CONSTRAINED MARKOV DECISION PROCESSES

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

To Tania and Einat

Preface

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

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

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

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

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

Eitan Altman, August 1998

Contents

CHAPTER 1

Introduction

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

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

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

1.1 Examples of constrained dynamic control problems

Telecommunications networks are designed to enable the simultaneous transmission of heterogeneous types of information: file transfers, interactive messages, computer outputs, facsimile, voice and video, etc... At the access to the network, or at nodes within the network itself, the different types of traffic typically compete for a shared resource. Typical performance measures are the transmission delay, the throughputs, probabilities of losses of packets (that stem from the fact that there are finite buffers at intermediate nodes of the network), etc... All these performance measures are determined by continuously monitoring and controlling the input flows into the network, by controlling the admission of new calls (or sessions), by controlling the allocation of the resources to different traffic, by routing decisions. Different types of traffic differ from each other by their statistical properties, as well as by their performance requirements. For example, for interactive messages it is necessary that the average end-to-end delay be limited. Strict delay constraints are important for voice traffic; there, we hardly distinguish between different delays as long as they are lower than some limit of the order of 0.1 second. When the delay increases beyond this limit, it becomes quickly intolerable. For non-interactive file transfer, we often wish to minimize delays or to maximize throughputs.

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

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

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

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

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

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

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

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

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

1.2 On solution approaches for CMDPs with expected costs

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

Several methods have been used in the past to solve this kind of CMDP. The first one, based on a Linear Program (LP), was introduced by Derman and Klein (1965), Derman (1970), and further developed by Derman and Veinott (1972), Kallenberg (1983), and Hordijk and Kallenberg (1984). It is based on an LP whose decision variables correspond to the occupation measure. The value of the LP is equal to the value of the CMDP, and there is a one to one correspondence between the optimal solutions of the LP and the optimal policies of the CMDP. This method is quite efficient (in terms of complexity of computations, and in the amount of decision variables, and hence memory requirements) for calculating the value of the CMDP (for the finite state and action space) for both the discounted or total cost, as well as the average cost with unichain structure. However, for the expected average cost with general multi-chain ergodic structure, the computation of an optimal policy is very costly and, as stated by Kallenberg (1983), it 'is unattractive for practical problems. The number of calculations is prohibitive' (p. 142). An alternative efficient way (again, in terms of complexity of calculations and memory requirements) for obtaining optimal policies from the LP for the average cost was obtained by Krass (1989). In Chapters 8, 10 and 11 we present the extension of the LP approach to the case of countable state space. (This is based on Altman and Shwartz, 1991a, and Altman, 1994, 1996, 1998).

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

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

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

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

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

1.3 Other types of CMDPs

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

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

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

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

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

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

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

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

An important problem in MDPs in general, and in constrained MDPs in particular, which is often encountered in applications, is of simultaneous learning and controlling. This occurs when some parameters of the problem are unknown to the decision maker. The standard cost criteria may be quite unsuitable for this type of situation. For example, the total expected discounted cost may not be well defined if we do not have any knowledge of the probability distribution. This required the introduction of new cost criteria. Schäl (1975) introduced an asymptotic discounted cost criterion for non-constrained MDPs, for which adaptive optimal policies combining estimation and control were investigated (Schäl, 1987, Hernandez-Lerma, 1989, and references therein). Altman and Shwartz (1991d) adapted these cost criteria to CMDPs and proposed several optimal adaptive techniques (1991b, 1991d). The solutions are based on ideas on sensitivity analysis of linear programs. An alternative solution approach based on stochastic approximations can be used to solve the adaptive MDP. This approach was used by Makowski and Shwartz (1992), Ma *et al.* (1992), Ma and Makowski (1988, 1992).

1.4 Cost criteria and assumptions

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

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

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

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

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

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

1. The first allows for quite general probabilistic assumptions, in particular, a tightness assumption, and yet requires the immediate costs to be

bounded from below; alternatively, even the tightness assumption may be relaxed and replaced by some stronger growth condition on the cost. This approach is due to Borkar (1983) and was adapted to constrained MDPs in Altman and Shwartz (1991a).

2. In the second framework we relax the boundedness on the immediate cost and require instead some tightness conditions as well as some uniform integrability ones. This framework will be shown to be equivalent to having a uniform Lyapunov function (due to Hordijk, 1977).
3. A further subclass of MDPs with uniform Lyapunov function that we shall briefly study is that of uniformly μ-recurrent MDPs, who were introduced and investigated by Dekker and Hordijk (1988), Spieksma (1990), and Dekker *et al.* (1994). This framework can be considered as the one corresponding to contracting MDPs.

