcal{M}} \mathcal{P}_{xwy}^n C_{tc}^n(y,w) - \mathcal{P}_{xwy} C_{tc}(y,w) \right| \\ \leq \quad & \frac{1}{\mu(x)} \sum_{y \notin \mathbf{X}_{g_1} \setminus \mathcal{M}} \mathcal{P}_{xwy}^n \left| C_{tc}^n(y,w) - \mathcal{P}_{xwy} C_{tc}(y,w) \right| \\ & + \frac{1}{\mu(x)} \sum_{y \notin \mathbf{X}_{g_1}} \left| \mathcal{P}_{xwy}^n C_{tc}^n(y,w) \right| + \left| \mathcal{P}_{xwy} C_{tc}(y,w) \right| \\ \leq \quad & \sum_{y \in \mathbf{X}_{g_1} \setminus \mathcal{M}} \frac{\mathcal{P}_{xwy}\mu(y)}{\mu(x)} \frac{\left| \mathcal{P}_{xwy}^n C_{tc}^n(y,w) - \mathcal{P}_{xwy} C_{tc}(y,w) \right|}{\mu(y)} \end{aligned}$$

$$+\frac{1}{\mu(x)}\sum_{y\notin\mathbf{X}_{g_1}}\mathcal{P}_{xwy}\mu(y)\left(\left|\frac{C^n_{tc}(y,w)}{\mu(y)}\right|+\left|\frac{C_{tc}(y,w)}{\mu(y)}\right|\right)$$

In the last inequality, the first term is bounded by $\xi \parallel C^n_{tc}(\cdot,w)-C_{tc}(\cdot,w)\parallel_\mu^{\mathbf{X}_{g_1}}$ (because of assumption (2.21)) and the second by $2\overline{b}\epsilon/(1-\xi)$ (due to the definition of the sequence $\mathbf{X}_{g_k}$ and by Theorem 3.3 (i)). We obtain

$$\parallel C^n_{tc}(\cdot,w)-C_{tc}(\cdot,w)\parallel_\mu^{\mathbf{X}_0} \;\le\; \xi \parallel C^n_{tc}(\cdot,w)-C_{tc}(\cdot,w)\parallel_\mu^{\mathbf{X}_{g_1}} +2\frac{\overline{b}\epsilon}{1-\xi}.$$

In the same way we get for $k\le m^\nu(\epsilon)\le n$

$$\parallel C^n_{tc}(\cdot,w)-C_{tc}(\cdot,w)\parallel_\mu^{\mathbf{X}_{g_k}} \;\le\; \xi \parallel C^n_{tc}(\cdot,w)-C_{tc}(\cdot,w)\parallel_\mu^{\mathbf{X}_{g_{k+1}}} +2\frac{\overline{b}\epsilon}{1-\xi}.$$

Since

$$\parallel C^n_{tc}(\cdot,w)-C_{tc}(\cdot,w)\parallel_\mu^{\mathbf{X}_{g_k}} \;\le\; \frac{2\overline{b}}{1-\xi},$$

we get for any integer ν with $n\ge m^\nu(\epsilon)$,

$$\parallel C^n_{tc}(\cdot,w)-C_{tc}(\cdot,w)\parallel_\mu^{\mathcal{X}} \;\le\; \xi^\nu\frac{2\overline{b}}{1-\xi}+\frac{2\overline{b}\epsilon}{1-\xi}\left(\frac{1-\xi^\nu}{1-\xi}\right). \tag{8.11}$$

Since ν 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 $D^k_{tc}(x,u)$, $k=1,...,K$, we obtain finally establish condition (**S**3). ∎

Remark 8.1 *Other results from Chapter 7 can be used to further characterize the convergence of values and policies. In particular, one may use Theorem 7.1 (ii) to further characterize the rate of convergence of $C^n_{tc}(\beta)$ to $C_{tc}(\beta)$. Moreover, a construction of almost optimal policies for* **COP** *based on policies that are optimal for* $\mathbf{COP}_n$ *can be carried as in Theorem 7.4 (ii).*