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

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

1.5 The convex analytical approach and occupation measures

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

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

This type of analysis of occupation measure goes back to Derman (1970) who also made use of it for studying constrained MDPs (in finite state and action spaces). It was further developed by Kallenberg (1983) and Hordijk and Kallenberg (1984), and Feinberg (1995) (who considered the semi-Markov case). The properties of occupation measures corresponding to the infinite state space were investigated by Borkar (1988, 1990), Altman and Shwartz (1988, 1991a), Altman (1994, 1996, 1998), Spieksma (1990), and Feinberg and Sonin (1993, 1995). The study of occupation measures arises also in other related areas in control. In particular, in the controlled diffusions they have already been studied by Krylov (1985) and later by Borkar and Ghosh (1990, 1993).

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

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

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

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

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

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

1.6 Linear Programming and Lagrangian approach for CMDPs

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

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

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

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

An alternative LP approach (which can also be obtained by the Lagrangian technique) is the one that corresponds to the restriction of the constrained problem to mixed stationary-deterministic policies. The fact that these policies are dominating is established in Chapters 8 and 11, so that the restriction is without loss of optimality. The decision variables here are the measures over all stationary deterministic policies. An advantage of such formulation is that, even when the ergodic structure is general multi-chain, the same type of Linear Program applies for the expected average cost as well as the discounted cost. This fact allowed Tidball and Altman (1996b) to obtain convergence of the values and policies of discounted CMDP to those of expected average MDPs, as the discount factor tends to 1, for a general multi-chain structure. This approach extends the one by Feinberg (1993) that was derived for the case of finite state and action spaces. The LP has the same form as the one introduced by Altman and Shwartz (1993) for computing optimal time-sharing policies. We present these LPs at the end of Chapters 9 and 12.

1.7 About the methodology

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

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

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

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

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

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

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

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

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

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

how to solve the inf-sup problem using a single LP, which turns out to be the dual of the one obtained by the first approach. We illustrate the use of this approach in Chapter 5.

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

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

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

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

The costs can be written as

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

where

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

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

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

Moreover, we show in Chapters 3 and 10 that any ρ satisfying (1.1) equals the occupation measure corresponding to any stationary policy w that satisfies the following: for any state y for which $\sum_{a'\in\mathsf{A}} \rho(y,a') > 0$, w chooses action a at state y with probability

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

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

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

subject to (1.1) and

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

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

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

C_α therefore satisfies the inequalities

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We end this section with a brief discussion on the last part of the book:

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

1.8 The structure of the book

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

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

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

Some of the sections in the monograph are marked with an asterisk.

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

CHAPTER 18

References

E. Altman (1993), 'Asymptotic properties of constrained Markov decision processes', *ZOR – Methods and Models in Operations Research*, **37**, Issue 2, pp. 151-170.

E. Altman (1994), 'Denumerable constrained Markov decision processes and finite approximations', *Math. of Operations Research*, **19**, pp. 169-191.

E. Altman (1996), 'Constrained Markov decision processes with total cost criteria: occupation measures and primal LP', *ZOR – Mathematical Methods in Operations Research*, **43**, Issue 1, pp. 45-72.

E. Altman (1998), 'Constrained Markov decision processes with total cost criteria: Lagrange approach and dual LP', *ZOR – Mathematical Methods in Operations Research*, **48**, pp. 387-417, 1998.

E. Altman and V. A. Gaitsgory (1993), 'Stability and singular perturbations in constrained Markov decision problems', *IEEE Transactions on Automatic Control*, **38**, pp. 971-975.

E. Altman and V. A. Gaitsgory (1995), 'A hybrid (differential-stochastic) zero-sum game with fast stochastic part', *Annals of the International Society of Dynamic Games*, **3**, pp. 47-59.

E. Altman, A. Hordijk and L. C. M. Kallenberg (1996), 'On the value in constrained control of Markov chains', *ZOR – Methods and Models in Operations Research*, **44**, Issue 3, pp. 387-400.

E. Altman, A. Hordijk and F. M. Spieksma (1997), 'Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards', *MOR*, **22** No. 3, pp. 588-618.

E. Altman and A. Shwartz (1988), 'Markov optimization problems: state-action frequencies revisited', *27th IEEE Conference on Decision and Control*, Austin, Texas, December 1988 (invited paper).

E. Altman and A. Shwartz (1989), 'Optimal priority assignment: a time sharing approach', *IEEE Transactions on Automatic Control* **AC-34**, pp. 1089-1102.

E. Altman and A. Shwartz (1991a), 'Markov decision problems and state-action frequencies', *SIAM J. Control and Optimization,* **29**, pp. 786-809.

E. Altman and A. Shwartz (1991b), 'Adaptive control of constrained Markov chains', *IEEE Transactions on Automatic Control*, **36**, pp. 454-462.

E. Altman and A. Shwartz (1991c), 'Sensitivity of constrained Markov Decision Problems', *Annals of Operations Research*, **32**, pp. 1-22.

E. Altman and A. Shwartz (1991d), 'Adaptive control of constrained Markov chains: criteria and policies', *Annals of Operations Research* **28**, special issue on 'Markov Decision Processes', Eds. O. Hernández-Lerma and J. B. Lasserre, pp. 101-134.

E. Altman and A. Shwartz (1993), 'Time-sharing policies for controlled Markov chains', *Operations Research*, **41**, pp. 1116-1124.

E. Altman and A. Shwartz (1995), 'Constrained Markov games: Nash equilibria', submitted to *Annals of Dynamic Games.*

E. Altman and F. Spieksma (1995), 'The Linear Program approach in Markov decision problems revisited', *ZOR – Methods and Models in Operations Research*, **42**, Issue 2, pp. 169-188.

E. Altman and O. Zeitouni (1994), 'Rate of convergence of empirical measures and costs in controlled Markov chains and transient optimality', *Math. of Operations Research*, **19**, pp. 955-974.

J. Anderson and P. Nash (1987), *Linear Programming in Infinite-Dimensional Spaces*, Wiley, England.

A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh and S. I. Marcus (1993), 'Discrete-time controlled Markov processes with average cost criterion: a survey', *SIAM J. Control and Optimization*, **31**, pp. 282-344.

J. P. Aubin (1993), *Optima and Equilibria, An Introduction to Nonlinear Analysis*, Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.

R. J. Aumann (1964), 'Mixed and behavior strategies in infinite extensive games', Advances in Game Theory, *Ann. Math. Study*, **52**, pp. 627-650.

J. Bather (1973), 'Optimal decision procedures for finite Markov chains. Part II: Communicating systems', *Advances in Applied Probability*, **5**, pp. 521-540.

M. Bayal-Gursoy and K. W. Ross (1992), 'Variability sensitive Markov decision processes', *Math. of Operations Research*, **17**, pp. 558-571.

P. Bernhard (1992), 'Information and strategies in dynamic games', *SIAM J. Cont. and Opt.*, **30**, pp. 212-228.

F. J. Beutler and K. W. Ross (1985), 'Optimal policies for controlled Markov chains with a constraint', *J. Mathematical Analysis and Applications,* **112**, 236-252.

F. J. Beutler and K. W. Ross (1986), 'Time-average optimal constrained semi-Markov decision processes', *Advances of Applied Probability*, **18**, pp. 341-359.

P. Billingsley (1968), *Convergence of Probability Measures*, J. Wiley, New York.

J. R. Birge and R. J. Wets (1986), 'Designing approximating schemes for stochastic optimization problems', *Math. Programm. Study*, **27**, pp. 54-102.

V. S. Borkar (1983), 'On minimum cost per unit time control of Markov chains', *SIAM J. Control Optim.*, **22**, pp. 965-978.

V. S. Borkar (1988), 'A convex analytic approach to Markov decision processes', *Prob. Th. Rel. Fields*, **78**, pp. 583-602.

V. S. Borkar (1990), *Topics in Controlled Markov Chains*, Longman Scientific & Technical.

V. S. Borkar (1993), 'Controlled diffusions with constraints, II', *Journal of Math. Analysis and Appli.*, **176**, No. 2, pp. 310-321.

V. S. Borkar (1994), 'Ergodic control of Markov Chains with constraints — the general case', *SIAM J. Control and Optimization*, **32**, pp. 176-186.

V. S. Borkar and M. M. Ghosh (1990), 'Controlled diffusions with constraints', *Mathematical Analysis and Applications*, **152**, No. 1, pp. 88-108.

A. D. Bovopoulos and A. A. Lazar (1991), 'The effect of delayed feedback information on network performance', *Annals of Operations Res.* **36**, pp. 581-588.