8.3 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, $x \in \mathbf{X}$, $a \in \mathsf{A}(x)$,

$$q^n_{xay} \geq 0 \text{ for } y \in \mathbf{X}_n, \qquad q^n_{xay} = 0 \text{ for } y \notin \mathbf{X}_n, \qquad \sum_{y \in \mathbf{X}_n} (\mathcal{P}_{xay} + q^n_{xay}) = 1.$$

The transitions for the finite problems are then given by

$$\mathcal{P}^n_{xay} = \begin{cases} \mathcal{P}_{xay} + q^n_{xay} & x, y \in \mathbf{X}_n \\ 0 & \text{otherwize} \end{cases} \tag{8.12}$$

It follows that

$$\sum_{y \in \mathbf{X}_n} q^n_{xay} = \sum_{y \notin \mathbf{X}_n} \mathcal{P}_{xay}. \tag{8.13}$$

We make the following assumption on μ and on $\mathbf{X}_n$.

$$\text{For any } n > m \text{ and } x \in \mathbf{X}_n \setminus \mathbf{X}_m, \mu(x) \geq \sup_{y \in \mathbf{X}_m} \mu(x) =: \overline{\mu}_m.$$

As in the first approximation scheme, we have $C^n_{tc}(x, u) = D^{k,n}_{tc}(x, u) = 0, k = 1, ..., K$ for all $x \notin \mathbf{X}_n$.

Theorem 8.2 *(Convergence of the values and policies)*
Consider the contracting framework. Assume that there exists some policy v satisfying the Slater condition (8.10). Then under Scheme II, all the statements of Theorem 8.1 hold.

Proof: The Proof is obtained by applying again Theorems 7.1, 7.3 and 7.4. 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 8.1. For any stationary policy w, $C^n_{tc}(\cdot, w)$ and $C^n_{tc}(\cdot, w)$ are again the unique solutions in F^μ of the fixed point equations (8.9) (with the new transition probabilities $\mathcal{P}^n$), and are zero outside of $\mathbf{X}_n$.

We begin by obtaining a bound for $C^n_{tc}(x, w)$, uniformly over n and w.

$$\begin{aligned} C^n_{tc}(x, w) &= c(x, w) + \sum_{y \in \mathcal{X}_n \setminus \mathcal{M}} \mathcal{P}^n_{xwy} C^n_{tc}(y) \\ &= c(x, w) + \sum_{y \in \mathcal{X}_n \setminus \mathcal{M}} \mathcal{P}_{xwy} C^n_{tc}(y) + \sum_{y \in \mathcal{X}_n \setminus \mathcal{M}} q^n_{xwy} C^n_{tc}(y) \\ &\leq c(x, w) + \sum_{y \in \mathcal{X}_n \setminus \mathcal{M}} \mathcal{P}_{xwy} C^n_{tc}(y) + \sum_{y \in \mathcal{X}_n \setminus \mathcal{M}} q^n_{xwy} \overline{\mu}_n \sup_{y' \in \mathbf{X}_n} \frac{C^n_{tc}(y')}{\mu(y')} \\ &\leq c(x, w) + \sum_{y \in \mathcal{X}_n \setminus \mathcal{M}} \mathcal{P}_{xwy} \mu(y) \frac{C^n_{tc}(y)}{\mu(y)} + \sum_{y \notin \mathcal{X}_n} \mathcal{P}_{xay} \mu(y) \sup_{y' \in \mathbf{X}_n} \frac{C^n_{tc}(y')}{\mu(y')} \end{aligned}$$