E. B. N. Bui (1989), *Contrôle de l'allocation dynamique de trame dans un multiplexeur integrant voix et données*, TELECOM, Département Réseaux, Paris 89 E 005, June.

R. Cavazos-Cadena (1986), 'Finite-state approximations for denumerable state discounted Markov decision processes', *J. Applied Mathematics and Optimization*, **14**, pp. 27-47.

R. Cavazos-Cadena (1989), 'Weak conditions for the existence of optimal stationary policies in average cost Markov decision chains with unbounded cost', *Kybernetika*, **25**, 145-156.

R. Cavazos-Cadena (1992), 'Existence of optimal stationary policies in average Markov decision processes with a recurrent state', *Appl. Math. Optim.*, **26**, pp. 171-194.

R. Cavazos-Cadena and O. Hernández-Lerma (1992), 'Equivalence of Lyapunov stability criteria in a class of Markov decision processes', *Appl. Math. Optim.*, **26**, pp. 113-137.

R. Cavazos-Cadena and L. I. Sennott (1992), 'Comparing recent assumptions for the existence of average optimal stationary policies', *Operations Research Letters*, **11**, pp. 33-37.

K. L. Chung (1967), *Markov chains with stationary transition probabilities*, 2nd edition, Springer-Verlag, New York.

D. J. Daley and D. Vere-Jones (1988), *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York.

G. B. Dantzig, J. Folkman and N. Shapiro (1967), 'On the continuity of the minimum set of a continuous function', *J. Math. Anal. and Applications,* **17**, pp. 519-548.

G. T. De Ghellinck (1960), 'Les problèmes de décisions séquentielles', *Cahiers du Centre de Recherche Opérationelle*, **2**, pp. 161-179.

R. Dekker and A. Hordijk (1988), 'Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards', *Mathematics of Operations Research*, **13**, pp. 395-421.

R. Dekker, A. Hordijk and F. M. Spieksma (1994), 'On the relation between recurrence and ergodicity properties in denumerable Markov decision chains', *Math. Operat. Res.*, **19**, pp. 539-559.

E. V. Denardo (1970), 'On linear programming in a Markov decision problem', *Management Science*, **16**, pp. 281-288.

E. V. Denardo and B. L. Fox (1968), 'Multichain Markov renewal programs', *SIAM J. of Applied Math.*, **16**, pp. 468-487.

F. D'Epenoux (1960), 'Sur un problème de production et de stockage dans l'aléatoire', *Revue Française de Recherche Opérationelle*, **14**, pp. 3-16.

F. D'Epenoux (1963), 'A probabilistic production and inventory problem', *Management Science*, **10**, 98-108.

C. Derman (1970), *Finite State Markovian Decision Processes,* Academic Press, New York and London.

C. Derman and M. Klein (1965), 'Some remarks on finite horizon Markovian decision models', *Operations Research*, **13**, pp. 272-278.

C. Derman and R. E. Strauch (1966), 'On memoryless rules for controlling sequential control processes', *Ann. Math. Stat.*, **37**, pp. 276-278.

C. Derman and A. F. Veinott, Jr. (1972), 'Constrained Markov decision chains', *Management Science*, **19**, pp. 389-390.

J. L. Doob (1994), *Measure Theory*, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest.

N. Dunford and J. T. Schwartz (1988), *Linear operators*, part I, John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore.

E. Dynkin and A. Yushkevich (1979), *Controlled Markov Processes*, Springer-Verlag, Berlin.

A. Federgruen (1979), 'Geometric convergence of value-iteration in multichain Markov decision problems', *Adv. Appl. Prob.,* **11**, pp. 188-217.

E. A. Feinberg (1982), 'Non-randomized Markov and semi-Markov strategies in dynamic programming', *Theor. Probab. and its Applications,* **27**, pp. 116-126.

E. A. Feinberg (1986), 'Sufficient classes of strategies in discrete dynamic programming I: decomposition of randomized strategies and embedded models', *SIAM Theory Probab. Appl.*, **31**, pp. 658-668.

E. A. Feinberg (1986), 'Sufficient classes of strategies in discrete dynamic programming', *Theory Probability Appl.,* **31**, pp. 658-668.

E. A. Feinberg (1991), 'Non-randomized strategies in stochastic decision processes', *Annals of Operations Research*, **29**, pp. 315-332.

E. A. Feinberg (1995), 'Constrained semi-Markov decision processes with average rewards', *ZOR – Methods and Models in Operations Research*, **39**, pp. 257-288.

E. A. Feinberg and D. J. Kim (1996), 'Bicriterion optimization of an M/G/1 queue with a removable server', *Probab. in the Eng. and Inf. Sciences*, **10**, pp. 57-73.

E. A. Feinberg and M. I. Reiman (1994), 'Optimality of randomized trunk reservation', *Probability in the Engineering and Informational Sciences*, **8**, pp. 463-489.

E. A. Feinberg and I. Sonin (1983), 'Stationary and Markov policies in countable state dynamic programming', *Lecture Notes in Mathematics*, **1021**, pp. 111-129.

E. A. Feinberg and I. Sonin (1993), 'The existence of an equivalent stationary strategy in the case of discount factor equal one', unpublished draft.

E. A. Feinberg and I. Sonin (1995), 'Notes on equivalent stationary policies in Markov decision processes with total rewards', *ZOR – Methods and Models in Operations Research*, **44**, pp. 205-221.

E. A. Feinberg E. and A. Shwartz (1995), 'Constrained Markov decision models with weighted discounted rewards', *Math. of Operations Research*, **20**, pp. 302-320.

E. A. Feinberg E. and A. Shwartz (1996), 'Constrained discounted dynamic programming', *Math. of Operations Research*, **21**, pp. 922-945.

A. V. Fiacco (1974), 'Convergence properties of local solutions of convex optimization problems', *J. Optim. Theory Appl.*, **13**, pp. 1-12.

J. A. Filar, L. C. M Kallenberg and H. M. Lee (1989), 'Variance-penalized Markov decision processes', *Math. of Operations Research*, **14**, pp. 147-161.

J. A. Filar and H. M. Lee (1985), 'Gain/variability tradeoffs in undiscounted Markov decision processes', *Proceedings of 24th Conference on Decision and Control IEEE*, pp. 1106-1112.

L. Fisher (1968), 'On recurrent denumerable decision processes', *Ann. Math. Stat.*, **39**, pp. 424-434.

L. Fisher and S. M. Ross (1968), 'An example in denumerable decision processes', *Ann. Math. Stat.,* **39**, pp. 674-675.

V. A. Gaitsgory and A. A. Pervozvanskii (1986), 'Perturbation theory for mathematical programming problems', *JOTA*, pp. 389-410.

K. Golabi, R. B. Kulkarni and G. B. Way (1982), 'A statewide Pavement Management System', *Interfaces*, **12**, pp. 5-21.

M. Haviv (1995), 'On constrained Markov decision processes', *OR Letters*, **19**, Issue 1, pp. 25-28.

W. R. Heilmann (1977), 'Generalized linear programming in Markovian decision problems', *Bonner Math. Schriften*, **98**, pp. 33-39.

W. R. Heilmann (1978), 'Solving stochastic dynamic programming 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, New York, Berlin, Heidelberg, London, Paris, Tokyo.

O. Hernández-Lerma and D. Hernández-Hernández (1994), 'Discounted cost Markov decision processes on Borel spaces: the linear programming formulation', *J. of Math. Anal. and Appl.*, **183**, pp. 335-351.

O. Hernández-Lerma and J. B. Lasserre (1994), 'Linear programming and average optimality on Borel spaces-unbounded costs', *SIAM J. Control and Optimization*, **32**, pp. 480-500.

O. Hernández-Lerma and J. B. Lasserre (1995), *Discrete-Time Markov Control Processes, Basic Optimality Criteria*, Springer-Verlag, New York, Berlin, Heidelberg.

K. Hinderer (1970), *Foundation of Non-Stationary Dynamic Programming with Discrete Time Parameter*, Vol. 33, Lecture Notes in Operations Research and Mathematical Systems, Springer-Verlag, Berlin.

A. Hordijk (1977), *Dynamic Programming and Markov Potential Theory*, second edition, Mathematical Centre Tracts 51, Mathematisch Centrum, Amsterdam.

A. Hordijk and L. C. M. Kallenberg (1979), 'Linear programming and Markov decision chains', *Management Science*, **25**, pp. 352-362.

A. Hordijk and L. C. M. Kallenberg (1984), 'Constrained undiscounted stochastic dynamic programming', *Mathematics of Operations Research,* **9**, pp. 276-289.

A. Hordijk and J. B. Lasserre (1994), 'Linear programming formulation of MDPs in countable state space: the multichain case', *ZOR – Methods and Models in Operations Research*, **40**, pp. 91-108.