We thus conclude that

$$\|C^n_{tc}(\cdot,w)\|_\mu \le \overline{b} + \xi \, \|C^n_{tc}(\cdot,w)\|_\mu$$

so that

$$\|C^n_{tc}(\cdot,w)\|_\mu \le \frac{\overline{b}}{1-\xi}.$$

Let $n \ge m^\nu(\epsilon)$, where ν is some given integer (hence, in particular, $n \ge g_1$). For $x \in \mathbf{X}_{g_0}$ (and thus, in particular, for $x \in \mathcal{X}$),

$$\begin{aligned}
&\frac{1}{\mu(x)} \left| C^n_{tc}(x,w) - C_{tc}(x,w) \right| \\
&= \quad \frac{1}{\mu(x)} \left| \sum_{y \notin \mathcal{M}} \mathcal{P}^n_{xwy} C^n_{tc}(y,w) - \mathcal{P}_{xwy} C_{tc}(y,w) \right| \\
&\le \quad \frac{1}{\mu(x)} \sum_{y \in \mathbf{X}_{g_1} \setminus \mathcal{M}} \mathcal{P}_{xwy} |C^n_{tc}(y,w) - C_{tc}(y,w)| \\
&\qquad + \frac{1}{\mu(x)} \sum_{y \in \mathbf{X}_n \setminus \mathcal{M}} q^n_{xwy} \left| C^n_{tc}(y,w) \right| \\
&\qquad + \frac{1}{\mu(x)} \sum_{y \notin \mathbf{X}_{g_1}} \mathcal{P}_{xwy} |C^n_{tc}(y,w)| + \mathcal{P}_{xwy} |C_{tc}(y,w)| \\
&\le \quad \sum_{y \in \mathbf{X}_{g_1} \setminus \mathcal{M}} \frac{\mathcal{P}_{xwy}\mu(y)}{\mu(x)} \frac{\left| \mathcal{P}^n_{xwy} C^n_{tc}(y,w) - \mathcal{P}_{xwy} C_{tc}(y,w) \right|}{\mu(y)} \\
&\qquad + \frac{1}{\mu(x)} \sum_{y \in \mathbf{X}_n \setminus \mathcal{M}} q^n_{xwy} \mu(y) \sup_{y' \in \mathbf{X}_n} \frac{|C^n_{tc}(y',w)|}{\mu(y')} \\
&\qquad + \frac{1}{\mu(x)} \sum_{y \notin \mathbf{X}_{g_1}} \mathcal{P}_{xwy} \mu(y) \left(\left| \frac{C^n_{tc}(y,w)}{\mu(y)} \right| + \left| \frac{C_{tc}(y,w)}{\mu(y)} \right| \right) \\
&\le \quad \xi \parallel C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \parallel_\mu^{\mathbf{X}_{g_1}} \\
&\qquad + \frac{1}{\mu(x)} \sum_{y \notin \mathbf{X}_{g_1}} \mathcal{P}_{xwy} \mu(y) \sup_{y' \in \mathbf{X}_n} \frac{|C^n_{tc}(y',w)|}{\mu(y')} + \frac{2\overline{b}\epsilon}{1-\xi} \\
&\le \quad \xi \parallel C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \parallel_\mu^{\mathbf{X}_{g_1}} + \frac{3\overline{b}\epsilon}{1-\xi}
\end{aligned}$$

Thus,

$$\parallel C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \parallel_\mu^{\mathbf{X}_0} \;\le\; \xi \; \parallel C^n_{tc}(\cdot,w) - C_{tc}(\cdot,w) \parallel_\mu^{\mathbf{X}_{g_1}} + 3\frac{\overline{b}\epsilon}{1-\xi}.$$

As in (8.11), we get for any integer ν with $n \geq m^\nu(\epsilon)$,

$$\| C_{tc}^n(\cdot, w) - C_{tc}(\cdot, w) \|_\mu^{\mathcal{X}} \leq \xi^\nu \frac{2\bar{b}}{1-\xi} + \frac{3\bar{b}\epsilon}{1-\xi}\left(\frac{1-\xi^\nu}{1-\xi}\right). \tag{8.14}$$

This establishes (**S**3). ■

Again, one may use Theorem 7.1 (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 $\mathbf{COP}_n$ can be carried as in Theorem 7.4 (ii).