A. Hordijk and F. Spieksma (1989), 'Constrained admission control to a queuing system', *Advances of Applied Probability*, **21**, pp. 409-431.

R. Horn and C. R. Johnson (1985), *Matrix Analysis*, Cambridge Univ. Press.

M. T. Hsiao and A. A. Lazar (1991), 'Optimal decentralized flow control of Markovian queueing networks with multiple controllers', *Performance Evaluation*, **13**, pp. 181-204.

Y. Huang and L. C. M. Kallenberg (1994), 'On finding optimal policies for Markov decision chains: A unifying framework for mean-variance tradeoffs', *Math. of Operations Research*, **19**, pp. 434-448.

D. Kadelka (1983), 'On randomized policies and mixtures of deterministic policies in dynamic programming', *Methods of Operations Research*, **46**, pp. 67-75.

L. C. M. Kallenberg (1983), *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tracts 148, Amsterdam.

L. C. M. Kallenberg (1994), 'Survey of linear programming for standard and nonstandard Markovian control problems, Part I: Theory', *ZOR – Methods and Models in Operations Research*, **40**, pp. 1-42.

P. Kanniappan and S. M. A. Sastry (1974), 'Uniform convergence of convex optimization problems', *J. Math. Anal. Appli.*, **96**, pp. 1-12.

H. Kawai (1987), 'A variance minimization problem for a Markov decision process', *European Journal of Operations Research*, **31**, pp. 140-145.

J. G. Kemeney, J. L. Snell and A. W. Knapp (1976), *Denumerable Markov Chains*, Springer-Verlag.

L. Kleinrock (1976), *Queuing systems, Volume I.* John Wiley, New York.

P. Kolesar (1970), 'A Markovian model for hospital admission and scheduling', *Management Science*, **16**, pp. 384-396.

G. M. Koole (1988), *Stochastische Dynamische Programmering met Bijvoorwaarden* (translation: Stochastic dynamic programming with additional constraints) Master thesis, Leiden University, The Netherlands.

Y. A. Korilis and A. Lazar (1995a), 'On the existence of equilibria in noncooperative optimal flow control', *J. of the Association for Computing Machinery*, **42**, No. 3, pp. 584-613.

Y. A. Korilis and A. Lazar (1995b), 'Why is flow control hard: optimality, fairness, partial and delayed information', preprint.

D. Krass (1989), *Contributions to the Theory and Applications of Markov Decision Processes*, Ph.D. thesis, Department of Mathematical Sciences, Johns Hopkins Univ., Baltimore, MD.

M. Krein and D. Milman (1940), 'On extreme points of regularly convex sets', *Studia Math.*, **9**, pp. 133-138.

N. Krylov (1985), ‘Once more about the connection between elliptic operators and Ito’s stochastic equations’, *Statistics and Control of Stochastic Processes*, Steklov Seminar 1984 (Krylov N. *et al.*, Eds.), Optimization Software, New York, 69-101.

H. W. Kuhn (1953), ‘Extensive games and the problem of information’, *Ann. Math. Stud.*, **28**, pp. 193-216.

H. Kushner and J. Kleinman (1971), ‘Mathematical programming and the control of Markov chains’, *Internat. J. Control*, **13**, pp. 801-820.

J. B. Lasserre (1994), ‘Average optimal stationary policies and linear programming in countable state Markov decision processes’, *J. Math. Anal. Appl.*, **183**, pp. 233-249.

A. Lazar (1983), ‘Optimal flow control of a class of queuing networks in equilibrium’, *IEEE Transactions on Automatic Control*, **28**, pp. 1001-1007.

M. B. Lignota and J. Morgan (1992), ‘Convergences of marginal functions with dependent constraints’, *Optimization*, **23**, pp. 189-213.

R. Lucchetti and R. J. B Wets (1993), ‘Convergence of minima of integral functionals, with applications to optimal control and stochastic optimization’, *Statistics and Decisions*, **11**, pp. 69-84.

D.-J. Ma and A. M. Makowski (1988), ‘A class of steering policies under a recurrence condition’, *27th IEEE Conference on Decision and Control*, Austin, TX, December, pp. 1192-1197.

D.-J. Ma and A. M. Makowski (1992), ‘A class of two-dimensional stochastic approximations and steering policies for Markov decision processes’, *31st IEEE Conference on Decision and Control*, Tucson, Arizona, pp. 3344-3349.

D.-J. Ma, A. M. Makowski and A. Shwartz (1990), ‘Stochastic approximations for finite state Markov chains’, *Stochastic Processes and Their Applications*, **35**, pp. 27-45.

B. Maglaris and M. Schwartz (1982), ‘Optimal fixed frame multiplexing in integrated line- and packet-switched communication networks’, *IEEE Transactions on Information Theory*, **IT-28**, pp. 263-273.

A. M. Makowski and A. Shwartz (1987), ‘Recurrence properties of a system of competing queues with applications’, Research report EE Pub. No. 627, Technion, Haifa, Israel.

A. M. Makowski and A. Shwartz (1992), ‘Stochastic approximations and adaptive control of a discrete-time single server network with random routing', *SIAM J. Control and Optimization*, **30**, pp. 1476-1506.

A. S. Manne (1960), 'Linear programming and sequential decisions', *Management Science*, **6**, pp. 259-267.

S. Meyn and R. Tweedie (1994), *Markov Chains and Stochastic Stability*, Springer-Verlag, New York.

P. Nain and K. W. Ross (1986), 'Optimal priority assignment with hard constraint', *Transactions on Automatic Control,* **31**, pp. 883-888.

A. S. Nowak (1985), 'Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space', *JOTA*, **45**, pp. 592-602.

A. A. Pervozvanskii and V. A. Gaitsgory (1988), *Theory of Suboptimal Decision: Decomposition and Aggregation*, Kluwer Academic Publishers, Dordrecht.

A. B. Piunovskiy (1993), 'Control of random sequences in problems with constraints', *Theory Probab. Appl.*, **38**, No. 4, translated from Russian.

A. B. Piunovskiy (1994), 'Control of jump-like processes in constrained problems', *Avtomatika i Telemekhanika*, **4**, pp. 75-89. Translated into English in *Automation and Remote Control*, **55**, No. 4, 1994.

A. B. Piunovskiy (1995), 'Multicriteria control problems for stochastic jump processes', *Proceedings of 3rd European Control Conference*, Rome, Italy, September, pp. 492-495.

A. B. Piunovskiy (1996), 'A multicriteria model of optimal control of a stochastic linear system', *Automation and Remote Control* **57**, No. 6, Part 1, pp. 831-842.

A. B. Piunovskiy (1997a), *Optimal Control of Random Sequences in Problems with Constraints*, *Mathematics and its Applications*, Kluwer Academic Publishers, Dordrecht, Boston, London.

A. B. Piunovskiy (1997b), 'Optimal control of stochastic sequences in sequences with constraints', *Stochastic Analysis and Applications*, No. 2.

M. Puterman (1994), *Markov Decision Processes*, John Wiley & Sons, New York.

T. E. S. Raghavan and J. A. Filar (1991), 'Algorithms for stochastic pames – a survey', *ZOR – Methods and Models in Operations Research*, **35**, pp. 437-472.

D. Revuz (1975), *Markov Chains*, North-Holland, Amsterdam, The Netherlands.

R. T. Rockafellar (1989), *Conjugate Duality and Optimization*, Society for Industrial and Applied Mathematics, 2nd printing, Philadelphia.

K. W. Ross (1989), 'Randomized and past-dependent policies for Markov decision processes with multiple constraints', *Operations Research*, **37**, pp. 474-477.

K. W. Ross and B. Chen (1988), 'Optimal scheduling of interactive and non-interactive traffic in telecommunication systems', *IEEE Transactions on Automatic Control,* **33**, pp. 261-267.

K. Ross and R. Varadarajan (1989), 'Markov decision processes with sample path constraints: the communicating case', *Operations Research*, **37**, pp. 780-790.

K. Ross and R. Varadarajan (1991), 'Multichain Markov decision processes with a sample path constraint: a decomposition approach', *Math. of Operations Research*, **16**, pp. 195-207.

H. L. Royden (1988), *Real Analysis*, 3rd edition, Macmillan Publishing Company, New York.

M. Schäl (1975), 'Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal', *Z. Wahrscheinlichkeitstheorie und verw. Geb.,* **32**, pp. 179-196.

M. Schäl (1987), 'Estimation and control in discounted dynamic programming', *Stochastics*, **20**, pp. 51-71.

I. E. Schochetman (1990), 'Pointwise versions of the maximum theorem with applications to optimization', *Appl. Math. Lett.*, **3**, pp. 89-92.

I. E. Schochetman and R. L. Smith (1991), 'Convergence of selections with applications in optimization', *J. Math. Anal. Appl.*, **155**, pp. 278-242.