8.4 Scheme III: The total cost

The basic idea of the approximation scheme is to fix some stationary policies for both players, and use them 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 first fix some arbitrary policies $u \in U_S$. We note that in this approach, the set of policies depends on n (see Remark 7.1):

$$U_n = \{w \in U_S : w_x = u_x, \ \forall x \notin \mathbf{X}_n\}.$$

To avoid this problem, we introduce the projection $\pi : U_S \to U_n$:

$$\pi_x^n(w) = \begin{cases} w(x) & \text{if } x \in \mathbf{X}_n, \\ & \\ u(i) & \text{if } x \notin \mathbf{X}_n; \end{cases}$$

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 8.1 hold also for Scheme III.

8.5 The expected average cost

All the results of previous Sections hold also for the expected average cost. This is summerized in the following:

Theorem 8.3 *(Convergence of values and policies)*
Consider the contracting framework. Assume that there exists some policy v satisfying the Slater condition

$$D_{ea}(\beta, v) < V. \tag{8.15}$$

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 $\epsilon > 0$, there exists a stationary policy w (characterized in Theorem 7.4 (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* $\mathbf{COP}$.

Proof: The Proof is obtained by applying Theorems 7.1, 7.3 and 7.4. We show that the assumptions there indeed hold. (**S**1) holds by assumption (8.10); (**S**2) is established in Corollary 6.2 and Theorem 6.8; Lemma 5.4 (ii) implies (SENS4). (**S**5) follows from Theorem 6.6 (ii). It remains to establish (**S**3). We prove it for C_{ea}; a same proof holds for D^k_{ea}.

Since we are in the contracting framework (see Remark 5.1) some finite set $\mathcal{M}$ for which

$$\sum_{y \notin \mathcal{M}} [P^{n_0}(u)]_{xy} \mu(y) \leq \xi \mu(x). \tag{8.16}$$

(for some integer n_0). It follows (see Spieksma 1990) that one may choose some state, say 0, with $0 \in \mathcal{M}$, some μ' and ξ' such that (8.16) 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

$$T := \inf_{t>0}\{X_t = 0\}, \qquad \mathcal{T}(w) := E^w_0 T.$$

For any stationary policy, say w, we have

$$C_{ea}(x,w) = \frac{C_{tc}(0,w)}{\mathcal{T}(w)}$$

(see Chung 1967, pp. 91-92), where by $C_{tc}(0,w)$ we mean the standard total costs till we hit the set $\mathcal{M} = \{0\}$. Similarly, we have

$$C^n_{ea}(x,w) = \frac{C^n_{tc}(0,w)}{\mathcal{T}^n(w)}, \tag{8.17}$$

where both C^n_{tc} and $\mathcal{T}^n(w)$ 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 $\mathcal{T}^n(w)$ converges to $\mathcal{T}(w)$ uniformly in U_S (this is obtained by identifying $\mathcal{T}^n(w)$ as the total expected cost till hitting $\mathcal{M} = \{0\}$, for the immediate cost of $c'(x,a) = 1\{x \neq 0\}$). This, together with (8.17), implies that $C^n_{ea}(x,w)$ converges to $C_{ea}(x,w)$ uniformly in $w \in U_S$, which establishes (**S**3). ■

CHAPTER 9

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", to appear in *ZOR - Methods and Models in Operations Research*, **43**, issue 1.

E. Altman (1995a), "Constrained Markov decision processes with total cost criteria: Lagrange approach and dual LP", submitted.

E. Altman (1995b), "Constrained Markov decision processes with expected average cost criteria: Lagrange approach and dual LP", submitted.

E. Altman and V. A. Gaitsgory (1993), "Stability and Singular Perturbations in Constrained Markov Decision Problems", *IEEE Trans. Auto. Control*, **38**, pp. 971-975.

E. Altman and V. A. Gaitsgory (1995), "A hybrid (differential-stochastic) zero-sum game with fast stochastic part", *Advances of dynamic games and applications*.

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 *Int. J. of Game Theory*.

E. Altman and F. Spieksma (1995), "The Linear Program approach in Markov Decision Problems revisited", to appear in *ZOR - Methods and Models in Operations Research*, **42**, Issue 2.

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-Dimentional 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.

R. J. Aumann (1964), "Mixed and behavior strategies in infinite extensive games", Advances in Game Theory, *Ann. Math. Study*, **52**, pp. 627-650.

M. Bayal-Gursoy and K. W. Ross (1992), "Variability sensitive Markov decision processes", *Math. of Operations Research*, **17**, pp. 558-571.

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.

P. Billingsley (1968), *Convergence of Probability Measures*, J. Wiley, New-York.

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.

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 (1994), "Ergodic control of Markov Chains with constraints – the general case", *SIAM J. Control and Optimization*, **32**, pp. 176-186.

A. D. Bovopoulos and A. A. Lazar (1991), "The effect of delayed feedback information on network performance", *Annals of Operations Res.*

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 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.

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**, 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. 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.

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.

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 (1995), "Constrained semi-Markov decision processes with average rewards", *ZOR - Methods and Models in Operations Research*, **39**, pp. 257-288.

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 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", submitted to *ZOR - Methods and Models in Operations Research*.

E. A. Feinberg E. and A. Shwartz (1995a), "Constrained discounted dynamic programming", to appear in *Math. of Operations Research*.

J. A. Filar and L. C. M Kallenberg (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. Statit.*, **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*, 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", submitted to *OR letters.*

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 problems 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.

O. Hernández-Lerma and D. Hernández-Hernández (1994), *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.

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.*

A. Hordijk and F. Spieksma (1989), "Constrained admission control to a queuing system" *Advances of Applied Probability* **21**, pp. 409-431.

M. T. Hsiao and A. A. Lazar (1991), "Optimal decentralized flow control of Markovian queueing networks with multiple controllers", *Performance evaluation*, **13**, 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, "On randomized policies and mixtures of deterministic policies", manuscript.

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.

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.

P. Kolesar (1970), "A Markovian model for hospital admission and scheduling", *Management Science* **16**, pp. 384-396, 1970.

Y. A. Korilis and A. Lazar (1995a), "On the Existence of Equilibria in Noncooperative Optimal Flow Control", To appear *J. of the ACM*. Available from ftp.ctr.columbia.edu as CTR-Research/comet/public/papers/93/KOR93.ps.gz

Y. A. Korilis and A. Lazar (1995b), "Why is flow control hard: optimality, fairness, partial and delayed information", prepreint.

D. Krass (1989), *Contributions to the Theory and Applications of Markov Decision Processes*, Ph.D. thesis, Department of Mathematical Sciences, Johns Hopkins Univ., Baltimore, USA.

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 (1995), "Average optimal stationary policies and Linear programming in countable state Markov decision processes", to appear in *J. Math. Anal. Appl.*.

A. Lazar (1983), "Optimal flow control of a class of queuing networks in equilibrium", *IEEE Transactions on Automatic Control*, **28**, pp. 1001-1007.

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.

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 Trans. Inform. Theory, IT-28*, pp. 263-273.

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.

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.

Pervozvanskii A. A. and V. A. Gaitsgory (1988), *Theory of suboptimal Decision: Decomposition and Aggregation*, Kluwer Academic Publisher, Dordrecht.

A. B. Piunovskiy (1994), "Control of jump processes with constraints", *Avtomatika i telemekhanika*, **4** pp. 75-89.

M. Puterman (1994), *Markov decision processes*, John Wiley & Sons, New York.

T. E. S. Raghavan and J. A. Filar (1991), "Algorithms for Stochastic Games - A survey", *ZOR - Methods and Models in Operations Research*, **35**, pp. 437-472.

R. T. Rockafelar (1989), *Conjugate Duality and Optimization*, Society for Industrial and Applied Mathematics, 2nd printing, Philadelphia.

K. W. Ross, "Randomized and past-dependent policies for Markov decision processes with multiple constraints", *Operations Research* **37**, pp. 474-477, 1989.

K. W. Ross and B. Chen (1988), "Optimal scheduling of interactive and non interactive traffic in telecommunication systems", *IEEE Trans. on Auto. 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.