L. I. Sennott (1989), 'Average cost optimal stationary policies in average cost Markov decision processes', *Operations Research*, **37**, pp. 626-633.

L. I. Sennott (1991), 'Constrained discounted Markov decision chains', *Probability in the Engineering and Informational Sciences*, **5**, pp. 463-475.

L. I. Sennott (1993), 'Constrained average cost Markov decision chains', *Probability in the Engineering and Informational Sciences*, **7**, pp. 69-83.

L. I. Sennott (1997), 'On computing average optimal policies with application to routing to parallel queues', *Mathematical Methods of Operations Research*, **45**, pp. 45-62.

L. S. Shapley (1953), 'Stochastic games', *Proceedings Nat. Acad. of Science USA*, **39**, pp. 1095-1100.

N. Shimkin (1994), 'Stochastic games with average cost constraints', *Annals of the International Society of Dynamic Games, Vol. 1: Advances in Dynamic Games and Applications*, Eds. T. Basar and A. Haurie, Birkhauser, Boston.

M. J. Sobel (1985), 'Maximal mean/standard deviation ratio in undiscounted MDP', *OR Letters*, **4**, pp. 157-159.

M. J. Sobel (1994), 'Mean-variance tradeoffs in an undiscounted MDP', *Operations Research*, **42**, pp. 175-188.

F. M. Spieksma (1990), *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, Ph.D. thesis, University of Leiden, The Netherlands.

R. Sznadjer and J. A. Filar (1992), 'Some comments on a theorem of Hardy and Littlewood', *J. Optim. Theory Appl.*, **75**, pp. 210-218.

L. C. Thomas and D. Stengos (1985), 'Finite state approximation algorithms for average cost denumerable state Markov decision processes', *OR Spectrum*, **7**, pp. 27-37.

M. Tidball and E. Altman (1996a), 'Approximations in dynamic zero-sum games, I', *SIAM J. Control and Optimization*, **34**, No. 1, pp. 311-328.

M. Tidball, O. Pourtallier and E. Altman (1996b), 'Continuity of optimal values and solutions of convex optimization, and constrained control of Markov chains', submitted to *SIAM J. Control and Optimization*.

M. Tidball, O. Pourtallier and E. Altman (1997), 'Approximations in dynamic zero-sum games, II', *SIAM J. Control and Optimization*, **35**, pp. 2101-2117.

F. Vakil and A. A. Lazar (1987), 'Flow control protocols for integrated networks with partially observed voice traffic', *IEEE Transactions on Automatic Control*, **AC-32**, pp. 2-14.

J. Van Der Wal (1981a), *Stochastic Dynamic Programming*, Mathematical Centre Tract 139, Mathematisch Centrum, Amsterdam.

J. Van Der Wal (1981b), 'On stationary strategies', Eindhoven Univ. of Technology, Dept. of Math., Memorandum-COSOR 81-14, 1981.

J. Wessels (1977), 'Markov Games with unbounded rewards', *Dynamische Optimierung*, M. Schäl (Editor) Bonner Mathematische Schriften, Nr. 98, Bonn.

D. J. White (1980), 'Finite state approximations for denumerable state infinite horizon discounted Markov decision Processes', *J. Mathematical Analysis and Applications*, **74**, pp. 292-295.

D. J. White (1982), 'Finite state approximations for denumerable state infinite horizon discounted Markov decision processes with unbounded rewards', *J. Mathematical Analysis and Applications*, **86**, pp. 292-306.

D. J. White (1987), 'Utility, probabilistic constraints, mean variance of discounted rewards in Markov decision processes', *OR Spectrum*, **9**, pp. 13-22.

D. J. White (1994), 'A mathematical programming approach to a problem in variance penalized Markov decision processes', *OR Spectrum*, **15**, pp. 225-230.

W. Whitt (1978), 'Approximations of dynamic programs, I', *Mathematics of Operations Research*, **3**, No. 3, pp. 231-243.

W. Whitt (1980), 'Representation and approximation of noncooperative sequential games', *SIAM J. Control and Opt.*, **18**, No. 1, pp. 33-43.

C. V. Winden and R. Dekker (1994), 'Markov Decision Models for Building Maintenance: A Feasibility Study', Report 9473/A, ERASMUS University Rotterdam, The Netherlands.

D. Williams (1992), *Probability and Martingales*, Cambridge University Press, Cambridge.

A. A. Yushkevich (1973), 'On a class of strategies in general Markov decision models', *Theory Probab. Appl.* **18**, pp. 777-779.

Index