L. S. Shapley (1953), "Stochastic Games", *proccedings Nat. Acad. of Science USA*, **39**, pp. 1095-1100.

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 (1995), "Computation of average optimal policies in denumerable state Markov Decision Chains", preprint.

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.

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.

M. Sion (1958), "On general minimax theorems", *Pacific J. Math* **8**, pp. 171-176.

M. J. Sobel (1985), "Maximal mean/standard deviation ratio in undiscounted MDP", *OR Letters*, **4**, pp. 157-159.

M. J. Sobel (1994), "Mean-vatiance 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.

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 (1995), "Approximations in dynamic zero-sum games, I", INRIA report No. 2166, to appear in *SIAM J. Control and Optimization*.

M. Tidball and A. Altman (1996), "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 (1995), "Approximations in dynamic zero-sum games, II", INRIA report No. RR-2348, Submitted to *SIAM J. Control and Optimization*.

F. Vakil and A. A. Lazar (1987), "Flow control protocols for integrated networks with partially observed voice traffic", *IEEE Trans. on Automatic Control*, **AC-32**, pp. 2-14.

J. Van Der Wal, *Stochastic Dynamic Programming*, Mathematisch Centrum, Amsterdam, 1990.

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 Spektrum*, **9**, pp. 13-22.

D. J. White (1994), "A mathematical programming approach to a problem in variance penalised Markov decision processes", *OR Spectrum* **15**, pp. 225-230.

W. Whitt (1978), "Approximations of Dynamic Programs, I", *Mathematics of Operations Research*, Vol. 3 No. 3, pp. 231-243, 1978.

W. Whitt (1980), "Representation and Approximation of Noncooperative Sequential Games", *SIAM J. Control and Opt.*, Vol 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.

CHAPTER 10

List of Symbols and Notation

$\langle q_1, q_2 \rangle$:= scalar product between two vectors.
$q_1 \le q_2$:= componentwize ordering between two vectors.

A^c: the complement of a set A.

a, A_t, A: actions, actions at time t, actions space, Section 2.1.

B, $\mathcal{B}$ a borel set, set of Borel subsets, Section 2.1.

$\underline{b}$: bound on the costs, eq. (5.1). b: constants, used for bounds, eq. (8.6).

$\overline{B}$ upper bound on $C(x,u)$, eq. (7.1).

c- immediate cost, Section 2.1.

$C^T(\beta,u)$, $C_\alpha^T(\beta,u)$, $C_{tc}(\beta,u)$, $C_\alpha(\beta,u)$, $C_{ea}(\beta,u)$, $C_{av}(\beta,u)$:= finite horizon expected cost, finite horizon discounted expected cost, total expected cost, total discounted expected cost, expected average cost, average cost.

$\mathcal{C}(\rho) := c \cdot \rho$ - linear expressions in the primal LP.

d^k- immediate costs, Section 2.1.

$D^{k,T}(\beta,u)$, $D_\alpha^{k,T}(\beta,u)$, $D(\beta,u)$, $D_\alpha^k(\beta,u)$, $D_{ea}^k(\beta,u)$, $D_{av}^k(\beta,u)$:= the other costs.

$\mathcal{D}^k(\rho) := d^k \cdot \rho$:= the linear expressions in the primal LP.

$\mathbf{DP}_i$: Linear Programs related to **COP**.

E^u_β - expectation related to initial distribution β and policy u.

f, $\bar{f}$ - occupation measures, expected occupation measure.
F_{ea} limit set of occupation measures for the expected average cost.
F^μ:= the set of μ-bounded functions (defined in Section 2.5).

g - staionary deterministic policy, Section 2.1.

$G, \mathcal{G}$:= denotes a set together with its σ-algebra, Section 2.1.

h_α:= some difference of optimal discounted costs, Section6.2.
$\underline{h}$:= lower bound on h_α (Section 6.2).
h_t, H_t, $\mathbf{H}_t$: history till t, the space of histories, Section 2.1.

J^λ:= the Lagrangian. Section 4.

k, K: indices (of constraints), Number of constraints.
$\mathcal{K}, \kappa$:= state action pairs, generic element. Section 2.1.

$\mathbf{L}, \mathcal{L}, \mathbf{L}^\alpha$ - set of occupation measures for total cost eq. (3.1), expected average cost eq. (5.5), discounted cost eq. (3.17). 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.

$\mathbf{LP}_i$: Linear Programs which are equivalent to **COP**.

$\overline{m}(\cdot)$:= upper bound on h_α (Section 6.2 and **??**).
$\mathbf{M}$ - set of achievable costs for total cost eq. (4.18), expected average cost eq. (6.23). 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, relates to the definition of contracting MDPs, Section 2.5.
M^μ:= the set of measures q with $E^q\mu$ finite (defined in Section 2.5).
$M_1(G)$, $M(G)$, $\overline{M}(G)$:= set of probability measures over a set G, the set of measures over G. Mixed strategies over $G \subset U$.

$p^u_\beta(t; x) = P^u_\beta(X_t = x)$ and $p^u_\beta(t; x, a) = P^u_\beta(X_t = x, A_t = a)$, Section 2.1.

$\mathcal{P}$, P^u_β, E^u_β - transition probabilities; probablity generated by initial distribution β and policy u, and the corresponding expectation.

$\mathbf{Q}_{tc}$, $\mathbf{Q}_{ea}$, $\mathbf{Q}^{\alpha}$- the feasible set for the primal LPs. eq. (3.2), (5.5), (3.18).
Q- matrix, Section 2.5, Definition 3.2.
$q, \hat{q}$:= probability distribution; mixed policy with parameter q, Section 2.1.

r - deviation measure.

($\mathbf{S}1$)-($\mathbf{S}5$) conditions defined in Section 7.2 for the convergence of values and policies.

s, t, T: time, horizon length.

T_{tc}:= the dynamic programming operator for the total cost, defined in (4.1).
$T, \mathcal{T}$:= hitting times, and expected hitting times see eq. (3.8), (6.11), Example 4.1, and Section 8.5.

$u, v, U, \mathcal{U}, U_M, U_S, U_D$: policies, set of behavioral and non-behavioral policies, Markov, stationary, stationary deterministic policies, Section 2.1.

$\mathbf{V}_{tc}$, $\mathbf{V}_{ea}$ - sets related to primal achievable costs. eq. (4.19), (6.24).
$V = (V_1, ..., V_K)$- r.h.s. of constraints, Section 2.1.

w: stationary (randomized) policy, Section 2.1.
$W_\alpha(u; x)$:= the total cost from x to y, see (6.11).

x, y, z, X_t, $\mathbf{X}$ states, state at time t, state space, Section 2.1.

$\mathbf{Y}(x)$ sets of neighbouring states of x, Chapter 8.

α: discount factor.

β: initial distribution.

γ: parameter for time-sharing policy, probability measure.

λ: Lagrange multiplier.

$\mu, \xi, \tilde{\xi}$:= involved in the definition of contracting MDPs, Section 2.5, and Definitions 5.3 and 5.4 of uniform geometric recurrence and ergodicity.

σ := a constant in the Definition 5.4 of uniform geometric ergodicity.

ν:= measure. Used in showing that f_{tc} is continuous in $u \in U_S$.

χ: used in the proof of the compactness of L_s for the average cost.

ρ, ζ - decision variables in the primal LP; ρ corresponds to the occupation measure (3.2) and Section 3.4, (5.5) and Section 5.6. ζ will appear later in the multi-chain case, expected average cost and will correspond to the deviation measure.

(ϕ, ψ) - decision variables in the dual LP. ϕ correponds to the value for the finite horizon case and ψ- to the relative cost.

$\pi(\cdot)$: steady sate probabilities. $\pi(g)$:= steady state probailities corresponding to a stationary policy g.
$\pi^n(\cdot)$:= projection, Section 8.4.

Θ: value of the dual program.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399